Aplikacyjne metody obliczeniowe oraz zarządzanie danymi

Aplikacyjne metody obliczeniowe oraz zarządzanie danymi

Redakcja:
Jakub Pizoń
Beata A. Nowak

Lublin 2017

Recenzenci:

- prof. dr hab. inż. Piotr Kacejko
- prof. dr hab. inż. Jerzy Lipski
- prof. dr hab. Jolanta Rzymowska
- dr hab. inż. Marcin Chodźko
- dr inż. Arkadiusz Gola
- dr inż. Grzegorz Kłosowski
- dr inż. Mirosław Szala
- dr inż. Monika Ostapiuk
- dr inż.Tomasz Cieplak
- dr Lubomira Wengler
- dr Magdalena Smoleń-Wawrzusiszyn
- dr n. med. Marta Łuczyk
- dr inż. Rafał Sochaczewski

Wszystkie opublikowane rozdziały otrzymały pozytywne recenzje.

Skład i łamanie:
Agnieszka Pytka

Projekt okładki:
Marcin Szklarczyk

ISBN 978-83-65598-94-3

Wydawca:
Wydawnictwo Naukowe TYGIEL sp. z o.o.
ul. Głowackiego 35/341, 20-060 Lublin
www.wydawnictwo-tygiel.pl

Spis treści

Kamila Szewczyk[1], Anna Duda[2], Anna Kosiacka[3]

Diagnostyka współliniowości w modelu regresji liniowej

1. Wstęp

Analiza regresji jest metodą statystyczną, która pozwala wykryć zależności między zmiennymi na podstawie zgromadzonego zbioru danych rzeczywistych [1, 2]. W praktyce rzadko spotykane są zbiory danych, które idealnie sprawdzałby się we wszystkich aspektach analizy regresji. Zbiory te często posiadają jakieś wady, zazwyczaj są to źle zanotowane dane, bądź zbiór posiada punkty oddalone, czy też występują w nim relacje współliniowe [3-6]. Ostatnim przypadkiem zajmiemy się w tej pracy.

Omówimy czym dokładnie jest współliniowość w modelu regresji, wymienimy kilka najbardziej popularnych, znanych wszystkim badaczom analizy regresji, metod wykrywania współliniowości oraz wskażemy metodę, która wykazuje się największą precyzją w naszym problemie. Jest to niezwykle ważne z punktu widzenia jakości otrzymanego modelu [3, 4]. Nasze rozważania poprzemy odpowiednim przykładem.

Przedstawiona metoda jest stosunkowo nowa, rzadko omawiana w polskiej literaturze oraz wykazuje się lepszą skutecznością od metod tradycyjnych.

2. Wprowadzenie

Diagnostyką współliniowości będziemy się zajmować w kontekście liniowego modelu regresji [1, 2]:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{1}$$

gdzie:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{2}$$

[1] kamila.szewczyk@ajd.czest.pl, Instytut Fizyki, Wydział Matematyczno-Przyrodniczy, Akademia im. Jana Długosza w Częstochowie, www.ajd.czest.pl
[2] anna.duda@wip.pcz.pl, Instytut Fizyki, Wydział Inżynierii Produkcji i Technologii Materiałów, Politechnika Częstochowska, www.pcz.pl
[3] akosiacka@vp.pl, Instytut Fizyki, Wydział Inżynierii Produkcji i Technologii Materiałów, Politechnika Częstochowska, www.pcz.pl

$\boldsymbol{Y}$ - wektor kolumnowy $n \times 1$ zmiennych objaśnianych (zmienna losowa),
$\boldsymbol{X}$ - macierz zmiennych objaśniających $n \times p$ (zmienne deterministyczne),
$\boldsymbol{\beta}$ - wektor kolumnowy $p \times 1$ parametrów regresji,
$\boldsymbol{\epsilon}$ - wektor kolumnowy $n \times 1$ składników losowych (zmienne niezależne).

Ponadto, wektor $\boldsymbol{b} = \boldsymbol{XTX} - \mathbf{1}\boldsymbol{XTY}$ jest estymatorem wyznaczonym metodą najmniejszych kwadratów [1, 2, 7]. Będziemy analizować miary diagnozujące współliniowość jednej lub więcej relacji wśród kolumn macierzy $\boldsymbol{X}$ [3, 4].

Czym tak naprawdę jest współliniowość w kontekście analizy regresji? Zanim to wyjaśnimy zastanówmy się najpierw jak definiujemy zwykłą zależność liniową dwóch wektorów. Więc, dwa wektory są współliniowe, jeśli leżą w podprzestrzeni wymiaru 1. Można tą definicję uogólnić dla kilku wektorów. Zatem, wektorów jest liniowo zależnych, jeśli jeden z tych wektorów jest dokładną kombinacją liniową innych, czyli leży w przestrzeni mniejszego wymiaru niż [3, 4]. Są to definicje dokładnej liniowej zależności. W rzeczywistości bardzo rzadko zdarza się tak, żeby w zbiorze danych, które są zebrane na podstawie obserwacji wszelakich sytuacji życiowych, występowały dokładne relacje współliniowe. Jednak, nawet jeśli dokładna zależność liniowa nie występuje w praktyce, okazuje się, że relacje, które są bliskie współliniowości powodują kłopoty w poprawnym oszacowaniu modelu. Dlatego w kontekście analizy regresji, zmiennymi liniowo zależnymi będziemy określać zmienne, które są "prawie" współliniowe. Zatem, dwie zmienne są liniowo zależne, jeśli leżą one prawie w tej samej linii, to znaczy, jeśli kąt pomiędzy nimi jest mały. Można to również uogólnić do kilku zmiennych współliniowych. Więc zmiennych jest współliniowych, jeśli jedna z nich leży prawie w przestrzeni rozpiętej przez pozostałe − 1 zmienne, czyli jeśli kąt, pomiędzy jedną zmienną, a jej rzutem prostopadłym, jest mały [3, 4].

Zastanówmy się teraz dlaczego współliniowość jest tak bardzo niepożądana przez analityków regresji. Relacje liniowo zależne w modelu regresji kreują sytuacje, w których statystyczna estymacja metodą najmniejszych kwadratów jest poważnie zagrożona. Oznacza to, że występowanie współliniowości w zbiorze danych może bardzo zaburzyć poprawność oszacowania całego modelu regresji.

3. Tradycyjne techniki wykrywania relacji współliniowych w modelu regresji liniowej

Wymienimy tutaj tylko kilka popularnych technik diagnozowania relacji współliniowych, ale zapewnimy że takich praktyk jest znacznie więcej oraz dla każdej można znaleźć jakieś wady [3, 4].

Pierwszą metodą, bardzo często omawianą w polskiej literaturze dotyczącej analizy regresji jest diagnostyka oparta na odwrotności macierzy korelacji $\boldsymbol{R}^{-1}$, która bada dokładnie występowanie relacji współliniowych na podstawie współczynnika VIF (variance inflation factor). Załóżmy, że macierz $\boldsymbol{X}$ jest scentrowana i skalowana do jednostkowej długości, wtedy $\boldsymbol{R}^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Elementy przekątnej macierzy $\boldsymbol{R}^{-1}$, r^{ii} wyznaczamy następująco:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{3}$$

Gdzie R_i^2 jest wielokrotną korelacją współczynnika $\boldsymbol{X}_i$ oszacowaną dla pozostałych zmiennych objaśniających. Powyższy wzór pochodzi z faktu, iż wariancja i-tego oszacowanego współczynnika $\sigma_{b_i}^2$ z regresji pełnego zbioru danych $\boldsymbol{X}$ obejmuje relację:

$$\sigma_{b_i}^2 = \frac{\sigma^2}{\boldsymbol{X}_i^T \boldsymbol{X}_i} VIF_i, \tag{4}$$

gdzie σ^2 jest wariancją z zakłóceń modelu $\boldsymbol{\epsilon}$. W przedstawionej metodzie jest jednak kilka poważnych wad. Po pierwsze, okazuje się, że wysoki VIF jest wystarczający do wykrycia relacji współliniowych, ale nie jest konieczny! Wynika to z faktu, że miara ta jest oparta na macierzy korelacji. Często uważa się, że wysoka korelacja dwóch zmiennych wchodzących do modelu wskazuje na relacje współliniowe, co jest poważnym błędem. Wróćmy jeszcze na chwilę do definicji relacji liniowo zależnych. Mały kąt pomiędzy dwoma wektorami (lub wektorem a jego rzutem prostopadłym) nie jest równoważny wysokiej korelacji pomiędzy nimi. Wynika to z faktu, że wysoka korelacja z pewnością wskazuje na mały kąt, ale odwrotność tego stwierdzenia nie jest już prawdą. Zdarzają się przypadki, że kąt pomiędzy dwoma wektorami jest mały wskazując tym samym współliniowość, natomiast pomiędzy ich scentrowanymi wektorami występuje kąt prosty wskazując zero korelacji. Zatem współliniowość i korelacja nie są tymi samymi rzeczami. Jest to ważne stwierdzenie, ponieważ te dwa terminy są często mylone, ze szkodą dla wszystkich późniejszych analiz i dyskusji. Po drugie miara ta, nie pokazuje, które dokładnie relacje są zaangażowane we współliniowość. Nie wskazuje też czy w liniową zależność włączone tylko dwie zmienne, czy też więcej. Po trzecie, nie wymyślono jeszcze żadnego sprawdzonego sposobu, który określałby jaka jest granica dla określenia wysokiego bądź niskiego. Wszystkie kryteria jakie przyjmuje się w literaturze są określone czysto heurystycznie [3, 4].

Następna technika skupia się na badaniu $\det(\boldsymbol{X}^T\boldsymbol{X})$ lub $\det(\boldsymbol{R})$. Współliniowe relacje pośród kolumn macierzy $\boldsymbol{X}$ oznaczają małą wartość wyznacznika $\boldsymbol{X}^T\boldsymbol{X}$, sugerując $\det(\boldsymbol{X}^T\boldsymbol{X})$ jako miarę współliniowości. Niestety wyznacznik ten ma wiele braków w swojej roli. Po pierwsze, jak mała ma być wartość wyznacznika, aby wskazywał na wystąpienie współliniowości? Wyznacznik ten również nie wskazuje numerów obserwacji zaangażowanych w liniowe zależności. Obecność jakiejkolwiek współliniowości może powodować małą wartość wyznacznika, a tym samym maskować istnienie innych. Najważniejsze jest to, że mały wyznacznik może być konieczny dla współliniowości, ale nie wystarczający. Zdarzają się bowiem przypadki, w których macierze mają małe wyznaczniki, a wcale nie posiadają

relacji współliniowych. To samo dotyczy det($\boldsymbol{R}$), zachowuje on bowiem pierwsze dwie słabości dane dla det($\boldsymbol{X}^T\boldsymbol{X}$) [3, 4].

Ostatnia z podanych tutaj technik dotyczy badania wartości własnych oraz wektorów własnych macierzy $\boldsymbol{X}^T\boldsymbol{X}$ lub $\boldsymbol{R}$. Jest popularną techniką, stosowaną od wielu lat. System własny (wektory własne i wartości własne) macierzy $\boldsymbol{X}^T\boldsymbol{X}$ lub $\boldsymbol{R}$ był sugerowany jako miara dla wykrycia współliniowości. Wektory własne macierzy $\boldsymbol{X}^T\boldsymbol{X}$ tworzą zbiór p niezerowych wektorów $\boldsymbol{\xi}$, które spełniają warunek $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\xi}=\lambda\boldsymbol{\xi}$, gdzie λ jest wartością własną, która odpowiada danemu wektorowi własnemu. Zależność liniowa występuje wtedy, gdy wektor własny ma wartość własną równą zero, tzn. $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\xi}=\boldsymbol{0}$ lub równoważnie $\boldsymbol{X}\boldsymbol{\xi}=\boldsymbol{0}$, co oznacza, że macierz $\boldsymbol{X}$ posiada dokładną współliniowość wśród swoich kolumn. W związku z tym sugeruje się, że małe wartości własne $\boldsymbol{X}^T\boldsymbol{X}$ wskazują bliską zależność liniową, tzn. $\boldsymbol{X}\boldsymbol{\xi}\approx\boldsymbol{0}$. Sposób ten ma przewagę nad wymienionymi wcześniej, ponieważ jest zdolny do wykrycia kilku współistniejących współliniowych relacji. Dostaniemy bowiem, każdą małą wartość własną dla każdej bliskiej zależności liniowej. Niestety, tu znów pojawia się problem, jak mała musi być ta wartość, aby wskazywała na relację współliniowości? Powstało kilka teorii na ten temat, lecz żadna nie wydaje się wartościową, ponieważ każda ma jakieś wady [3, 4].

Metody opisane wyżej nie są w pełni satysfakcjonujące w diagnozowaniu współliniowości. Jednak podstawa techniki wykrywającejzależności liniowe w modelu regresji jest na wyciągnięcie ręki. Okazuje się, że koncepcje rozwijane w teorii analizy numerycznej są zdolne do oddania prawdziwego znaczenia ostatniej miary opisanej wyżej, czyli opartej na systemie własnym. Uzasadnieniem tego stwierdzenia jest to, że współliniowość dotyczy numerycznej lub geometry cznej charakterystyki danej macierzy $\boldsymbol{X}$. Za to nie dotyczy żadnego statystycznego aspektu, który może generować macierz $\boldsymbol{X}$ lub może być istotny w liniowym modelu regresji. Oznacza to, że współliniowość jest problemem danych, nie statystycznym problemem. Analiza numeryczna, przykładowo, zajmuje się właściwościami zbioru **A** w liniowym systemie równań **Az=c**, które pozwalają uzyskać rozwiązanie z numeryczną stabilnością. W kontekście średniokwadratowym, współliniowość jest istotna. Estymatory wyznaczone metodą najmniejszych kwadratów są rozwiązaniem liniowego systemu równań normalnych $(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{Y}$. Macierzą wariancji-kowariancji tego estymatora jest $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Zatem współliniowość kolumn macierzy $\boldsymbol{X}$, ma swoje konsekwencje dla macierzy $\boldsymbol{A}=\boldsymbol{X}^T\boldsymbol{X}$, której złe uwarunkowanie powoduje numeryczną niestabilność rozwiązań dla $\boldsymbol{b}$ oraz dla jego macierzy kowariancji. Techniki oparte na systemie własnym, można rozszerzyć do diagnostyki współliniowości, a mianowicie wartości własne mogą służyć do formowania zbioru warunkowych indeksów, które pozwalają nam na określenie siły oraz numeru obserwacji zaangażowanych w relacje bliskich zależności liniowych oraz wektory własne i ich wartości własne, użyte razem, kształtują zbiór ilorazu wariancji (variance-decomposition proportion), które pozwalają nam określić zmienne zaangażowane [6, 7].

4. Procedura diagnostyki współliniowości w modelu regresji liniowej

Aby w pełni zrozumieć poniższą procedurę należy najpierw przeprowadzić kilka prostych rozważań [3, 4].

4.1. Rozkład na wartości osobliwe macierzy $\boldsymbol{X}$

Rozkład na wartości osobliwe macierzy $\boldsymbol{X}$ ma powiązanie z systemem własnym macierzy $\boldsymbol{X}^T\boldsymbol{X}$. W rozkładzie na wartości własne, każdą macierz $\boldsymbol{X}$ wymiaru $n \times p$ gdzie $n \geq p$, można zapisać następująco:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \tag{5}$$

gdzie $\boldsymbol{U}^T\boldsymbol{U}=\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}_p$ oraz **D** jest macierzą diagonalną, z nieujemnymi elementami na przekątnej $\mu_1,\ldots,\mu_p$, nazwanymi osobliwymi wartościami macierzy $\boldsymbol{X}$. Powyższy układ posiada interesujące właściwości:

- macierz $\boldsymbol{U}$ posiada ortogonalne kolumny;
- macierz $\boldsymbol{V}$ ma ortogonalne kolumny i wiersze, tzn. że $\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{V}\boldsymbol{V}^T=\boldsymbol{I}_p$;
- $\boldsymbol{D}$ jest macierzą diagonalną, z nieujemnymi elementami na przekątnej.

Zakładamy tutaj, że $\boldsymbol{U}$ jest wymiaru $n \times p$ oraz $\boldsymbol{D}$ i $\boldsymbol{V}$ mają wymiar $p \times p$. Inne założenia są tutaj jak najbardziej możliwe, lecz na potrzeby naszej procedury przyjmijmy nasze wytyczne. Badając dokładniej powyższe wyrażenie (5), możemy zauważyć, że kolumny $\boldsymbol{X}$ są liniową kombinacją kolumn $\boldsymbol{U}$ oraz wiersze $\boldsymbol{X}$ są liniową kombinacją kolumn $\boldsymbol{V}$, a więc kolumny $\boldsymbol{U}$ są ortogonalną bazą dla kolumn macierzy $\boldsymbol{X}$ oraz kolumny (i wiersze) $\boldsymbol{V}$ są ortogonalną bazą przestrzeni wierszy macierzy $\boldsymbol{X}$.

Użycie rozkładu na wartości osobliwe preferuje się bardziej, niż systemu własnego z kilku prostych powodów:

- rozkład osobliwy koncentruje się na macierzy $\boldsymbol{X}$, która jest źródłem naszej uwagi, a nie na $\boldsymbol{X}^T\boldsymbol{X}$;
- długość $\text{II}\boldsymbol{a}\text{II}$ liniowej kombinacji macierzy $\boldsymbol{X}$, której szukamy do minimalizacji wyrażenia $\boldsymbol{X}\boldsymbol{v}=\boldsymbol{a}$, jest dobrze określona, zarówno w pierwiastku kwadratowym wartości własnych macierzy $\boldsymbol{X}^T\boldsymbol{X}$ jak i w wartościach osobliwych macierzy $\boldsymbol{X}$;
- system własny oraz rozkład na wartości osobliwe wydają się, dla naszych celów, być równoważne w sensie matematycznym, lecz obliczeniowo nie są. Istniejące algorytmy, które wyznaczają rozkład osobliwy są numerycznie bardziej stabilne od tych, które wyznaczają system własny macierzy $\boldsymbol{X}^T\boldsymbol{X}$.

4.2. Warunkowe indeksy

Z naszych rozważań wynika, że każda bliska zależność liniowa będzie miała niedużą wartość osobliwą macierzy $\boldsymbol{X}$ (lub wartość własną macierzy $\boldsymbol{X}^T\boldsymbol{X}$). Lecz wciąż nie wiemy jak niska musi być wartość osobliwa, aby wskazywała na bliską zależność liniową. Stopień złego uwarunkowania zależy od tego, jak niska jest minimalna wartość osobliwa w stosunku do maksymalnej wartości osobliwej. Pomocny tutaj okazuje się zbiór indeksów warunkowych, który określa, dla każdej relacji liniowej, stosunek niskiej wartości osobliwej do tej maksymalnej. Zbiór indeksów warunkowych dla macierzy $\boldsymbol{X}$ przedstawia się następująco:

$$\eta_k = \frac{\mu_{max}}{\mu_k}, \qquad k = 1, \dots, p, \tag{6}$$

dla wszystkich k, $\eta_k \geq 1$. Największa wartość η_k jest również indeksem warunkowym macierzy $\boldsymbol{X}$. Zatem możemy powiedzieć, że mamy tak wiele bliskich zależności wśród kolumn macierzy $\boldsymbol{X}$, jak wiele jest wysokich indeksów warunkowych. Zatem, słabe zależności są związane z indeksami warunkowymi na poziomie 5-10, a silne relacje liniowe wiążą się z indeksami warunkowymi na poziomie 30-100. Ponadto, jednoczesne występowanie kilku dużych wartości η_k wskazuje na obecność większej niż jedna ilości bliskich zależności liniowych. Zatem warunkowe indeksy są miarą diagnozującą współliniowość oraz określają jak dużo tych relacji jest.

4.3. Rozkład wariancji

Dobrze znaną, ale nie bezpośrednią, konsekwencją występowania współliniowości w modelu regresji jest to, że wariancje jej oszacowanych współczynników dążą do nieskończoności, przy estymacji metodą najmniejszych kwadratów. Ten fakt wykorzystamy przy określeniu, które zmienne są zaangażowane w relacje współliniowe. Pokażemy jak można rozszerzyć powyższą analizę, zapewniając rozkład wariancji każdego oszacowanego współczynnika regresji w sumę, która związana jest z indeksami warunkowymi każdej zmiennej. Duża wariancja występuje dla każdej zmiennej zaangażowanej w bliską liniową niezależność, a to zapewnia miarę dla diagnostyki zmiennych zaangażowanych w oddzielne współistniejące bliskie zależności. Rozkład wariancji jest powiązany z numeryczną analizą macierzy $\boldsymbol{X}$, która zawarta jest w rozkładzie na wartości osobliwe macierzy $\boldsymbol{X}$.

Przy naszych założeniach macierz kowariancji-wariancji $\boldsymbol{V}(\boldsymbol{b})$ estymatorów wyznaczonych metodą najmniejszych kwadratów $\boldsymbol{b}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ ma postać:

$$\boldsymbol{V}(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \tag{7}$$

gdzie σ^2 jest oczywiście wariancją składników losowych $\boldsymbol{\epsilon}$ liniowego modelu $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\epsilon}$.

Wykorzystując rozkład na wartości osobliwe $\boldsymbol{X}=\boldsymbol{UDV}^T$, powyższą macierz kowariancji-wariacji (7), możemy zapisać następująco:

$$\boldsymbol{V(b)} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T. \tag{8}$$

Więc wariancja współczynnika, b_k, który jest -tym diagonalnym elementem powyższego wyrażenia, ma postać:

$$var(b_k) = \sigma^2 \sum_j \frac{v_{kj}^2}{\mu_j^2}, \tag{9}$$

gdzie v_{ij} jest poszczególnym elementem macierzy $\boldsymbol{V}$, a μ_j, jest wartością osobliwą macierzy $\boldsymbol{X}$ Z powyższego wyrażenia widać, że $var(b_k)$ jest sumą składników. Każdy składnik jest związany z jedną i tylko jedną (z różnych) wartością osobliwą μ_j macierzy **X** Pojawienie się μ_j^2 w mianowniku powoduje, że składniki związane ze współliniowością będą duże w stosunku do innych. To zaś sugeruje, że niezwykle wysoka proporcja wariancji dwóch lub więcej współczynników, koncentruje się w składnikach zawiązanych z tymi samymi małymi wartościami osobliwymi. To zapewnia dowód, że zmienne, odpowiadające tym współczynnikom, są zaangażowane w bliską liniową zależność, odpowiadającą małym wartościom własnym. Wyraźmy powyższe rozumowanie w zapisie matematycznym. Zdefiniujmy zatem (k,j)-tą proporcję rozkładu wariancji, która opisuje stosunek wariancji -tego regresyjnego współczynnika do j- tego składnika jej rozkładu z równania (7). Wtedy:

$$\phi_{kj} = \frac{v_{kj}^2}{\mu_j^2} \quad oraz \quad \phi_k = \sum_{j=1}^{p} \phi_{kj}\,, \qquad k = 1,..,p. \tag{10}$$

Zatem proporcja rozkładu wariancji jest równa

$$\pi_{jk} = \frac{\phi_{kj}}{\phi_k}, \qquad k,j = 1,\dots,p. \tag{11}$$

Wyraźniej widać wyniki powyższej proporcji, w zestawieniu ich w macierz **Π**. W tej macierzy, każdy wiersz związany jest z warunkowym indeksem η_j. Wiersze te można ustawić tak, żeby warunkowe indeksy były rosnące (lub malejące). Naturalnie, kolumny macierzy **Π** powinny się sumować do wartości 1.

Tabela 1. Rozkład wariancji - macierz

Warunkowy Indeks ηk	$var(b1)$	$var(b2)$	...	$var(bp)$
$\eta 1$	$\pi 11$	$\pi 12$	...	$\pi 1p$
η 2	$\pi 21$	π 22	...	$\pi 2p$
...	⋮	⋮	⋱	⋮
ηp	$\pi p1$	$\pi p2$	...	πpp

Źródło: [3, 4]

4.4. Skalowanie macierzy $\boldsymbol{X}$

Aby nasze rozważania były w całości poprawne należy zrobić jeszcze jeden mały krok, który w zasadzie powinno się wykonać na samym początku. Początkowe dane, które tworzą macierz $\boldsymbol{X}$ różnią się od siebie skalą przydzieloną do każdej kolumny. Mimo tego reprezentują zasadniczo równoważne informacje, nie jest ważne to, czy dane są podane, np. w złotówkach czy w milionach złotych albo w metrach, kilometrach czy stopach długości. Jednak takie skale zmian mają wpływ na numeryczne właściwości macierzy, skutkując bardzo różnymi wartościami rozkładu proporcji wariancji oraz warunkowych indeksów. Natomiast nie wpływają na występowanie współliniowości w macierzy $\boldsymbol{X}$, ponieważ dla każdej nieosobliwej macierzy $\boldsymbol{B}$, istnieje niezerowy element $\boldsymbol{c}$, taki że $\boldsymbol{Xc} = \boldsymbol{0}$ wtedy i tylko wtedy, gdy $[\boldsymbol{XB}][\boldsymbol{B}^{-1}\boldsymbol{c}] = \bar{\mathbf{X}}\bar{\mathbf{c}} = \boldsymbol{0}$, gdzie $\bar{\mathbf{X}} = \boldsymbol{XB}$ oraz $\bar{\mathbf{c}} = \boldsymbol{B}^{-1}\boldsymbol{c}$. W takiej sytuacji, warunkowe indeksy mogą zapewnić niestabilne informacje, dla użytkownika liniowej regresji, o stopniu współliniowości wśród kolumn macierzy $\boldsymbol{X}$. Dlatego, w tym przypadku, należy unormować macierz danych do równoważnego modelu strukturalnego. W tym celu, wykonamy skalowanie na macierzy $\boldsymbol{X}$, które zrównoważy nam długość każdej jej kolumny do długości równej 1. Każde inne skalowanie nie odzwierciedla pożądanych właściwości.

Dalej kontynuujemy definicje rozkładu na wartości osobliwe macierzy $\boldsymbol{X}$, warunkowych indeksów oraz rozkładu wariancji, dla macierzy skalowanej. Zapiszmy matematycznie powyższe rozumowanie. Niech $\boldsymbol{X}$=[$\boldsymbol{X}_1$... $\boldsymbol{X}_p$],s_i=$(\mathbf{X}_i^T$ $\boldsymbol{X}_i)^{-\frac{1}{2}}$ oraz $\boldsymbol{S}$=diag(s_1,...,s_p) wtedy mamy:

$$\tilde{\eta}_i(\boldsymbol{X}) = \eta_i(\boldsymbol{XS}) \quad i = 1, \dots, p. \tag{12}$$

Takim samym sposobem możemy zdefiniować skalowaną proporcję rozkładu wariancji, gdy zamiast macierzy $\boldsymbol{X}$ użyjemy macierzy skalowanej$\boldsymbol{XS}$.

4.5. Algorytm postępowania w celu wykrywania relacji współliniowych

Dzięki powyższym rozważaniom, możemy rozwinąć diagnostyczną procedurę dla wykrycia jednej lub więcej obserwacji współliniowych, do znalezienia zmiennych, które są zaangażowane w każdą taką relację. Belsley (1991) sugeruje podążanie za dwoma następującymi warunkami:

- skalowane indeksy warunkowe o wysokiej wartości związane z
- wysoką skalowaną proporcją rozkładu wariancji dla dwóch lub więcej wariancji oszacowanych współczynników regresji.

Wartość skalowanych indeksów warunkowych, która jest wysoka, (powyżej 15-30), w pierwszym warunku, rozpoznaje liczbę relacji współliniowych wśród kolumn macierzy. $\boldsymbol{X}$ W drugim warunku, wysoka proporcja rozkładu wariancji (większa niż 0,5), związana z każdą wysoką wartością skalowanych indeksów warunkowych, rozpoznaje te zmienne, które są zaangażowane w każdą bliską zależność.

Należy zatem postępować według poniższego algorytmu.

Krok 1. Skalowanie macierzy $\boldsymbol{X}$. Należy kolumny macierzy $\boldsymbol{X}$ unormować do długości równej 1. Inne skalowanie może nie przynieść pożądanych efektów.

Krok 2. Uzyskanie rozkładu na wartości osobliwe macierzy oraz wyznaczenie:
warunkowych indeksów η_k
macierzy proporcji rozkładu wariancji **Π**.

Krok 3. Wskazanie relacji współliniowych oraz ich względnego natężenia. W tym celu używamy indeksów warunkowych, które przekraczają pewną wartość krytyczną. Wartość tą wybieramy w zależności od pożądanego efektu, może to być np. $\eta^* = 10$, $\eta^* = 15$ lub $\eta^* = 30$. Zależności, które mają równe indeksy warunkowe, przekraczające nieznacznie zadany próg, mają podobne niskie natężenie, tak samo jest z zależnościami, których warunkowe indeksy znacznie przekraczają zadany próg - mają silne natężenie współliniowości.

Krok 4. Określenie zmiennych zaangażowanych w liniowe zależności. W tym kroku należy ustalić próg dla proporcji rozkładu wariancji π^* (w praktyce najczęściej używa się $\pi^* = 0.5$)

5. Analiza diagnostyki współliniowości na przykładzie

Dane w tabeli 2 reprezentują kształtowanie się ceny lodówek w zależności od kilku ich cech. Rozważymy czy w danych tych występują jakieś współliniowe relacje.

Zmienne w powyższych danych są następujące:
$\boldsymbol{C}$ - cena lodówki
$\boldsymbol{K}$ - koszt rocznego utrzymania (koszt zużycia energii itp.)
$\boldsymbol{W}$ - pojemność lodówki (mierzona w stopach sześciennych)
$\boldsymbol{PZ}$ - pojemność komory zamrażarki (mierzona w stopach sześciennych)
$\boldsymbol{LP}$ - liczba półek w drzwiach lodówki
$\boldsymbol{PP}$ - pojemność półek (mierzona w stopach sześciennych)
$\boldsymbol{LF}$ - liczba funkcji lodówki

Dla powyższych danych, oszacowany model regresji prezentuje się następująco:

$$\begin{aligned}\widehat{\boldsymbol{C}} = -874.3571 - 6.1294\boldsymbol{K} + 65.06118\boldsymbol{W} \\ + 124.5467\boldsymbol{PZ} + 46.0744\boldsymbol{LP} \\ + 8.9159\boldsymbol{PP} + 23.40153\boldsymbol{LF}.\end{aligned} \tag{13}$$

Zbadajmy te zmienne pod względem współliniowości. W pierwszym kroku przeprowadzamy skalowanie macierzy zmiennych $\boldsymbol{X}$, tak aby jej kolumny miały długość równą 1. Postać macierzy unormowanej jest dana w tabeli 3. Mając daną unormowaną macierz $\boldsymbol{X}$ można teraz przejść do naszej procedury detekcji współliniowości w zaproponowanym modelu. Rozkład na wartości osobliwe oraz dalsze rachunki zostały wykonane w pakiecie matematycznym Maple 15. Przytoczymy tutaj tylko najistotniejsze wyniki. Jak wiemy, decyzję o istnieniu współliniowości zmiennych dokonujemy zawsze w stosunku do skalowanych warunkowych indeksów Zatem macierz $\boldsymbol{\Pi}$ jest podana w tabeli 4. Dla trzech największych wartości skalowanych indeksów warunkowych, występuje tylko jedna relacja współliniowości, lecz jest ona dość słaba. Dla indeksu warunkowego równego 59.9305 widać, że zmienna $\boldsymbol{LP}$ i $\boldsymbol{PP}$ mają między sobą słabą relację współliniowości. Zauważamy również, że zmienna $\boldsymbol{PP}$ jest nieistotna w modelu, co widać w wartościach danych dla testu istotności (14). To może sugerować, że zmienna $\boldsymbol{PP}$ jest zbędna w naszym modelu. Jednak, aby lepiej poznać jej naturę należałoby faktycznie sprawdzić jak wyglądają wskaźniki jakości modelu po jej odjęciu. Może okazać się, że usunięcie jej znacznie pogorszy jakość modelu. Dodatkowo relacja współliniowości jest dosyć słaba (wartości rozkładu wariancji są blisko wartości krytycznej równej 0,5, ale jednak jej nie przekraczają). Wskaźniki jakości modelu są zaprezentowane w tabeli 5. Widzimy, że po odjęciu zmiennej $\boldsymbol{PP}$ z modelu, współczynnik determinacji nam minimalnie zmalał, lecz nie jest to na tyle wielka zmiana, aby móc się tym martwić. Zaś błąd standardowy modelu, dany w trzeciej kolumnie, nieznacznie się zwiększył, co też nie jest zbyt ważne. Największą zmianę możemy dostrzec we wskaźniku wyrazistości modelu, który zmienił się na korzyść modelu bez zmiennej $\boldsymbol{PP}$ oraz w błędach oszacowań poszczególnych zmiennych. Też nie są to duże zmiany, jednak nieznacznie poprawiają model. Właśnie ta poprawa modelu oraz podejrzenie o współliniowość mogą być potwierdzeniem na to, że wyrzucenie zmiennej $\boldsymbol{PP}$ będzie słuszną decyzją.

Tabela 2. Kształtowanie się ceny lodówek w 1992 roku.

LP	C	K	W	PZ	LP	PP	LF
1	595	75	12,8	5,7	3	25,4	2
2	685	75	12,9	5,7	3	26,7	1
3	535	67	13,3	4,5	1	24	6
4	600	67	13,2	4,5	3	23,5	5
5	605	67	13,3	4,5	3	24	3
6	665	67	13,3	4,5	1	25,6	10
7	515	67	13,9	4,1	2	23	1
8	485	68	12,9	5,1	2	21,7	3
9	550	68	13,1	5,1	2	24,4	3
10	555	68	13,1	5,1	2	23,4	2
11	610	61	13,1	5,1	2	23,6	3
12	580	60	14,3	4,3	2	24,7	4
13	700	60	14,3	4,3	4	24,1	4

14	505	68	13,1	5,1	2	23,1	2
15	555	68	13,1	5,1	2	23,3	2
16	500	72	13,9	4,1	2	25,7	1
17	675	61	13,2	4,8	3	23,6	3
18	545	75	12,8	5,7	3	25,4	3
19	600	75	12,8	5,7	3	25,4	1
20	700	69	12,9	5,7	3	26,2	6
21	760	69	12,8	5,7	1	28,3	5
22	530	75	12,9	5,7	3	25,2	2
23	550	68	13,1	5,1	2	23,1	2
24	580	68	13,1	5,1	2	23,3	1
25	530	73	13,7	4,4	3	21,2	2
26	550	73	13,7	4,4	3	23,6	1
27	615	60	14,3	4,3	2	26,2	1
28	710	60	14,3	4,3	4	25,2	3
29	555	62	13,2	4,8	3	23,6	1
30	590	62	13,2	4,8	3	23,6	2
31	460	66	14,2	4,4	1	25,5	3
32	520	66	14,2	4,4	1	30,2	5
33	1200	81	12,6	7,3	5	24	12
34	745	90	12,9	6,8	5	20,6	6
35	800	90	12,9	6,8	5	20,6	7
36	840	94	14,7	7,4	1	28,3	5
37	880	94	14,7	7,4	1	28,3	5

Źródło danych: Consumer Reports, 1992, July. "*Refrigerators: A Comprehensive Guide to the Big White Box*". Dane te są dołączone do podręcznika E. Freesa [5]

Tabela 3. Skalowana macierz

0,164	0,17325	0,15744	0,17784	0,1797	0,169418	0,0772
0,164	0,17325	0,15867	0,17784	0,1797	0,178089	0,0386
0,164	0,15477	0,16359	0,1404	0,0599	0,16008	0,2316
0,164	0,15477	0,16236	0,1404	0,1797	0,156745	0,193
0,164	0,15477	0,16359	0,1404	0,1797	0,16008	0,1158
0,164	0,15477	0,16359	0,1404	0,0599	0,170752	0,386
0,164	0,15477	0,17097	0,12792	0,1198	0,15341	0,0386
0,164	0,15708	0,15867	0,15912	0,1198	0,144739	0,1158
0,164	0,15708	0,16113	0,15912	0,1198	0,162748	0,1158
0,164	0,15708	0,16113	0,15912	0,1198	0,156078	0,0772
0,164	0,14091	0,16113	0,15912	0,1198	0,157412	0,1158
0,164	0,1386	0,17589	0,13416	0,1198	0,164749	0,1544
0,164	0,1386	0,17589	0,13416	0,2396	0,160747	0,1544
0,164	0,15708	0,16113	0,15912	0,1198	0,154077	0,0772
0,164	0,15708	0,16113	0,15912	0,1198	0,155411	0,0772
0,164	0,16632	0,17097	0,12792	0,1198	0,171419	0,0386

0,164	0,14091	0,16236	0,14976	0,1797	0,157412	0,1158
0,164	0,17325	0,15744	0,17784	0,1797	0,169418	0,1158
0,164	0,17325	0,15744	0,17784	0,1797	0,169418	0,0386
0,164	0,15939	0,15867	0,17784	0,1797	0,174754	0,2316
0,164	0,15939	0,15744	0,17784	0,0599	0,188761	0,193
0,164	0,17325	0,15867	0,17784	0,1797	0,168084	0,0772
0,164	0,15708	0,16113	0,15912	0,1198	0,154077	0,0772
0,164	0,15708	0,16113	0,15912	0,1198	0,155411	0,0386
0,164	0,16863	0,16851	0,13728	0,1797	0,141404	0,0772
0,164	0,16863	0,16851	0,13728	0,1797	0,157412	0,0386
0,164	0,1386	0,17589	0,13416	0,1198	0,174754	0,0386
0,164	0,1386	0,17589	0,13416	0,2396	0,168084	0,1158
0,164	0,14322	0,16236	0,14976	0,1797	0,157412	0,0386
0,164	0,14322	0,16236	0,14976	0,1797	0,157412	0,0772
0,164	0,15246	0,17466	0,13728	0,0599	0,170085	0,1158
0,164	0,15246	0,17466	0,13728	0,0599	0,201434	0,193
0,164	0,18711	0,15498	0,22776	0,2995	0,16008	0,4632
0,164	0,2079	0,15867	0,21216	0,2995	0,137402	0,2316
0,164	0,2079	0,15867	0,21216	0,2995	0,137402	0,2702
0,164	0,21714	0,18081	0,23088	0,0599	0,188761	0,193
0,164	0,21714	0,18081	0,23088	0,0599	0,188761	0,193

Źródło: Opracowanie własne

Tabela 4. Zestawienie warunkowych indeksów oraz macierzy rozkładu proporcji

$\widetilde{\eta_i}$	$\widetilde{\pi_{ij}}$						
1.0	0.00816	0.00916	0.00795	0.07841	0.04767	0.84862	2.2716
4.80980	0.00448	0.00198	0.00285	0.08422	0.43250	0.47394	0.07742
6.76093	0.01739	0.02127	0.02897	0.25531	0.39538	0.28167	5.8197
16.6643	0.00628	0.00061	0.00189	0.35386	0.25308	0.38427	0.10681
40.522	0.02084	0.00640	0.88233	0.02566	0.02206	0.04269	0.00365
59.9305	0.00437	0.00421	0.01555	0.03945	0.45292	0.48348	0.01757
97.9095	0.02547	0.92683	0.00392	0.03912	0.00462	0.00002	0.00036

Źródło: Opracowanie własne

Tabela 5. Wskaźniki jakości modelu regresji

	R^2	S_z	V	S_{b_i}
Dla modelu pełnego	0.848	59.7	9,54%	S_{b_0} = 272.7 S_{b_1} = 2.31 S_{b_2} = 20.8 S_{b_3} = 25.14 S_{b_4} = 11.3 S_{b_5} = 6.41 S_{b_6} = 4.44
Dla modelu bez zmiennej **PP**	0.839	60.7	9,70%	S_{b_0} = 271.63 S_{b_1} = 2.277 S_{b_2} = 19.457 S_{b_3} = 23.782 S_{b_4} = 9.894 S_{b_6} = 4.515

Źródło: Opracowanie własne

6. Posumowanie i końcowe wnioski

W pracy tej zostały wymienione oraz dokładnie opisane popularne metody wykrywania współliniowości wraz ze wskazaniem ich wad. Wykazano, że źródłem tego typu obserwacji jest źle uwarunkowana macierz danych. Złe uwarunkowanie dotyczy numerycznej lub geometrycznej charakterystyki tej macierzy – takie obserwacje są problemem danych, a nie problemem statystycznym. Głównym wnioskiem płynącym z tej pracy jest to, że wysoka korelacja jest wystarczająca do zdiagnozowania relacji współliniowych, ale nie jest konieczna.

Literatura

1. Bartosiewicz J. *Wykłady ze statystyki matematycznej*, PWN, Warszawa 1989.
2. Pawłowski Z. *Statystyka matematyczna*, PWN, Warszawa 1976.
3. Belsley D. A. *Conditioning diagnostics. Collinearity and Weak Data in Regression*, John Wiley &Sons, New York 1991.
4. Belsley D. A., Kuh E., Roy W. E, *Regression Diagnostics*: Identifying Influential Data and Sources of Collinearity, John Wiley &Sons, New York 1980.
5. Frees E. W. *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press, New York 2010.
6. Green W. *Econometric Analysis*, Upper Saddle River, New Jersey 2003.
7. Rao C.R, *Modele liniowe statystyki matematycznej*, PWN, Warszawa 1982.

Diagnostyka współliniowości w modelu regresji liniowej

Streszczenie

W pracy przedstawiono etapy diagnozowania współliniowości, mogącej pojawiać się w modelu regresji liniowej oraz konsekwencje jej występowania. Przeanalizowane zostały miary wykrywające obecność jednej lub więcej współliniowych relacji wśród kolumn macierzy **X**, rozważone zostały miary identyfikujące podzbiór zmiennych objaśniających, zaangażowanych w każdą taką relację. Omówiona metoda jest coraz szerzej stosowana oraz różni się od innych lepszą efektywnością i skutecznością. W pracy tej zostały również wymienione oraz dokładnie opisane popularne metody wykrywania współliniowości wraz ze wskazaniem ich wad. Wykazano, że źródłem tego typu obserwacji jest źle uwarunkowana macierz danych. Złe uwarunkowanie dotyczy numerycznej lub geometrycznej charakterystyki tej macierzy – takie obserwacje są problemem danych, a nie problemem statystycznym. Głównym wnioskiem płynącym z tej pracy jest to, że wysoka korelacja jest wystarczająca do zdiagnozowania relacji współliniowych, ale nie jest konieczna.

Słowa kluczowe: regresja liniowa, współliniowość zmiennych objaśniających, zbiór danych, zmienne objaśniające.

Collinearity diagnostics in a linear regression model

Abstract

In this paper, we have presented diagnostic of collinearity stage, that can appear in linear regression model and consequences its occurrence. We have analyzed measures detecting the presence of one or more collinear relationships among the columns of the matrix **X**, measures indentifying explanatory variables subset involved in each such relations has been considered. Discussed method is increasingly used and it differs from others having better effectiveness and efficiency. In this paper popular methods of detecting collinearity with the indication of its weaknesses have been mentioned and exactly described. It has been showed that the source of the collinear relations among the column of the matrix **X** is an ill-conditioned data matrix. Such ill-conditioning concerns numerical or geometric characteristic of the data matrix and it is a data problem not a statistical one. Main conclusion of this work is that high correlation is sufficient to diagnose collinear relationships, but isn't necessary.

Keywords: linear regression, collinearity of explanatory variables, data set, explanatory varialbes

Izabela A. Domagalska [1], Ewa A. Drzazga[2], Anna H. Kosiacka [3]

Równania Eliashberga i wzory semi-analityczne w opisie właściwości termodynamicznych stanu nadprzewodzącego indukującego się w wanadzie

1. Wprowadzenie

Nadprzewodnictwo to zjawisko elektryczne, które jest związane z zanikiem oporu w temperaturze niższej niż charakterystyczna dla danego materiału temperatura krytyczna. Jednocześnie, jeśli pod uwagę weźmie się zanik pola magnetycznego wewnątrz materiału będącego w stanie nadprzewodzącym, nadprzewodnictwo to zjawisko magnetyczne. Charakterystyczna zerowa wewnętrzna indukcja magnetyczna nazywana jest efektem Meissnera-Ochsenfelda.

Dobre nadprzewodniki powinny charakteryzować się czterema cechami, którymi są: jak najwyższa gęstość prądu, tania technologia wytwarzania, możliwie duże pole krytyczne, jednak najważniejszą z nich jest osiąganie jak najwyższej temperatury krytycznej.

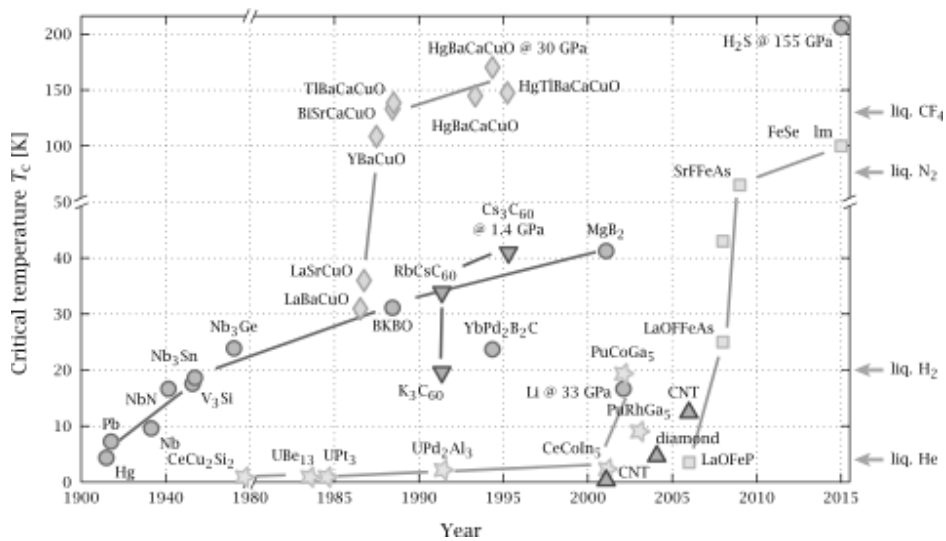
Stan nadprzewodzący jest przedmiotem badań naukowców od ponad stu lat. Jako pierwszy gwałtowny zanik oporu elektrycznego materiału zaobserwował holenderski fizyk Heike Kamerlingh-Onnes. W 1908 w Lejdzie badacz ten sprawdzał przewodnictwo rtęci w temperaturze ciekłego helu. Zaobserwował on gwałtowny spadek oporności w próbce, którą schłodzono poniżej temperatury $4{,}2\ K$, a za swoje badania został uhonorowany nagrodą Nobla.

Nadprzewodnictwo stanowi znaczącą wartość dla możliwości rozwoju techniki, jak również oszczędności energii. Obecnie nadprzewodniki cieszą się dużą popularnością w niektórych rozwiązaniach technologicznych, między innymi w medycznych urządzeniach diagnostycznych. Warto w tym miejscu wspomnieć o czułym detektorze pola magnetycznego, czyli nadprzewodnikowym interferometrze kwantowym (SQUID – ang. superconducting quantum interference device). Wysoka czułość tego elementu pozwala przeprowadzać nieinwazyjne badania mózgu, ze względu na możliwość wykrycia pola magnetycznego generowanego przez ten ludzki organ. Wcześniejszym a nadal aktualnym zastosowaniem nadprzewodników w medycynie było użycie ich w rezonansie magnetycznym, gdzie nadprzewodzące magnesy pozwalają stworzyć obrazy, które umożliwiają dokładne i nieinwazyjne sprawdzenie ciała pacjenta. Nie są

[1] iza.domagalska@wp.pl, Studenckie Koło Naukowe QUBIT, Instytut Fizyki, Wydział Inżynierii Produkcji i Technologii Materiałów, Politechnika Częstochowska, www.fizyka.wip.pcz.pl
[2] drzazga.ewa@wip.pcz.pl, Instytut Fizyki, Wydział Inżynierii Produkcji i Technologii Materiałów, Politechnika Częstochowska, www.fizyka.wip.pcz.pl
[3] akosiacka@vp.pl, Instytut Fizyki, Wydział Inżynierii Produkcji i Technologii Materiałów, Politechnika Częstochowska, www.fizyka.wip.pcz.pl

to jedyne zastosowania nadprzewodników. Materiały te używa się również między innymi w transformatorach nadprzewodnikowych, przy konstruowaniu kriokabli oraz w kolei magnetycznej.

Nadrzędnym celem, do którego dążą badacze zajmujący się stanem nadprzewodzącym, jest znalezienie materiału, którego temperatura krytyczna byłaby zbliżona do wartości pokojowej. Historycznie po pierwszych odkryciach nadprzewodników niskotemperaturowych stwierdzono, że temperatura krytyczna nie przekroczy $30\ K$. Dopiero w 1986 roku Georg Bednorz oraz Alex Müller odkryli wysokotemperaturowe nadprzewodnictwo w miedzianach. Podczas badań ceramiki o strukturze perowskitu zmierzono T_C równe 35 K. W kolejnych latach badano wiele związków chemicznych oraz pierwiastków. Zaczęto sprawdzać, jaką temperaturę krytyczną będą miały materiały, które znajdowałyby się pod działaniem wysokiego ciśnienia. Zwrócono uwagę na nadprzewodnictwo, które jest indukowane oddziaływaniem elektron-fonon, jak również na związki wodorowane. Nową granicą dla nadprzewodnictwa było $200\ K$. Dopiero kilka lat temu dokonano w tej dziedzinie kolejnego skoku jakościowego. W 2015 roku grupa badaczy odkryła bowiem takie związki chemiczne, których wartość temperatury krytycznej przewyższa $200\ K$. Problemem jest jednak ciśnienie, pod jakim związki H_2S i H_3S przechodzą w stan nadprzewodzący, bowiem sięga ono 150 GPa [1,2].



Rysunek 1. Chronologia odkryć związków nadprzewodzących [3]

Kolejnym problemem, jaki pojawia się podczas badania wysokotemperaturowego stanu nadprzewodzącego, jest brak odpowiednich teorii, które opisywałyby sposób indukowania się stanu nadprzewodzącego oraz jego właściwości. Stąd też liczne grono fizyków teoretyków pracuje nad dopasowaniem i modyfikacją znanych formalizmów tak, aby odpowiadały otrzymanym wartościom eksperymentalnym.

W pracy skupiono się na badaniu właściwości termodynamicznych stanu nadprzewodzącego indukującego się w wanadzie przy użyciu dwóch podejść – formalizmu Eliashberga oraz wzorów semi-analitycznych. Uwagę poświęcono temu pierwiastkowi, ponieważ należy on do grupy metali przejściowych. Chociaż wanad jest nadprzewodnikiem niskotemperaturowym a w dodatku nie cechuje się najwyższą temperaturą krytyczną pośród pierwiastków w bloku d, jest on przedmiotem badań pod kątem nadprzewodnictwa od 1934 roku. Wtedy to nagły spadek rezystancji tego pierwiastka do niemierzalnie małych wartości w temperaturze $4{,}3\ K$ odkryli Hans Westerhoff oraz Walter Meissner. W roku 1996 dokładną analizę stanu nadprzewodzącego w wanadzie przy pomocy metody *ab initio* przeprowadził Sergej Y. Savrasov wraz z innymi autorami [4]. Szczegółowe badania teoretyczne sugerowały stosunkowo wysoką wartość stałej sprzężenia elektron-fonon ($\lambda = 1{,}19$). Korzystając z tej wartości na podstawie wzoru Wiliama M. McMillana uzyskano temperaturę krytyczną na poziomie $6{,}68\ K$. We wspomnianej pracy arbitralnie założono istnienie bardzo silnych efektów deparujących, dla których μ^* utrzymywało się na poziomie $0{,}3$. W 2010 roku natomiast w eksperymencie tunelowym Shawn Westerdale zmierzył T_C dla wanadu równe $5{,}38\ K$ [5]. Istotne wyniki otrzymano, gdy ten pierwiastek bloku d poddano badaniom pod wysokim ciśnieniem. W 2000 roku troje badaczy: M. Ishizuka, M. Iketani oraz S. Endo przeprowadzili badania w zakresie ciśnień do $120\ GPa$. Wanad wykazał w nich silny liniowy wzrost temperatury krytycznej wraz ze zwiększaniem się ciśnienia a maksymalna zmierzona wartość T_C kształtowała się na poziomie $16{,}5\ K$. Nie były to jedyne ciśnieniowe badania tego pierwiastka. Również w 2000 roku Ganapathy Vaitheeswaran wraz z innymi autorami otrzymał pod ciśnieniem $130\ GPa$ jak odtąd najwyższą temperaturę krytyczną dla wanadu, która wynosiła $25\ K$.

2. Cel pracy

Celem niniejszej publikacji jest porównanie wyników, które opisują właściwości termodynamiczne stanu nadprzewodzącego, otrzymanych w ramach formalizmu równań Eliashberga oraz przy użyciu wzorów semi-analitycznych. W pracy przedstawiono szczegółowe rezultaty analizy numerycznej przeprowadzonej dla wanadu, który jest pierwiastkiem grupy przejściowej.

3. Opis stanu nadprzewodzącego

W pracy przedstawione zostaną wyniki otrzymane w ramach formalizmu równań Eliashberga, który jest uogólnieniem teorii BCS, oraz wzorów semi-analitycznych. W niniejszym rozdziale opisane zostaną ogólne założenia wspomnianych teorii.

3.1. Podstawy mikroskopowej teorii BCS

Mikroskopowa teoria nadprzewodnictwa, której autorami byli John Bardeen, Leon Cooper oraz Robert Schriffer okazała się przełomem w jakościowym opisie stanu nadprzewodzącego. Stanowi ona jedno z ważniejszych osiągnięć teoretycznych w fizyce ciała stałego. Nazwa teorii pochodzi od inicjałów naukowców, którzy ją stworzyli. Za opracowanie jej Bardeen, Cooper oraz Schriffer otrzymali Nagrodę Nobla w dziedzinie fizyki w 1972 roku.

Podstawą teorii BCS jest założenie, iż fermiony mogą łączyć się w pary Coopera a skondensowane pary są w stanie bezoporowo poruszać się wewnątrz nadprzewodnika. Opis całej teorii rozpoczyna się od zdefiniowania hamiltonianu, który modeluje potencjał parujący i energię stanu elektronowego. Operator Hamiltona jest wysoce złożony matematycznie, dlatego do opisu wykorzystuje się jego uproszczoną wersję, w której uwzględnia się przybliżenie średniego pola. W kolejnych krokach oblicza się parametr porządku Δ oraz wyznacza pozostałe wielkości w tym stosunki bezwymiarowe R_Δ, R_H oraz R_C o których więcej informacji przedstawionych zostanie w kolejnych częściach niniejszej pracy.

Warto jednak zwrócić uwagę na fakt, że teoria ta pozwala na jakościowe wyznaczenie parametrów termodynamicznych. Dodatkowo, badać można jedynie takie nadprzewodniki, które mają elektronowo-fononowy mechanizm parowania. Kolejnym obostrzeniem jest to, iż stała sprzężenia elektron-fonon nie powinna być wyższa niż 0,3.

3.2. Zarys formalizmu równań Eliashberga

Formalizm równań Eliashberga, który wykorzystywany jest w badaniach stanu nadprzewodzącego między innymi w wanadzie, to uogólnienie teorii Bardeena, Coopera oraz Schriffera. Formalizm Eliashberga uogólnia teorię BCS ze względu na uwzględnienie efektów retardacyjnych oraz silnosprzężeniowych. Analizę teorii Eliashberga należy rozpocząć od zapisania operatora pola Frölicha [6], czyli hamiltonianu elektronowo-fononowego w postaci:

$$H = \sum_{k\sigma} \bar{\varepsilon}_k \, c^+_{k\sigma} c_{k\sigma} + \sum_q \omega_q b^+_q b_q + \sum_{kq\sigma} g_{k,k+q} c^+_{k+q\sigma} c_{k\sigma} \left(b^+_{-q} + b_q\right), \quad (1)$$

gdzie $\bar{\varepsilon}_k$ opisuje energię stanu elektronowego, zdefiniowane przez różnicę: $\bar{\varepsilon}_k = \varepsilon_k - \mu$, gdzie μ to potencjał chemiczny. Symbol $c^{(+)}_{k\sigma}$ oznacza operator anihilacji (kreacji) elektronu w stanie Blocha o pędzie k i spinie σ, $b^{(+)}_q$ reprezentuje operator anihilacji (kreacji) fononu o pędzie q, $g_{k,k+q}$ opisuje elementy macierzowe oddziaływania elektron-fonon, natomiast ω_q - fononowa relacja dyspersyjna.

By móc wyprowadzić równania Eliashberga należy operator Hamiltona (1) zapisać w notacji macierzowej, używając do tego celu spinorów Nambu. W kolejnym kroku trzeba zdefiniować macierzową funkcję Greena typu Matsubary, jak również wyprowadzić macierzowe równanie Dysona [7,8]. Po przeprowadzenia samouzgodnienia do energii własnej, które szczegółowo przedstawione zostało w pracy Julesa Carbotte [9] uzyskuje się układ Eliashberga.

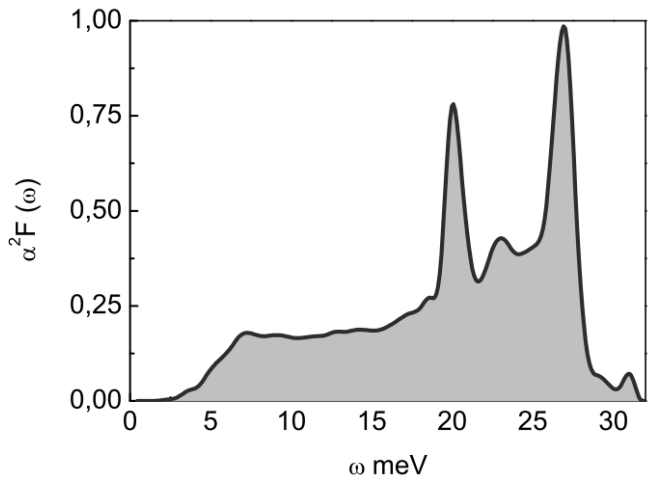
Na poziomie ilościowym analizę stanu nadprzewodzącego przeprowadza się korzystając z równań Eliashberga określanych na osi urojonej:

$$\phi_n = \frac{\pi}{\beta} \sum_{m=-M}^{M} \frac{\lambda(i\omega_n - i\omega_m) - \mu^* \theta(\omega_c - |\omega_m|)}{\sqrt{\omega_m^2 Z_m^2 + \phi_m^2}} \phi_m, \quad (2)$$

$$Z_n = 1 + \frac{1}{\omega_n} \frac{\pi}{\beta} \sum_{m=-M}^{M} \frac{\lambda(i\omega_n - i\omega_m)}{\sqrt{\omega_m^2 Z_m^2 + \phi_m^2}} \omega_m Z_m, \quad (3)$$

gdzie parametr porządku określa wzór $\Delta_n = \phi_n / Z_n$, λ to stała sprzężenia elektron-fonon, μ^* - pseudopotencjał kulombowski a θ reprezentuje funkcję Heaviside'a.

Funkcję Eliashberga, która potrzebna jest do przeprowadzenia wszystkich obliczeń, można otrzymać przy użyciu danych otrzymanych w ramach eksperymentu tunelowego. Znacznie wyższym skomplikowaniem cechuje się obliczanie tej funkcji z pierwszych zasad. Funkcja Eliashberga, która posłużyła do analizy numerycznej stanu nadprzewodzącego w wanadzie, została wyznaczona metodą DFTP (Density Functional Perturbation Theory). Funkcja $\alpha^2 F(\omega)$ pochodzi z pracy [10] a jej dokładny przebieg można prześledzić na rysunku 2.



Rysunek 2. Przebieg funkcji Eliashberga dla wanadu [opracowanie własne]

3.3. Wzory semi-analityczne

W literaturze obok wysoce skomplikowanego matematycznie formalizmu równań Eliashberga istnieją wzory, pozwalające obliczyć temperaturę krytyczną (T_C), różnicę energii swobodnej (ΔF), ciepło właściwe stanu nadprzewodzącego $\left(\mathrm{C^S}\right)$ i termo dynamiczne pole krytyczne $(\mathrm{H_C})$. Dodatkowo występują również zmodyfikowane formuły, przy pomocy których można sprawdzić wartości bezwymiarowych parametrów termodynamicznych $(\mathrm{R}_\Delta, R_C$ oraz $R_H)$ oraz porównać otrzymane wyniki do wartości określonych jako stałe w ramach formalizmu BCS.

Jedną z najczęściej używanych formuł, która służy do obliczenia temperatury krytycznej, jest wzór Philipa B. Allena i Roberta C. Dynesa. W pracy [11] naukowcy ci zaproponowali modyfikację znanego wcześniej wzoru Wiliama M. McMillana, który jest jedną z najprostszych formuł umożliwiających oszacowanie temperatury krytycznej. Zależność na temperaturę krytyczną według McMillana przedstawiono poniżej:

$$T_c = \frac{\omega_{ln}}{1{,}20} exp\left[\frac{-1{,}04(1+\lambda)}{\lambda-\mu^*(1+0{,}62\lambda)}\right], \quad (4)$$

Gdzie:

$$\omega_{ln} \equiv exp\left[\frac{2}{\lambda}\int_0^{+\infty} d\Omega \frac{\alpha^2 F(\Omega)}{\Omega} ln(\Omega)\right] \quad (5)$$

Uwzględniając poprawki Allena-Dynesa wzór pozwalający obliczyć T_C prezentuje się następująco:

$$k_B T_c = f_1 f_2 \frac{\omega_{ln}}{1{,}2} exp\left[\frac{-1{,}04(1+\lambda)}{\lambda-\mu^*(1+0{,}62\lambda)}\right], \quad (6)$$

gdzie ω_{ln} to fononowa częstość logarytmiczna, λ to stała sprzężenia elektron-fonon, natomiast $f_1 f_2$ to funkcje korekcji kolejno silnego sprzężenia oraz kształtu.

Kolejną formułą semi-analityczną, która opisuje właściwości termodynamiczne stanu nadprzewodzącego jest wzór na różnicę energii swobodnej między stanami nadprzewodzącym i normalnym [12]:

$$\Delta F = F^{(S)} - F^{(N)} = -2\pi k_B T \sum_{n=1}^{M} \sum_{\alpha} \rho_{\alpha}(0) \left[\sqrt{\omega_n^2 + (\Delta_n^{\alpha})^2} - |\omega_n|\right] \left[Z_n^{\alpha,(S)} - \right.$$

$$\left. Zna,N\omega n\omega n2+\Delta n\alpha 2\right]. \quad (7)$$

Aby obliczyć termodynamiczne pole krytyczne, można skorzystać z następującej zależności:

$$H_C = \sqrt{-8\pi\Delta F} \quad (8)$$

Do otrzymania wartości ciepła właściwego stanu nadprzewodzącego można wykorzystać poniższy wzór, który opisuje skok ciepła właściwego:

$$\Delta C = C^S - C^N = -\frac{d^2\Delta F}{dT^2}. \quad (9)$$

Wartości bezwymiarowych parametrów termodynamicznych można obliczyć na podstawie wzorów, które zaproponowali kolejno: Bozidar Mitrović, H. G. Zarate i J. P. Carbotte [12], Frank Marsiglio i J. P. Carbotte [13] oraz Jules P. Carbotte [9]. Warto zwrócić uwagę, iż poniższe zależności są zmodyfikowanymi wzorami wynikającymi z teorii BCS:

$$R_{\Delta} = \frac{2\Delta(0)}{k_B T_c} = 3{,}53\left[1 + 12{,}5\left(\frac{k_B T_c}{\omega_{ln}}\right)^2 ln\left(\frac{\omega_{ln}}{2k_B T_c}\right)\right], \quad (10)$$

$$R_c = \frac{\Delta C(T_c)}{C^N(T_c)} = 1{,}43\left[1 + 53\left(\frac{k_B T_c}{\omega_{ln}}\right)^2 ln\left(\frac{\omega_{ln}}{3k_B T_c}\right)\right], \quad (11)$$

$$R_H = \frac{T_c C^N(T_c)}{H_c^2(0)} = 0{,}168\left[1 - 12{,}2\left(\frac{k_B T_c}{\omega_{ln}}\right)^2 ln\left(\frac{\omega_{ln}}{3k_B T_c}\right)\right], \quad (12)$$

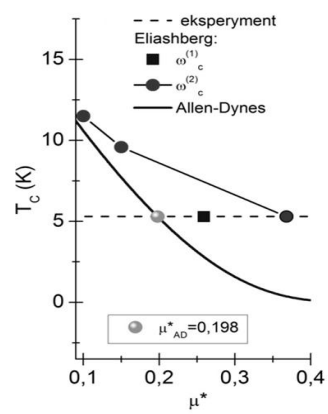
gdzie $\Delta C(T_C) = C^S(T_C) - C^N(T_C)$ jest skokiem ciepła właściwego w temperaturze krytycznej, C^N - ciepło właściwe stanu normalnego.

4. Termodynamika stanu nadprzewodzącego

Analiza numeryczna termodynamiki stanu nadprzewodzącego w wanadzie polegała na określeniu pseudopotencjału kulombowskiego oraz sprawdzeniu zależności temperatury krytycznej od μ^*. W kolejnych krokach zbadano bezwymiarowe parametry termodynamiczne oraz określono właściwości termodynamiczne. Szczególną uwagę zwrócono na porównanie wyników uzyskanych w ramach formalizmu równań Eliashberga oraz przy pomocy wzorów semi-analitycznych.

4.1. Pseudopotencjał kulombowski oraz temperatura krytyczna

Badania wybranych parametrów stanu nadprzewodzącego w wanadzie rozpoczęto od określenia odpowiedniej wartości pseudopotencjału kulombowskiego. Parametr ten dopasowano tak, aby pozwalał on na jak najdokładniejsze odzwierciedlenie wartości temperatury krytycznej, która wynika z badań eksperymentalnych ($T_C = 5{,}3\ K$) [15]. Po przeprowadzeniu odpowiedniej analizy numerycznej stwierdzono, że μ^* silnie zależy od częstości odcięcia (ω_c). Ze względu na to, że w literaturze przyjmuje się $\omega_c \in < 3\Omega_{max}, 10\Omega_{max} >$, gdzie ω_c to częstość odcięcia, natomiast $\Omega_{\max}$ opisuje maksymalną częstość drgań fononów, obliczono pseudopotencjał kulombowski dla skrajnych wartości tego przedziału. Dla niższej częstości odcięcia $\omega_c^{(1)} = 3\Omega_{max}$ otrzymano $\mu^*\left(\omega_c^{(1)}\right) = 0{,}259$, co już uznaje się za wysoką wartość. Przy uwzględnieniu najwyższej częstości odcięcia równiej dziesięciokrotności maksymalnej częstości drgań fononowych, wartość pseudopotencjału kulombowskiego również wzrasta i wynosi $0{,}368$. W kolejnym kroku korzystając z teorii równań Eliashberga oraz wzoru (6) obliczono wartości T_C dla wybranych pseudopotencjałów. Szczegółowy przebieg zależności temperatury krytycznej od μ^* przedstawiono na rysunku 3.

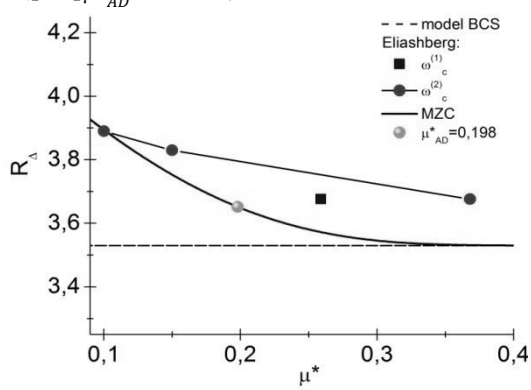


Rysunek 3. Temperatura krytyczna w funkcji pseudopotencjału kulombowskiego [opracowanie własne]

Na podstawie otrzymanych rezultatów stwierdzono, że pseudopotencjał kulombowski dla formuły Allena-Dynesa powinien być rozpatrywany jako parametr dopasowujący. Podobnie jest w równaniach Eliashberga, gdy wartość μ^* znacznie przewyższa 0,1. Należy również zwrócić uwagę, iż dla pseudopotencjału kulombowskiego odpowiadającego skrajnym wartościom przedziału częstości odcięcia otrzymano równie dokładne odwzorowanie temperatury krytycznej.

4.2. Bezwymiarowe parametry termodynamiczne

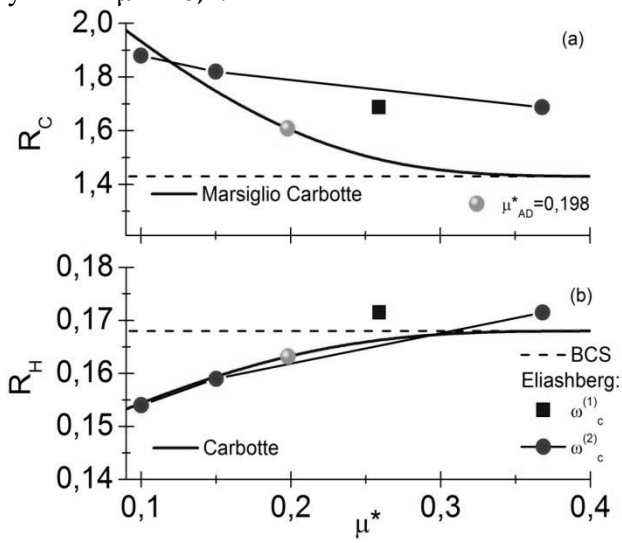
W kolejnym etapie badań obliczono bezwymiarowe parametry termodynamiczne. Początkowo sprawdzono wartość parametru R_Δ przy użyciu formuły Bozidara Mitrovicia, H. G. Zarate i J. P. Carbotte (MZC), która opisana jest wzorem (10). Szczegółowe wyniki przedstawiono na rysunku 3. Otrzymane rezultaty porównano z wynikami uzyskanymi przy pomocy formuły Eliashberga dla dwóch wspomnianych wcześniej wartości częstości odcięcia. Należy zwrócić uwagę, iż dla $\omega_c^{(1)}$ oraz $\omega_c^{(2)}$ otrzymano takie same wartości równe $R_\Delta = 3{,}67$, które nieznacznie odbiegają od tych, wynikających z teorii BCS ($R_\Delta = 3{,}53$). Wraz z obniżeniem wartości pseudopotencjału kulombowskiego wzrastają różnice pomiędzy wynikami otrzymanymi w ramach różnych formalizmów a teorią BCS. Warto również skierować swoją uwagę na fakt, iż formuła semi-analityczna dobrze odtwarza wyniki uzyskane w ramach równań Eliashberga, jeśli za pseudopotencjał kulombowski przyjmiemy wartość otrzymaną ze wzoru Allena-Dynesa ($[R_\Delta]_{\mu^*_{AD}} = 3{,}65$).



Rysunek 4.Bezwymiarowy parametr R_Δ w funkcji pseudopotencjału kulombowskiego [opracowanie własne]

W kolejnej części badań zanalizowano przebieg parametrów R_C oraz R_H, korzystając ze wzorów Franka Marsiglio i J. P. Carbotte (11) oraz Julesa P. Carbotte (12). Dla wanadu o pseudopotencjale oddającym wyniki eksperymentalne wspomniane parametry bezwymiarowe wynoszą odpowiednio 1,68 i 0,171. W teorii BCS natomiast wartości te są na poziomie: $R_C = 1{,}43$ i $R_H = 0{,}168$. Podczas analizy zauważono, podobnie jak w pierwszym przypadku, rosnące wpływy efektów silnosprzężeniowych oraz retardacyjnych, które powodują wzrost różnicy między wartościami teoretycznymi i otrzymanymi w ramach formalizmu równań Eliashberga. Ponownie

szczegółowe rezultaty przedstawiono na wykresie (rysunek 4.). Warto również skierować swoją uwagę na fakt, że oba prezentowane w pracy podejścia prezentują porównywalne wyniki dla $\mu^* \approx 0{,}1$.



Rysunek 5. (a) Wartości parametru R_c w funkcji pseudopotencjału kulombowskiego, (b) wartości parametru R_H w funkcji pseudopotencjału kulombowskiego [opracowanie własne]

4.3. Właściwości termodynamiczne

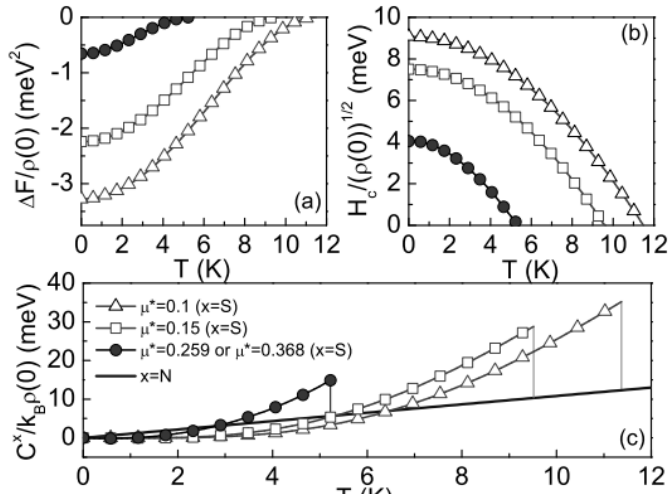
Końcowy etap analizy termodynamiki stanu nadprzewodzącego indukującego się w wanadzie polegał na wyznaczeniu wartości różnicy energii swobodnej stanu normalnego oraz nadprzewodzącego, termodynamicznego pola krytycznego, jak również skoku ciepła właściwego w funkcji temperatury. Do obliczeń wykorzystano rozwiązania równań Eliashberga na osi urojonej. Otrzymane rezultaty zostały przedstawione na rysunku 6.

Na panelu (a) zaprezentowano różnicę energii swobodnej stanu normalnego oraz nadprzewodzącego, która została obliczona na podstawie następującej formuły:

$$\frac{\Delta F}{\rho(0)} = -\frac{2\pi}{\beta}\sum_{n=1}^{M}\left[\sqrt{\omega_n^2+(\Delta_n)^2}-|\omega_n|\right]\left[Z_n^S - Z_n^N\frac{|\omega_n|}{\sqrt{\omega_n^2+(\Delta_n)^2}}\right] \quad (13)$$

gdzie $\rho(0)$ reprezentuje wartość elektronowej gęstości stanów na poziomie Fermiego, natomiast Z_n^S i Z_n^N to czynniki re normalizujące funkcję falową dla stanu nadprzewodzącego oraz normalnego.

Wzrost wartości pseudopotencjału kulombowskiego spowodował spadek wartości różnicy energii swobodnej ponad pięciokrotnie (dokładnie 5,33), co może odpowiadać wpływowi deparujących korelacji elektronowych.



Rysunek 6. (a) Różnica energii swobodnej stanu normalnego oraz nadprzewodzącego; (b) Termodynamiczne pole krytyczne w funkcji temperatury; (c) Skok ciepła właściwego w funkcji temperatury [opracowanie własne]

Na podstawie otrzymanych wyników można wyznaczyć termodynamiczne pole krytyczne. Zależność $H_C(T)$ przedstawiono na rysunku 6. (b). Widać, że wartość termo dyna-micznego pola krytycznego zmniejszyła się ponad dwukrotnie wraz ze wzrostem pseudopotencjału kulombowskiego.

Jako ostatnia została wyznaczona różnica ciepła właściwego pomiędzy stanem nadprzewodzącym a normalnym. Wyniki przedstawiono na panelu (c) rysunku 6. Na wykresie można zauważyć charakterystyczny skok ciepła właściwego w temperaturze krytycznej, który został oznaczony pionową linią. Dodatkowo, ponownie wraz ze wzrostem pseudopotencjału kulombowskiego wartość parametru termodynamicznego zmniejszyła się. Podobnie jak termodynamiczne pole krytyczne, również ostatnia z wyznaczonych wielkości zmalała dwukrotnie.

5. Podsumowanie

W pracy przedstawione zostało porównanie wybranych parametrów termo dyna-micznych wanadu, będącego w stanie nadprzewodzącym. Wanad to pierwiastek z grupy metali przejściowych, którego temperatura krytyczna wynosi $5{,}4\ K$. Wyniki otrzymano po przeprowadzeniu analizy numerycznej w ramach formalizmu równań Eliashberga, jak również korzystając ze wzorów semi-analitycznych.

Interesujący jest fakt, iż mimo tego że wanad charakteryzuje się wysoką wartością stałej sprzężenia elektron-fonon ($\lambda = 0{,}91$), otrzymane wartości w ramach formuły Eliashberga jedynie nieznacznie odbiegają od tych, uzyskanych na podstawie wzorów semi-analitycznych, opierających się na teorii BCS. Wzory semi-analityczne pomimo mniejszego poziomu skomplikowania niż równania Eliashberga, pozwalają z dobrym przybliżeniem odtworzyć zaawansowane rezultaty, jeśli za μ^* zostanie przyjęta war-

tość wynikająca z formuły Allena-Dynesa. Pseudopotencjał dla równań Eliashberga jest znacznie wyższy niż wspomniany.

Rezultaty wyrażają zgodność z twierdzeniem mówiącym o tym, iż wzory semi-analityczne, które określają właściwości termodynamiczne stanu nadprzewodzącego, dobrze odtwarzają przewidywania uzyskane przy pomocy formuły Eliashberga w przypadku, gdy wartość pseudopotencjału kulombowskiego jest niska: $\mu^* < 0{,}2$. Więcej o tym zagadnieniu można przeczytać w pracach [15,16]. Biorąc pod uwagę uzyskane wyniki należy zauważyć, iż warto przeprowadzać porównania rezultatów otrzymanych z wykorzystaniem różnych formalizmów. Ciekawym kierunkiem dla dalszych badań byłoby wyrugowanie z używanych wzorów parametru dopasowującego, za jaki w wielu wypadkach można traktować pseudopotencjał kulombowski μ^*.

Literatura

1. Drozdov A. P., Eremets M. I, Troyan I. A., Ksenofontov V., Shylin S. I..*Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system*, Nature 525 (2015), s. 73-76.
2. Troyan I., Gavriliuk A., Ruffer R., Chumakov A., Mironovich A., Lyubutin I., Perekalin D., A. P. Drozdov, M. I. Eremets *Observation of superconductivity in hydrogen sulfide from nuclear resonant scattering*, Science 351 (2016), s. 1303-1306.
3. Ray P. J. *Structural investigation of La(2 − x)Sr(x)CuO(4 + y) - Following staging as a function of temperature Niels Bohr Institute*, Faculty of Science, University of Copenhagen, Kopenhaga, 2015.
4. Savrasov S. Y., Savrasov D. Y. *Electron-phonon interactions and related physical properties of metals from linear-response theory*, Physical Review B 54 (1969), s. 16487.
5. Westerdale S. *Superconducting Metals: Finding Critical Temperatures and Observing Phenomena*, Cambridge, 2010.
6. Fröhlich H. *Theory of the superconducting state. I. The ground state at the absolute zero of temperature*, Physical Review 79 (1950), s. 845.
7. Fetter A. L.,Walecka J. D. *Quantum theory of many-particle systems*, Courier Corporation, North Chelmsford, 2012.
8. Elk K., Gasser . *Die Methode der Greenschen Funktionen in der Festkörperphysik Akademie - Verlag*, Berlin, 1979.
9. Carbotte J. P. *Properties of boson-exchange superconductors*, Reviews of Modern Physics 62 (1990), s. 1027.
10. Wierzbowska. *Effect of spin fluctuations on Tc from density-functional theory for superconductors*, The European Physical Journal B - Condensed Matter and Complex Systems 48 (2005), s. 207-217.
11. Allen P. B., Dynes R. C.*Transition temperature of strong-coupled superconductors reanalyze*, . Physical Review B 12 (1975), s. 905
12. Bardeen J., Cooper L. N., Schrieffer J. R. *Microscopic theory of superconductivity*, Physical Review 106 (1957), s. 162.
13. Mitrović B., Zarate H. G., Carbotte J. P. *The ratio* $\frac{2\Delta_0}{k_B T_C}$ *within Eliashberg theory*, Physical Review B 29 (1984), s. 184.
14. Marsiglio F., Carbotte J. P. *Strong-coupling corrections to Bardeen-Cooper-Schrieffer ratios*, Physical Review B 33 (1986), s. 6141.

15. Radebaugh R., Keesom P. H.. Low-Temperature *Thermodynamic Properties of Vanadium. I. Superconducting and Normal States*, Physical Review 149 (1966), s. 209
16. Szczęśniak R., Jarosik M. W., Szczęśniak., D. *Pressure-induced superconductivity in the fcc phase of lithium: strong-coupling approach*, Physica B 405 (2010), s. 4897-4902.
17. Durajski A. P., Szczęśniak D., Szczęśniak R. *Study of the superconducting phase in silicene under biaxial tensile strain*, Solid State Communications 200 (2014), s. 17-21.

Równania Eliashberga i wzory semi-analityczne w opisie właściwości termodynamicznych stanu nadprzewodzącego indukującego się w wanadzie

Streszczenie

Temperatura krytyczna w wanadzie jest równa $T_C = 5{,}3K$, podczas gdy pseudopotencjał kulombowski obliczony przy pomocy równań Eliashberga ma nadzwyczaj wysoką wartość. Pomimo wysokiej stałej sprzężenia elektron-fonon, parametry stanu nadprzewodzącego nie odbiegają znacznie od przewidywań teorii BCS. Pokazaliśmy, że wyniki zaawansowanego formalizmu równań Eliashberga mogą być relatywnie dokładnie odtworzone przy użyciu wzorów semi-analitycznych, jeśli za wartość μ^* przyjmiemy parametr obliczony na podstawie wzoru Allena-Dynesa.

Słowa kluczowe: nadprzewodnictwo, równania Eliashberga, wzory semi-analityczne, wanad

Eliashberg equations and semi-analytical formulas description of the thermodynamic properties of the superconducting state inducing in vanadium

Abstract

The critical temperature in vanadium is equal to $T_C = 5{,}3\ K$, while the Coulomb pseudopotential, calculated by Eliashberg equations has anomalously high value. Despire the large electron-phonon coupling constant, the quantities of superconducting state not deviating much from the predictions of the BCS theory. We have shown that the results of the advanced Eliashberg formalism can be relatively precisely reproduced with the help of the semi-analytical formulas, if the value of μ^* is determined on the basis of the Allen-Dynes formula.

Keywords: superconductivity, Eliashberg equations, semi-analitical equations, vanadium

Kaja Wójcik[1] , Tomasz Magiera[2]

Model matematyczny symulujący zderzenia narciarzy z przeszkodami na stoku

1. Wprowadzenie

Z roku na rok zwiększa się liczba wypadków na stokach narciarskich. W Stanach Zjednoczonych Ameryki Północnej według NSAA – National Ski Areas Association rocznie odnotowuje się ok. 150000 wypadków. Średnio ok. 40 osób ponosi w nich śmierć, a ok. 50 osób odnosi groźne obrażenia. W Szwajcarii liczba wypadków (szacowana na ok. 36000 rocznie) przeliczana jest na koszty, które na skutek nich musiały zostać poniesione. Według SUVA – The Swiss National Accident Insurance Fund rocznie wydawana jest kwota ok. 240 milinów CHF (ok. 950 milinów złotych), co wynosi ok. 8% wszystkich kosztów wypadków zaistniałych na terenie Szwajcarii w danym roku [1]. W Polsce według badania „Aktywność fizyczna Polaków" przeprowadzonego przez CBOS: 30% obywateli RP deklaruje, że potrafi jeździć na nartach, a 11% deklaruje, że jeździ na nartach przynajmniej jeden raz w roku [2]. Przy tak dużej liczbie narciarzy do wypadków podczas trwania sezonu zimowego dochodzi codziennie. W sezonie narciarskim 2015/2016 szacuje się, że doszło do ok. 9000 wypadków, w tym co najmniej dwa z nich były ze skutkiem śmiertelnym [3]. Są to wypadki rejestrowane - takie, w których konieczna była interwencja TOPR, GOPR lub ratowników narciarskich i nierejestrowane, w których poszkodowany samodzielnie zapewnił sobie transport do szpitala/lekarza lub nie była potrzebna interwencja lekarska. Główną przyczyną wypadków narciarskich na jest niedostosowanie prędkości przez narciarzy do swoich umiejętności jazdy i warunków panujących na stokach. Dodatkowo rozwój sprzętu narciarskiego pozwala na uzyskiwanie dużo większej, niż jeszcze pręlat temu, prędkości narciarzom-amatorom. Większość wypadków są to kolizje z innymi uczestnikami, mniej liczną grupę wypadków stanowią zderzenia narciarzy z przeszkodami znajdującymi się wokół trasy narciarskiej [4]. Są to zwykle dużo poważniejsze wypadki nierzadko skutkujące złamaniami otwartymi, urazami kręgosłupa czy nawet zatrzymaniami akcji serca [3]. Przeciwdziałanie wypadkom/ kolizjom narciarz-narciarz jest trudne i wymaga ciągłego uświadamiania użytkownikom stoku zagrożeń. Najczęściej jest to realizowane poprzez zalecenie stosowania się do tzw. „Dekalogu narciarza FIS", stworzenie regulaminu korzystania ze stoku w ośrodku narciarskim czy przeprowadzania kampanii promującej bezpieczeństwo. Stosowana jest również ochrona bezpośrednia użytkowników stoku. Narciarze używają specjalnych zabezpieczeń poszczególnych części ciała: kask chroniący głowę lub tzw. „żółw" chroniący kręgosłup. Do podniesienia poziomu bezpieczeństwa narciarzy wykorzystywane są również nowe rozwiązania techniczne dotyczące wiązań

[1] wojcikaja@gmail.com, student, Katedra Transportu Linowego, Wydział Inżynierii Mechanicznej i Robotyki, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, http://ktl.imir.agh.edu.pl
[2] magiera@agh.edu.pl, Katedra Transportu Linowego, Wydział Inżynierii Mechanicznej i Robotyki, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, http://ktl.imir.agh.edu.pl

narciarskich, czy indywidualne systemy ostrzegające o wystąpieniu możliwego zagrożenia. Jednym z takich systemów, są tzw. „gogle audiowizualne" odbierające sygnały o niebezpieczeństwie z nadajników umieszczonych na stoku narciarskim [1]. Wykorzystywane jest również rozwiązanie bazujące na analizie danych zarejestrowanych przez czujniki parametrów ruchu narciarzy. Na ich podstawie określane są potencjalne miejsca niebezpieczne na stokach i wprowadzane są w nich rozwiązania minimalizujące prawdopodobieństwo wystąpienia sytuacji niebezpiecznych [5].

W celu zmniejszenia liczby wypadków pomiędzy poruszającymi się na stoku narciarzami, stosowane są modele matematyczne, bazujące na podejściu stosowanym w modelach dotyczących analizy przepływu pieszych. Wykorzystuje się je do predykcji zachowań użytkowników stoku i wpływaniu na nich, poprzez znaki lub innego rodzaju ograniczenia ustawiane na trasie [1].

Ze względu na występującą mniejszą liczbę zmiennych o charakterze stochastycznym, znacznie łatwiejszym powinno być ograniczenie skutków kolizji pary narciarz-przeszkoda. Niejednokrotnie lepsze zabezpieczenie przeszkód na stoku tj. armatki, podpory czy drzewa, poprzez zastosowanie odpowiednich siatek czy materacy zabezpieczających, mogłoby zapobiec tragicznym w skutkach wypadkom.

Na dzień dzisiejszy nie ma przepisów regulujących sposób ustawiania i dobór odpowiednich siatek i materacy zabezpieczających na stokach narciarskich (fot.1.). Ponadto normy europejskie regulujące parametry geometryczne elementów zabezpieczających przeszkody na stoku, nie określają poziomu absorbowanej energii przez nie podczas zderzenia z narciarzem [4]. Należy również mieć na uwadze, że poziom bezpieczeństwa obniżany jest przez właścicieli stoków poprzez instalowanie tańszych rozwiązań niezapewniających całkowitego wyhamowania narciarza przed przeszkodą (siatki typu B lub ostrzegawcze zamiast siatek typu A).

Opracowany model matematyczny oraz przeprowadzone na stokach narciarskich testy mają na celu określenie poziomu bezpieczeństwa, który zapewniają siatki zabezpieczające-ostrzegawcze (tzw. zintegrowane) oraz materace. Na tej podstawie, w późniejszym okresie, sformułowane zostaną zalecenia dotyczące rozstawiania elementów zabezpieczających na stokach narciarskich, w zależności od wartości parametrów ruchu narciarzy i warunków pogodowych w analizowanym miejscu.

Fotografia 1. Armatka zabezpieczona siatkami zintegrowanymi na stoku narciarskim [6]

2. Aktualny stan wiedzy dotyczący zabezpieczeń stosowanych na stokach narciarskich

Zagadnienie obniżenia negatywnego wpływu na zdrowie narciarzy podczas zderzenia z elementami zabezpieczającymi przeszkody na stoku, w ostatnich latach, było analizowane przez kilka zespołów naukowych.

Kompleksowe testy zderzeń z materacami narciarskimi zostały przeprowadzone przez zespół pod kierownictwem N. Petrone. W 2009 roku. Przeprowadził on testy materaców narciarskich typu A, B i C (fot.2.). Materace typu A są to materace wypełnione powietrzem, zwykle stosowane na zawodach narciarskich, materace typu B są to materace piankowe segmentowe, natomiast materace typu C są konstrukcji piankowej. Przeprowadzone testy polegały na pomiarze przyspieszeń masy uderzającej z odpowiednią prędkością w zabezpieczenia. Maksymalna prędkość uderzenia została ustalona na 60 km/h. Co odpowiada średniej prędkości zaawansowanego narciarza poruszającego się po stoku [4].

Fotografia 2. Manekin wyposażony w akcelerometr, testy materaców wykonywane przez zespół N. Petrone [6]

Wynikiem przeprowadzonych eksperymentów było określenie poziomu absorbcji energii zderzenia zapewnianego przez poszczególne rodzaje materaców oraz porównanie ich ze sobą.

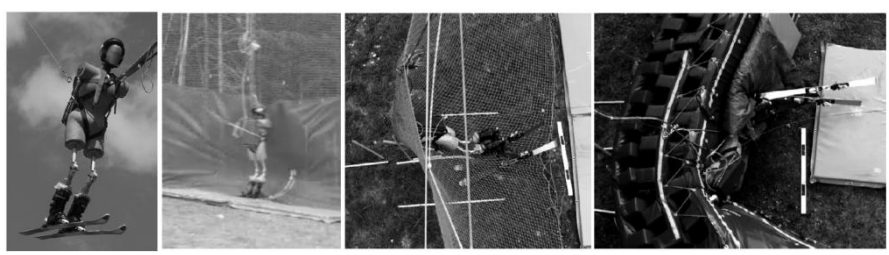
W 2010 roku zespół N. Petrone analizował i określał poziom zabezpieczeń przeszkód na stokach narciarskich poprzez zastosowanie siatek zabezpieczających. Testy zostały przeprowadzone przy użyciu manekina wzorowanego na manekinach z serii Hybrid II. W manekinie pomiarowym wprowadzono liczne uproszczenia, m.in. masa rąk została zredukowana do jego barków, a masa nóg do bioder. Podstawową zaletą zastosowania humanoidalnego manekina ubranego w narty było lepsze odwzorowanie zderzenia narciarza z siatką wyłapującą (typ A). Jednakże, w przeprowadzanych testach nie uwzględniono wpływu siły grawitacji na manekina, w ostatniej fazie zderzenia z siatką. Oddziaływanie siły grawitacji było równoważone przez siłę działającą w linie. Spowodowane to było przyjętym sposobem rozpędzania manekina. Polegał on na wykorzystaniu liny podwieszonej do dźwigu jako wahadła. Prędkość uderzenia manekina w siatki mieściła się w przedziale od 50 do 66 km/h. Zespół N. Petrone przeprowadził testy siatek wyłapujących typu A i materacy piankowych (fot.3.). Systemy zabezpieczające typu A spełniają wymogi FIS (Międzynarodowa Federacja Narciarska). Składają się one ze stalowych podpór osadzonych na fundamentach betonowych, siatki wyłapującej rozpiętej na nich poprzez zastosowanie

systemu naciągów, lin stalowych oraz elementów kotwiących siatkę do podłoża. Testy pozwoliły na określenie siły uderzenia manekina w zabezpieczenie oraz odległości ugięcia siatki, przy której następuje całkowite zatrzymanie manekina.

Bazując na wynikach przedstawionych przez zespół N. Petrone [6, 7], został opracowany model matematyczny, który symulował zderzenie narciarza z siatką wyłapująca typu A. Wykazano, poprzez dokładność jego wyników z pomiarami, że opracowywanie modeli matematycznych zderzeń pary narciarz-siatka, może zmniejszyć liczbę testów, przyspieszyć proces certyfikacji siatek oraz ocenić ich jakość, bez inwestowania dużych środków finansowych w tworzenie ich prototypów [4].

Testy i nowe rozwiązania konstrukcyjne siatek zabezpieczających typu B, zostały przeprowadzone przez [9]. Autorzy przyjęli jako priorytety w swoich badaniach, zmniejszenie wpływu oddziaływań dynamicznych na narciarzy oraz brak możliwości „przelecenia" narciarza przez siatkę. Teoretycznie zostało to zapewnione poprzez stworzenie tzw. „kieszeni" w siatkach. Są to strefy zwiniętej siatki, które podczas uderzenia narciarza otwierają się i wydłużają jego drogę hamowania, jednocześnie zmniejszając wartość opóźnienia działającego na niego. Natomiast zatrzymanie narciarza po uderzeniu w siatkę, było zrealizowane poprzez zaproponowanie nowych typów mocowań tyczek w śniegu, tak aby się nie przechylały i kładły na stoku, pozwalając narciarzowi prześlizgnąć się po nich.

W przedstawionych pracach badawczych nie znaleziono żadnej wzmianki o przeprowadzaniu jakichkolwiek testów z siatkami zabezpieczającymi-ostrzegawczymi (tzw. zintegrowanymi). Siatki te, są bardzo często stosowane jako jedno z zabezpieczeń przed uderzeniem w sztuczną przeszkodę znajdującą się na stoku (fot.1.).



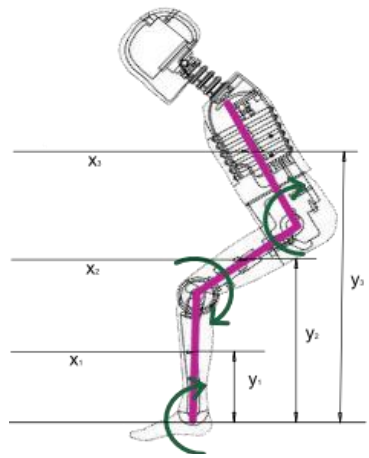
Fotografia 3. Manekin przygotowany do testu, testy siatek typu A i materaca typu C [7]

3. Model matematyczny symulujący zderzenie narciarza na stoku z przeszkodą

W celu ograniczenia kosztownych i czasami trudnych do wykonania testów na stoku narciarskim został opracowany model matematyczny symulujący zderzenia narciarzy z przeszkodami. Model matematyczny został oparty na równaniach Lagrange'a II rodzaju. Podczas procesu budowy modelu fizycznego, wprowadzano w nim kolejne uproszczenia mające na celu przyspieszenie procesu obliczeniowego modelu matematycznego, przy założonej jego dokładności podczas procesu

weryfikacji modelu. Uproszczenia modelu bazowały również na wynikach badań przeprowadzonych przez N. Petrone [7]. Głównymi uproszczeniami założonymi w modelu fizycznym narciarza było zredukowanie masy rąk i głowy do korpusu oraz zredukowanie masy dwóch nóg do jednej. Analizowany układ posiadał pięć stopni swobody (rys. 1.).

Zgodnie z teorią budowy równań dynamicznych Lagrange'a II rodzaju rozpisane zostały równania energii kinetycznej (1) oraz energii potencjalnej (2). Dyssypacja energii nie została uwzględniona w równaniach głównych, a dopiero w macierzy wyrazów wolnych. Pozwoliło to na łatwiejszą modyfikację modelu w tracie wprowadzania zmiennych danych początkowych.



Rysunek 1. Rysunek przedstawiający proste symulujące manekina i współrzędne użyte w równaniach głównych [opracowanie własne]

Energia kinetyczna:

$$E_k = \frac{1}{2}\begin{pmatrix} m_1 \cdot \dot{x}_1^2 + m_1 \cdot \dot{y}_1^2 + I_1 \cdot \dot{\varphi}_1^2 + m_2 \cdot \dot{x}_2^2 + m_2 \cdot \dot{y}_2^2 + I_2 \cdot \dot{\varphi}_2^2 + \\ +m_3 \cdot \dot{x}_3^2 + m_3 \cdot \dot{y}_3^2 + I_3 \cdot \dot{\varphi}_3^2 \end{pmatrix} \quad (1)$$

gdzie: m_1 - masa dwóch nóg do kolan, m_2- masa dwóch ud wraz z masą miednicy, m_3 - masa korpusu, rąk oraz głowy, $x_1, x_2, x_3, y_1, y_2, y_3$ -współrzędne opisujące środek poszczególnych elementów manekina, I_1, I_2, I_3 - momenty bezwładności poszczególnych elementów manekina, $\varphi_1, \varphi_2, \varphi_3$ - kąty obrotu na poszczególnych węzłach

Energia potencjalna:

$$E_p = m_1 \cdot g \cdot y_1 + m_2 \cdot g \cdot y_2 + m_3 \cdot g \cdot y_3 \qquad (2)$$

gdzie: m_1 - masa dwóch nóg do kolan, m_2 - masa dwóch ud wraz z masą miednicy, m_3 - masa korpusu, rąk oraz głowy, y_1, y_2, y_3 - współrzędne opisujące środek poszczególnych elementów manekina, g - przyspieszenie ziemskie

Ogólne równanie ruchu:

$$\frac{d}{dt}\left(\frac{\partial E_k}{\partial \dot{q}_i}\right) - \frac{\partial E_k}{\partial q_i} + \frac{\partial E_p}{\partial q_i} = Q_i \quad (3)$$

E_k - energia kinetyczna, E_p - energia potencjalna, q_i - współrzędna uogólniona $(x_1, x_2, x_3, y_1, y_2, y_3, \varphi_1, \varphi_2, \varphi_3)$, Q_i - siła uogólniona

Kolejnym krokiem tworzenia równań ruchu było napisanie równań więzów i zastosowanie ich w równaniu ogólnym (3). Po odpowiednim przekształceniu wzorów i zapisaniu równania w postaci macierzowej został napisany skrypt w programie MatLab. Równania zostały rozwiązane za pomocą funkcji ode4, opartej na metodzie Rungego-Kutty. Pierwszym krokiem weryfikacyjnym modelu narciarza, było obciążenie go działaniem siły grawitacji (fot. 4). Obecnie prowadzone są weryfikacje symulacji zderzenia manekina z przeszkodami (ciała nieodkształcalne i odkształcalne).

4. Weryfikacji modelu matematycznego – testy zderzeniowe

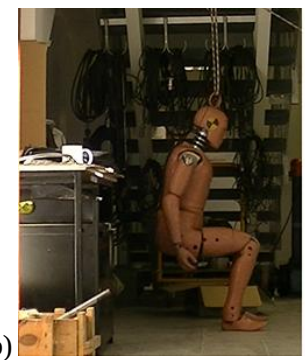
Najczęstszymi wielkościami fizycznymi analizowanymi podczas zderzeń, są przyspieszenia działające na narciarza, przyrost przyspieszenia oraz siła uderzenia działająca na jednostkę powierzchni. Uwzględnia się również kierunek działania składowych sił, względem usytuowania narciarza, jak również czas trwania oddziaływań dynamicznych [9]. Dodatkowo dla odpowiedniego przygotowania się do testów weryfikacyjnych na stoku narciarskim oraz odpowiedniego opracowania modelu matematycznego wzięto pod uwagę średnie prędkości uzyskiwane przez narciarzy na stoku narciarskim oraz siły działające na narciarza w trakcie zjazdu ze stoku [10]. Powołano się na artykuł, w którym zespół naukowców z Akademii Górniczo-Hutniczej przeanalizował siły działające na narciarza w zależności od postawy przyjętej przez zjeżdżającego. Dodatkowo na podstawie zapisu z kamer ośrodka narciarskiego w Jaworzynie Krynickiej zespół opracował mapę prędkości uzyskiwanych przez narciarzy na długości całego stoku oraz przeanalizował dystans zachowywany między narciarzami. Badania przeprowadzone przez zespół pomogły określić średnie prędkości narciarzy osiągane w miejscach newralgicznych tj. obrzeża stoku narciarskiego i okolice sztucznych przeszkód.

W celu przeprowadzenia testów potwierdzających poprawność opracowanego modelu, użyto manekina pomiarowego Hybrid III, mężczyzna 50 centyl. Manekin mierzy 180 cm i waży 80 kg. Wyposażony jest w dwa niezależne układy pomiarowe. Pierwszy układ składa się z pięciu akcelerometrów trójosiowych o zakresie pomiarowym $\pm$ 16 g. Drugi stanowią dwa akcelerometry trójosiowe o zakresie pomiarowym ± 200 g.

Początkowo wykonano badania laboratoryjne mające na celu weryfikację wpływu siły grawitacji na ruch jego poszczególnych elementów, podczas swobodnego spadku i zderzenia z podłożem manekina pomiarowego. Manekin był podwieszany na linie, która następnie była zwalniana. Wszystkie testy były nagrywane kamerą, w celu przeprowadzenia późniejszej dokładniej analizy ruchu poszczególnych elementów manekina. Testy wykazały poprawność wykonanego modelu matematycznego z założonym marginesem błędu. Różnice pomiędzy animacją z modelu, a zarejestrowanym obrazem wynikają z braku zamodelowanych w programie sił tarcia pochodzących od podłoża oraz niewielkiej ruchomości manekina w miednicy (fot.4.).

a)

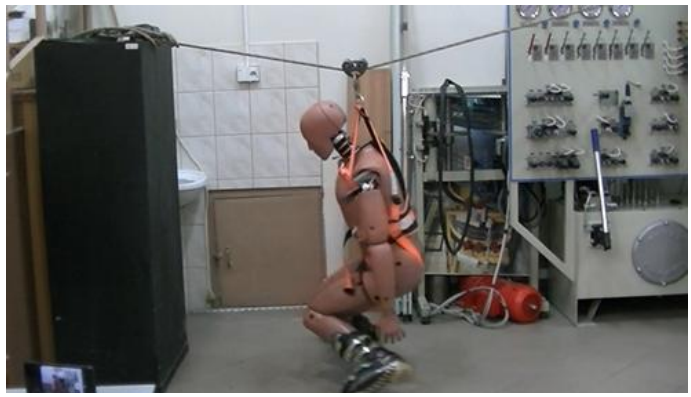
b)



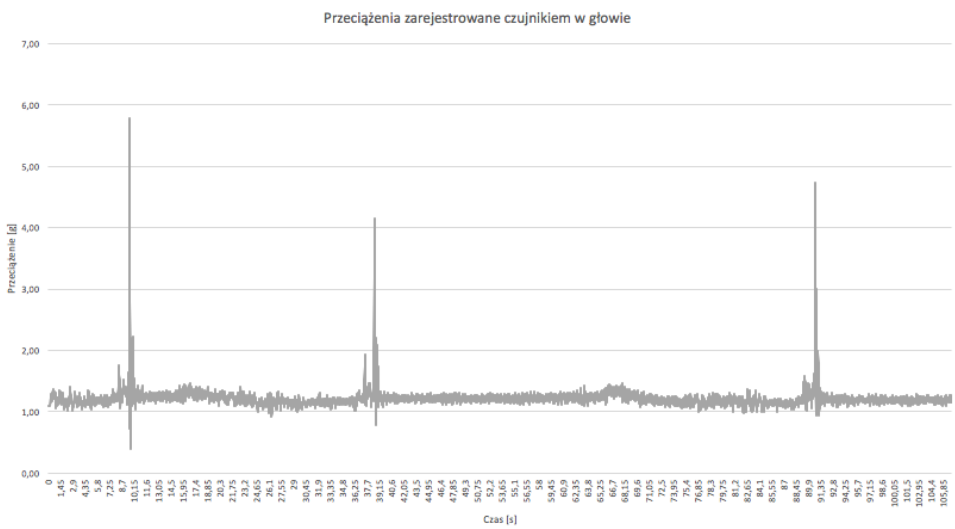
c)

Fotografia 4. Kolejne fazy (a)., b)., c).) upadku manekina: po lewej manekin w trakcie testu, po prawej symulacja przeprowadzona w programie MatLab [opracowanie własne]

Po wykonaniu podstawowych testów weryfikacyjnych opisanych w poprzednim akapicie, przeprowadzono dalsze badania laboratoryjne. Kolejne testy polegały na wykonaniu kilku zderzeń manekina z przeszkodą. W tym celu została rozpięta tyrolka alpinistyczna, na której rozpędzano manekina. Manekin zderzany był z szafą pancerną. Całe doświadczenie miało na celu określenie przeciążeń dynamicznych działających na manekina podczas zderzenia (wyk. 1.) oraz późniejsze zweryfikowanie modelu przy symulowaniu zderzenia z wysoką, sztywną przeszkodą. Manekin był rozpędzany do prędkości około 12 km/h (fot. 5).



Fotografia 5. Manekin rozpędzony na tyrolce alpinistycznej przed zderzeniem [opracowanie własne]

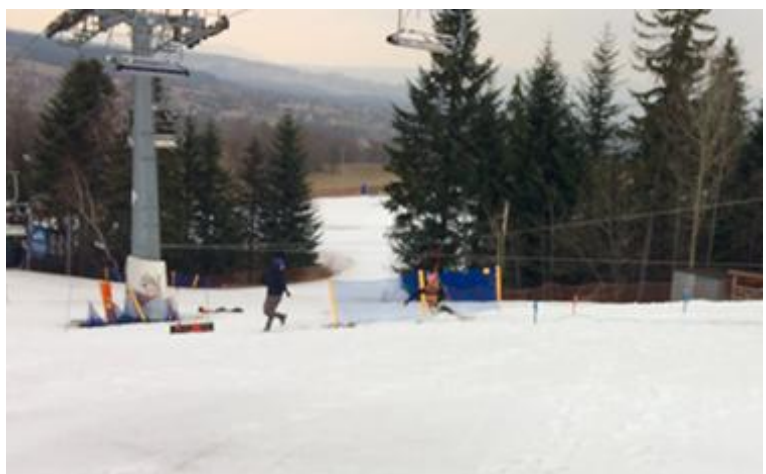


Wykres 1. Przeciążenia zarejestrowane podczas zderzenia manekina (trzy próby) z przeszkodą sztywną – wyniki z akcelerometru zamontowanego w głowie [opracowanie własne]

Wykres 1 przedstawia trzy próby zderzenia manekina z przeszkodą. W pierwszej próbie (pierwszy pik na wyk.1. patrząc od lewej strony) czujnik zarejestrował średnie przeciążenie o wartości 5,79 g, w drugiej (drugi pik na wyk.1. patrząc od lewej strony) – 4,16 g, a w trzeciej – 4,74 g. Różnice w przeprowadzonych testach są niewielkie i wynikają z różnej prędkości manekina uzyskiwanej w poszczególnych pomiarach.

Przeprowadzono również testy zderzeń manekina na stoku narciarskim, wykorzystując doświadczenie zderzenia manekina z przeszkodą wykonane w laboratorium. Na stoku narciarskim rozpięto tyrolkę alpinistyczną pomiędzy podporami kolei linowej. Manekin rozpędzany na tyrolce uderzał w siatkę zintegrowaną, która najczęściej ustawiana jest na stokach w celu odgrodzenia przeszkód tj. armatka śnieżna

lub podpora kolei linowej. Dodatkowo manekin ustawiany był tak aby pozycja w momencie najazdu na siatkę była jak najbardziej zbliżona do pozycji narciarza w podobnej sytuacji. Pozycja zjazdowa ma wpływ na opory powietrza oddziałujące na narciarza co za tym idzie w dużym stopniu determinuje prędkość z jaką uderza on w siatkę [10]. W trakcie testów uwzględniony został typ śniegu, po którym poruszał się manekin, co za tym idzie współczynnik tarcia miedzy nartą a podłożem. Testy wykazały, że nawet przy niskiej prędkości ok. 17 km/h manekin nie jest wyłapywany przez siatkę (fot. 6.), co może powodować w warunkach rzeczywistych uderzenie narciarza bezpośrednio w zabezpieczoną (lub nie) przeszkodę. Ponadto przeprowadzono zderzenia manekina z przeszkodami zabezpieczonymi materacami ochronnymi o grubości 10 cm. Otrzymane wyniki z testów są obecnie wykorzystywane do dalszej weryfikacji parametrów zawartych w opracowywanym modelu matematycznym.



Fotografia 6. Manekin wpadający w siatkę w trakcie testów przeprowadzonych na stoku narciarskim [opracowanie własne]

5. Wnioski

Badania mające na celu weryfikację poziomu bezpieczeństwa zapewnianego przez elementy zabezpieczające stoki narciarskie są niezbędne do zapewnienia większego bezpieczeństwa na stokach narciarskich. Przeprowadzone testy zderzeń manekina pomiarowego z siatkami i materacami zabezpieczającymi stoki, wykazały iż niezbędne jest sformułowanie zaleceń dotyczących ich ustawiania przed sztuczną przeszkodą na stokach. Ukończony i zweryfikowany model matematyczny będzie pomocny w ich określeniu. Umożliwi on zbadanie poszczególnych elementów zabezpieczających przy różnych parametrach określających ruch oraz właściwości fizyczne narciarzy (tj. prędkość, masa, wzrost narciarza) w sposób łatwy i nie wymagający przeprowadzania licznych testów na stoku.

$$\boldsymbol{X} = [\boldsymbol{X}_1 \dots \boldsymbol{X}_p],\ s_i = \left(\boldsymbol{X}_i^T \boldsymbol{X}_i\right)^{-\frac{1}{2}}$$

Literatura

1. Holleczek T., *Modeling and Shaping Skiing Traffic*, Praca doktorska ETH No. 20413, 2012.
2. http://www.cbos.pl/SPISKOM.POL/2013/K_129_13.PDF (data dostępu: 25.03.2017.)
3. Na podstawie rozmów przeprowadzonych z ratownikami TOPR i lekarzami ze Szpitala Powiatowego im. Dr Tytusa Chałubińskiego w Zakopanem, marzec 2016.
4. Anghileri M., Eralti D., Milanesea A., Pratoa A., i inni, *Nonlinear finite element analysis applied to the development of alpine ski safety net*, International Journal of Crashworthiness, 2014, s.161-171.
5. Quagliotto O. *Sensors for performance and safety analysis in alpine skiing*, praca magisterska, Dipartimento Ingegneria Meccanica, Padova University, 2015 .
6. http://stokinarciarskie.marabut.com/uploads/images/catalog_src/siatka-zintegrowana-5m_src_2.jpg (data dostępu: 25.03.2017) .
7. Petrone N., Pollazzon C., Morandin T., *Structural Behaviour of Ski Safety Barriers during Impacts of an Instrumented Dummy* , The Engineering of Sport, 7 (2009), s. 633-642 .
8. Petrone N., Ceolin F., Morandin T., *Full scale impact testing of ski safety barriers using an instrumented anthropomorphic dummy*, Procedia Engineering, 6 (2010), s. 2593-2598 .
9. Edovas T., Ostling J., *Energy Absorbing Barrier Project Number: CAB 1502*, projekt, Worcester Polytechnic Institute, 2015 .
10. Korecki T., Pałka D., Wąs J., *Adaptation of Social Force Model for simulation of downhill skiing*, Journal of Computational Science, 2016, s. 29-42.
11. Wang X., Yanming F., Zhao L., *The establishment of mathematical model of the take-off speed of aerials of freestyle skiing*, 33rd International Conference on Biomechanics in Sports, Poitiers, France, 6-7 2015.

Model matematyczny symulujący zderzenia narciarzy z przeszkodami na stoku

Streszczenie

Artykuł przedstawia problem niedostatecznego zabezpieczenia stoków narciarskich. W celu ograniczenia wypadków narciarz-przeszkoda autorzy podjęli pracę nad opracowaniem modelu matematycznego symulującego zderzenia narciarzy z przeszkodami. Bazowy model matematyczny został opracowany na podstawie równań Lagrange'a II rodzaju. Skrypt programu, rozwiązujący równania ruchu, został napisany w programie MatLab. W celu weryfikacji modelu przeprowadzono testy laboratoryjne oraz na stokach narciarskich. W trakcie testów wykorzystany został manekin Hybrid III rozpędzany na tyrolce alpinistycznej. Testy wykazały nieprawidłowości w stosowaniu niektórych zabezpieczeń stoków narciarskich i powinny być wstępem do dyskusji dotyczącej określenia zaleceń ustawiania elementów zabezpieczających w przyszłości.

Słowa kluczowe: siatki zabezpieczające, bezpieczeństwo na stokach, narciarstwo, manekin

A mathematical model simulating crashes of skiers with obstacles located on a ski slope

The article shows the problem of insufficient protection of ski slopes. In order to reduce number of accidents between skiers and obstacles, a basic mathematical model was created. The model bases on Lagrange'a motion equations, which are solved in Matlab. A verification of model was conducted by laboratory and on site tests. To perform it, the Hybrid III dummy was used and a special acceleration track was prepared using steel and material ropes. Tests showed that using safety nets may be at some point insufficient and the achieved results may be an introduction to discussion about setting new parameters for safety nets for ski slopes in the future.

Keywords: safety nets, safety on ski slopes, skiing, Hybrid III dummy

Leszek Sowa[1]

Wykorzystanie symulacji komputerowych do modelowania zjawisk cieplno-przepływowych procesu krzepnięcia wlewka COS

1. Wprowadzenie

Od wieków odlewnictwo stanowiło nie tylko dziedzinę nauki, ale również sztuki, co miało znaczący wpływ na rozwój naszej cywilizacji. Obecnie odlewnictwo należy do jednych z najczęściej stosowanych metod wytwarzania w przemyśle, szczególnie elementów metalowych. Proces odlewania stanowczo różni się od innych procesów wytwarzania. W przypadku procesu odlewania dąży się do otrzymania materiału o odpowiednim składzie chemicznym, jeszcze przed ukształtowaniem odlewu, natomiast w przypadku innych procesów wytwarzania z gotowego materiału poprzez kształtowanie otrzymuje się gotowy wyrób [1]. Z kolei dominującym sposobem przemysłowego wytwarzania wlewków stalowych jest metoda ciągłego odlewania, stąd potrzeba ciągłego jej udoskonalania. Powodem wzrostu popularności tego procesu na świecie była, przede wszystkim potrzeba produkowania dużej ilości wlewków o odpowiedniej jakości, co wiązało się z potrzebą podwyższania wydajności poprzez automatyzację produkcji.

Proces ciągłego odlewania stali (COS) polega na ciągłym wlewaniu ciekłego metalu, przez kilka godzin, do metalowej formy odlewniczej-krystalizatora oraz równoczesnym wysuwaniu z niej tworzącego się wlewka ciągłego. Dzięki temu możliwe jest wykonywanie wlewków, których rozmiary przewyższają znacznie gabaryty formy odlewniczej [2, 3]. Zastosowanie tej metody odlewania stali pozwoliło na eliminację z produkcji stali takich operacji jak: odlewanie stali do wlewnic, wygrzewanie wlewków w piecach wgłębnych oraz walcowanie wstępne w walcarce. W konsekwencji tego rozwiązania zmniejszono energochłonność wytwarzania stali i koszty jej produkcji a zwiększono uzysk materiałowy oraz wyeliminowało wstępne operacje obróbki plastycznej. Proces ciągłego odlewania stali mimo swojej długiej historii dalej jest traktowany jako jeden z najnowocześniejszych procesów odlewania stali. Dzięki właściwemu doborowi parametrów prowadzenia procesu COS możliwe jest wytwarzanie półwyrobów pozbawionych wad powierzchniowych i wewnętrznych, co stanowi dużą zaletę tej metody odlewania. Obecnie ponad 90% stali produkowanej w świecie jest odlewane tym sposobem [2].

[1] sowa@imipkm.pcz.pl, Instytut Mechaniki i Podstaw Konstrukcji Maszyn, Wydział Inżynierii Mechanicznej i Informatyki, Politechnika Częstochowska, www.wimii.pcz.pl

2. Technologia ciągłego odlewania stali

2.1. Ogólna charakterystyka procesu COS

Technologia ciągłego odlewania stali jest jednym z najbardziej efektywnych i nowoczesnych sposobów wytwarzania wlewków stalowych. Proces COS cechuje duża oszczędność materiału, dobra jakość wyrobu, związana z uzyskaniem odpowiedniej struktury, oraz możliwość mechanizacji i automatyzacji procesu. Równocześnie COS jest technologią umożliwiającą sterowanie przebiegiem procesu w stopniu znacznie większym niż technologie klasyczne.

Proces odlewania rozpoczyna się od dostarczenia stopionego metalu, w kadzi głównej, do urządzenia COS i jego zalania do kadzi pośredniej, w której następuje jego oczyszczenie i ujednorodnienie. Następnie metal spływa poprzez wylew zanurzony do wnęki krystalizatora. Dno krystalizatora w tej fazie stanowi drąg startowy, który jest nieruchomy. Po upływie określonego czasu, potrzebnego do utworzenia się wystarczająco wytrzymałej zakrzepłej warstwy wlewka, drąg startowy rozpoczyna swój ruch wzdłuż strefy chłodzenia wtórnego ciągnąc za sobą wlewek aż do strefy cięcia na odpowiedniej długości części.

Strefa chłodzenia pierwotnego zamyka się w krystalizatorze, w którym rozpoczyna się proces krzepnięcia, a ciepło od powierzchni wlewka przejmowane jest głównie przez schładzane wodą ścianki miedziane krystalizatora. Ilość odprowadzanego ciepła od ścian krystalizatora regulowana jest poprzez zmianę szybkości przepływu wody przez krystalizator lub temperaturą wody chłodzącej. Dla zapobieżenia przywieraniu metalu do powierzchni ścian krystalizatora wprowadza się zasypki smarujące lub ewentualnie olej oraz stosuje się ruch oscylacyjny krystalizatora w kierunku odlewania. Podstawową zasadą ruchu posuwisto-zwrotnego krystalizatora jest to, aby jego prędkość ruchu w dół była większa od prędkości wyciągania pasma, co uniemożliwia przyspawanie się naskórka do ścian krystalizatora. W przypadku przyspawania naskórka następuje jego rozerwanie pod wpływem sił ciągnących i może wówczas nastąpić wyciek stali z ciekłego rdzenia pasma. Wlewek po wyjściu z krystalizatora chłodzony jest, w strefie chłodzenia wtórnego, wodą lub rzadziej mgłą powietrzno-wodną. Wyciągane pasmo prowadzone jest pomiędzy parami rolek napędowych zabudowanych w ramach segmentów. Zakrzepnięty naskórek wlewka powinien mieć grubość zapewniającą możliwość bezpiecznego jego wyciągania. Przy zbyt małej grubości może nastąpić rozerwanie naskórka wskutek działania sił ciągnących. Następuje wtedy wyciek stali z ciekłego rdzenia pasma, co powoduje zniszczenie jednego lub kilku segmentów maszyny i przerwanie odlewania stali. Za komorą wtórnego chłodzenia pasmo przekazywane jest na samotoki, gdzie odbywa się cięcie wlewka na odpowiednią długość zazwyczaj za pomocą palników gazowo-powietrznych. Na krótszej ścianie wlewka nabijane jest oznakowanie za pomocą automatycznej znakownicy. Następnie wlewki spychane są popychaczami na stoły odbierające. Na stołach formowane są stosy składające się z kilku wlewków [3].

2.2. Podstawowe parametry ciągłego odlewania stali

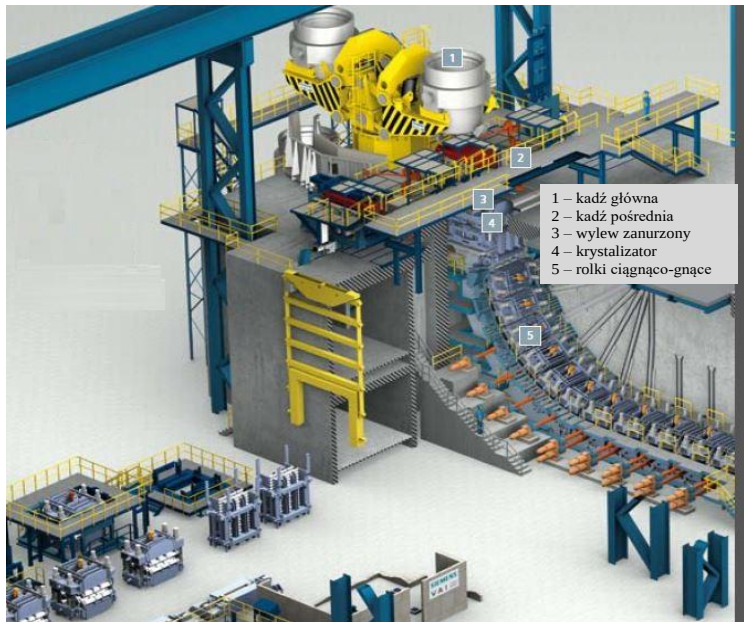
W procesie ciągłego odlewania stali do najważniejszych parametrów, które mają wpływ na jakość wytwarzanych wlewków należą [2, 4-6]:

- prędkość odlewania metalu – w zależności od struktury i zawartości odlewu stopy, które przejawiają dużą skłonność do pęknięć powinno odlewać się z małą prędkością, natomiast stopy plastyczne powinno odlewać się z wyższą prędkością; w procesie COS przedział prędkości odlewania metalu zawiera się od kilku mm/s do 0.5 m/s; z prędkością odlewania ściśle skorelowana jest prędkość zalewania krystalizatora spełniająca warunek ciągłości przepływu cieczy [5]; temperatura odlewania metalu – aby uzyskany stop nie został nasycony szkodliwymi gazami oraz aby warunki odlewania były optymalne temperatura ta powinna być wyższa o około 50K od temperatury likwidusu odlewanego metalu [4],
- intensywność chłodzenia ciekłego metalu w krystalizatorze – odlewany metal poddawany jest najintensywniejszemu chłodzeniu w okolicach ścianki krystalizatora; z powodu koncentracji domieszek stali może dojść do tak zwanego przechłodzenia stężeniowego, które może spowodować nietrwałość frontu krzepnięcia stali; aby uniknąć tego typu nietrwałości ściana krystalizatora musi mieścić się w odpowiednim przedziale grubości [6],
- smarowanie powierzchni ścianek krystalizatora – aby zredukować tarcie między ściankami krystalizatora oraz aby powstrzymać przywieranie naskórka do ścianek krystalizatora stosuję się smarowanie ścianek krystalizatora; podczas odlewania małych wlewków stosuje się smarowanie różnego rodzaju olejami, natomiast podczas odlewania większych wlewków stosuje się zasypkę krystalizatorową [7].

2.3. Urządzenie do procesu COS

Obecnie stosowane najczęściej urządzenia COS (rys.1) zbudowane są z następujących głównych elementów konstrukcyjnych a mianowicie z [3]:

- kadzi głównej – jej celem jest dostarczenie ciekłego metalu do urządzenia COS, który następnie spływa do kadzi pośredniej,
- kadzi pośredniej – która jest elementem urządzenia COS zapewniającym proces ciągłego odlewania, w przypadku występowania wielożyłowego urządzenia do ciągłego odlewania, kadź ta rozprowadza dodatkowo ciekłą stal na kilka krystalizatorów,
- krystalizatora – który jest najważniejszą częścią konstrukcyjną całego urządzenia, ponieważ w nim zachodzi proces krzepnięcia ciekłego metalu wskutek intensywnego chłodzenia wodą i powstawanie wlewka COS,
- strefy wtórnego chłodzenia – w której dochodzi do całkowitego zakrzepnięcia metalu,
- klatek ciągnąco-prostujących – których celem jest transport zakrzepniętego wlewka z odpowiednią szybkością do strefy cięcia,
- palnika – będącego ostatnim urządzeniem, które tnie wlewek na części, transportowane później do przeróbki plastycznej.



Rys. 1. Urządzenie do ciągłego odlewania stali [2]

2.4. Zasilanie krystalizatora ciągłego odlewania ciekłym metalem

Ciekły metal dawniej doprowadzany był do krystalizatora swobodną strugą, później przez zanurzony wylew z pojedynczym pionowym wylotem a następnie zanurzonym wylewem skrzynkowym z dwoma poziomymi wylotami. Zaletą wylewu skrzynkowego jest znikome utlenianie ciekłej stali, oraz równomierne i z małą prędkością doprowadzenie ciekłego metalu do krystalizatora. Nie występuje wówczas wypłukiwanie zakrzepłych warstw wlewka. Eliminuje się w ten sposób jedną z przyczyn pęknięć podłużnych wlewka wynikającą z osłabienia wytrzymałościowego naskórka. Jednocześnie skrzynkowe doprowadzenie metalu pozwala na osiągnięcie wystarczającego ruchu lustra metalu, topienia się zasypki odlewniczej i uniknięcia tworzenia się skrzepów na powierzchni ciekłej stali. Są to elementy pozwalające na odlewanie wlewków przy dużych prędkościach wyciągania [8, 9]. W tym względzie istotnego znaczenia nabiera także temperatura ciekłego metalu w kadzi pośredniej. Zbyt wysoka - przyczynia się do wymywania zakrzepłych warstw wlewka, niska - do tworzenia się skrzepów na powierzchni ciekłej stali. W rozwiązaniach z ostatnich lat doprowadzenie ciekłego metalu do krystalizatora następuje przy temperaturze nieznacznie przegrzanej, około 10-40K, lub nawet przy temperaturze likwidusu. Wiąże się to z dążeniem do obniżenia temperatury w obszarze krystalizatora a przez to można skrócić długość krystalizatora lub znacznie skrócić strefę chłodzenia wtórnego [3, 5].

3. Modelowanie zjawisk występujących w krystalizatorze COS

Złożoność zjawisk towarzyszących ciągłemu odlewaniu stali szczególnie wlewków wielkogabarytowych powoduje, że doświadczenia uzyskane przez obserwacje procesu odlewania do wlewnic lub form są niewystarczające. Proces tworzenia wlewka ciągłego jest znacznie bardziej złożony niż w przypadku wlewków stacjonarnych, zaś eksperymenty na rzeczywistych urządzeniach wymagają nie tylko pokonania istotnych trudności technicznych, lecz również charakteryzują się znaczną pracochłonnością i dużymi kosztami. Dlatego też, wyniki symulacji komputerowej stanowią tanie i częstokroć jedyne źródło informacji o podstawowych parametrach prowadzenia procesu COS (szczególnie przy odlewaniu nowych gatunków stali) pozwalających otrzymać wlewek o pożądanej jakości.

W ostatnich kilkudziesięciu latach opracowano wiele modeli matematycznych przepływu ciepła w procesie ciągłego odlewania [10-14], jednakże większość z nich mogła być stosowana tylko do symulacji stanów ustalonych. Dają one pole temperatury wlewka jako funkcję następujących parametrów: prędkości odlewania, przegrzania zalewanego metalu, odbierania ciepła przez krystalizator, intensywności natrysku wody chodzącej, gatunku stali i wymiarów wlewka. Analizie podawane są również zjawiska fizyko-chemiczne, na powierzchni górnej ciekłego metalu w krystalizatorze, związane z tworzeniem się żużla na tej powierzchni i jego zasysaniem do wnętrza wlewka [13]. Powoduje to powstawanie wad odlewniczych na ściankach lub w części środkowej wlewka. Wiele prac poświęcono zjawiskom cieplno-mechanicznym na styku wlewek-krystalizator, zwłaszcza tworzeniu się tam szczeliny skurczowej [15]. W literaturze przedmiotu można znaleźć prace, w których poddaje się śledzeniu cząstki zanieczyszczeń i rozważa się wpływ ich rozłożenia we wlewku na jego własności wytrzymałościowe [16]. Pomimo, że wykorzystywane w większości prac modele stanu ustalonego dostarczają ważnej wiedzy dotyczącej działania krystalizatora, to jednak nie mogą one być stosowane do symulacji stanów chwilowych krystalizatora, które występują dosyć często w procesie odlewania (rozruch procesu i sytuacje awaryjne).

W celu lepszego sterowania przepływem ciepła w całym cyklu procesu odlewania większość uwagi skupiona jest ostatnio na rozwoju modeli w czasie rzeczywistym, które są możliwe do wykorzystania w procesie nieustalonym [17-23]. Konieczne jest wtedy rozwiązywanie sprzężonego układu równań złożonego z równań pędu i równania ciągłości przepływu oraz z równania przewodnictwa w postaci ogólnej (z członem konwekcyjnym). Tak postawionemu problemowi poświęcona jest również i ta praca, w której na drodze symulacji numerycznej oceniany jest wpływ wariantu doprowadzenia ciekłego metalu do krystalizatora na pole prędkości i temperatury.

Wyznaczenie pola temperatury wlewka i krystalizatora jest zagadnieniem złożonym. Należy tu uwzględnić ruch wlewka, wydzielanie się ciepła krzepnięcia, nieliniowe warunki brzegowe i warunki ciągłości oraz zmienne z temperaturą własności termofizyczne odlewanego metalu. Rozważania, uwzględniające tylko pionową składową prędkości związaną z ruchem wlewka ku dołowi, stanowiące tzw. model podstawowy można znaleźć w pracach [11, 20]. Problem ten staje się bardziej

skomplikowany, gdy uwzględnia się wpływ ruchu cieczy we wlewku na jego pole temperatury. Takie sformułowanie problemu określa się mianem tzw. modelu złożonego [10, 17-19, 23].

4. Cel pracy

Celem pracy jest przeprowadzenie analizy wpływu sposobu doprowadzenia stopionego metalu do wnęki krystalizatora ciągłego odlewania na pole prędkości i kinetykę narastania fazy stałej wlewka oraz rozkład zanieczyszczeń w początkowych stanach pracy urządzenia COS.

5. Model matematyczny zjawisk cieplno-przepływowych w krystalizatorze ciągłego odlewania

5.1. Założenia przyjęte w modelowaniu procesu COS

W pracy rozważa się krzepnięcie płaskiego wlewka ciągłego w miedzianym krystalizatorze prostym. Z uwagi na zależność rozkładu pól prędkości od sposobu zalewania metalu do krystalizatora rozpatruje się różne możliwości jego doprowadzenia. Oprócz pól temperatury należy jednocześnie określać pola prędkości w fazie ciekłej, kształtujące front krzepnięcia i wpływające na kinetykę krzepnięcia. Zjawiska cieplno-przepływowe w krystalizatorze COS, towarzyszące procesowi krzepnięcia, rozpatruje się jako nieustalone, co pozwala dokonywać analizy chwilowych stanów pracy układu. Aby osiągnąć ten cel należało sformułować model matematyczny i opracować na jego bazie model numeryczny w metodzie elementów skończonych. W modelu matematycznym, przyjęto że ciekły metal ma cechy płynu lepkiego nieściśliwego i przewodzącego ciepło. Założono, że front krzepnięcia jest rozmyty, tzn. ciekły metal krzepnie w przedziale temperatur likwidus/solidus. Taki model krzepnięcia jest najczęściej stosowany, gdy rozważa się krzepnięcie stopów metali (stali lub staliwa). Uwzględniano zmianę parametrów termofizycznych od temperatury i od udziału fazy stałej w dwufazowej strefie przejściowej.

Przy konstrukcji algorytmu numerycznego, symulującego proces krzepnięcia i stygnięcia wlewka, zastosowano modyfikację opisu matematycznego tego procesu, która nie wymaga dokładnego określania położeń izoterm solidusu i likwidusu, przez co algorytm staje się dogodny i efektywny w stosowaniu. Założono ciągłość parametrów termofizycznych podobszarów (ciepła właściwego, gęstości, lepkości i współczynnika przewodzenia ciepła) w analizowanym przedziale temperatury. Rozpatruje się zatem pewien homogeniczny obszar o określonych parametrach termofizycznych i zamiast układu równań określających zjawiska cieplne w cieczy, strefie dwufazowej i części zakrzepłej mamy jedno równanie przewodnictwa obowiązujące dla całego układu.

5.2. Układ równań stanowiący model matematyczny

Proponowany w pracy model matematyczny transportu ciepła, z uwzględnieniem procesu wypełniania wnęki krystalizatora COS, opiera się na rozwiązywaniu następującego układu równań różniczkowych [17-23]:

- równania energii

$$\rho C_{ef}\left(\frac{\partial T}{\partial t}+v_j\frac{\partial T}{\partial x_j}\right)=\frac{\partial}{\partial x_j}\left(\lambda\frac{\partial T}{\partial x_j}+\frac{c\mu_t}{\sigma_t}\frac{\partial T}{\partial x_j}\right) \tag{1}$$

- równań pędu

$$\rho\left(\frac{\partial v_i}{\partial t}+v_j\frac{\partial v_i}{\partial x_j}\right)=\frac{\partial}{\partial x_j}\left((\mu+\mu_t)\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right)-\frac{2}{3}\rho k\delta_{ij}\right)-\frac{\partial p}{\partial x_i}+\rho g_i \tag{2}$$

- równania ciągłości przepływu

$$\frac{\partial v_j}{\partial x_j}=0 \tag{3}$$

Występujący w (1) i (2) dynamiczny współczynnik lepkości turbulentnej (μ_t) określany jest zależnością

$$\mu_t=c_\mu\rho\frac{k^2}{\varepsilon} \tag{4}$$

Wielkości k i ε występujące w powyższym wzorze obliczane są z wykorzystaniem równań [22]

- równania kinetycznej energii turbulencji

$$\rho\left(\frac{\partial k}{\partial t}+v_j\frac{\partial k}{\partial x_j}\right)=\frac{\partial}{\partial x_j}\left((\mu+\frac{\mu_t}{\sigma_k})\frac{\partial k}{\partial x_j}\right)+\mu_t\frac{\partial v_i}{\partial x_j}\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right)-\frac{\mu_t}{\rho\sigma_\rho}\frac{\partial\rho}{\partial x_i}g_i-\rho\varepsilon \tag{5}$$

- równania prędkości dyssypacji energii turbulencji

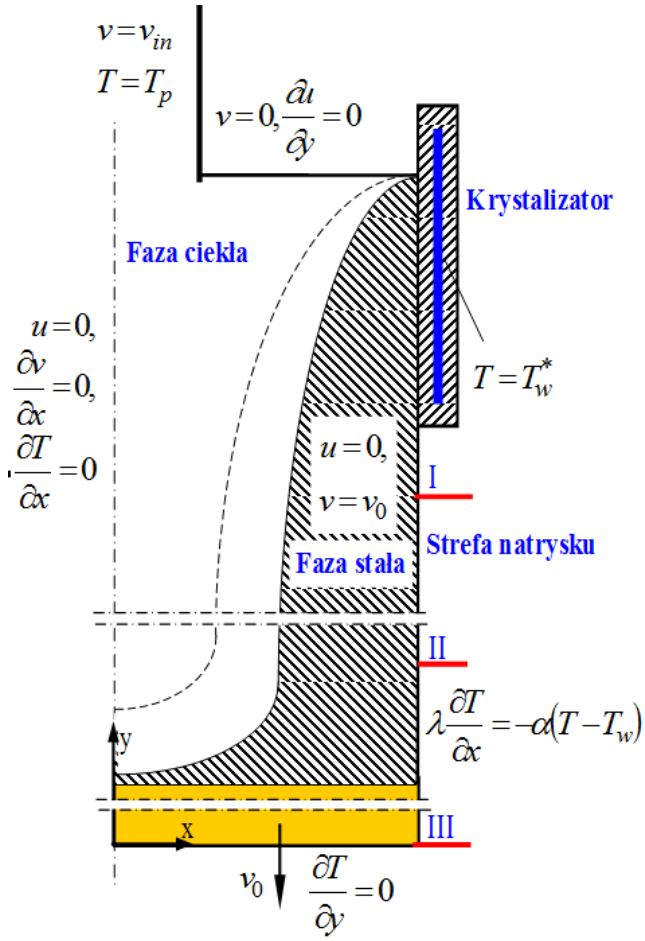
$$\rho\left(\frac{\partial \varepsilon}{\partial t}+v_j\frac{\partial \varepsilon}{\partial x_j}\right)=\frac{\partial}{\partial x_j}\left((\mu+\frac{\mu_t}{\sigma_\varepsilon})\frac{\partial \varepsilon}{\partial x_j}\right)+c_1\frac{\varepsilon}{k}\mu_t\frac{\partial v_i}{\partial x_j}\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right)-c_1(1-c_3)\frac{\varepsilon}{k}\frac{\mu_t}{\rho\sigma_\rho}\frac{\partial\rho}{\partial x_i}g_i-c_2\rho\frac{\varepsilon^2}{k} \tag{6}$$

gdzie: T – temperatura [K], t – czas [s], λ – współczynnik przewodzenia ciepła [W/(mK)], $\rho=\rho(T)$ – gęstość [kg/m^3], v_j – wektor prędkości przepływu metalu [m/s], $C_{ef}(T)=c_{LS}+L/(T_L-T_S)$ – efektywne ciepło właściwe strefy przejściowej [J/(kgK)] [17], L – ciepło krzepnięcia [J/kg], c_{LS} – ciepło właściwe strefy dwufazowej [J/(kgK)], $\mu(T)$ – dynamiczny współczynnik lepkości [Ns/m^2], p – ciśnienie [N/m^2], μ_t – dynamiczny współczynnik lepkości turbulentnej [Ns/m^2], g_i – wektor przyspieszenia [m/s^2], c – ciepło właściwe [J/(kgK)], k – kinetyczna energia turbulencji [m^2/s^2], ε - prędkość dyssypacji energii turbulencji [m^2/s^3], x_j – wektor położenia rozważanego punktu [m], c_μ = 0.09, c_1 = 1.44, c_2 = 1.92, c_3 = 0.8, σ_k = 1, σ_t = 0.9, σ_ε = 1.3, σ_p = 0.9 – stałe empiryczne [22].

5.3. Warunki brzegowe i początkowe

Powyższy układ równań uzupełniono warunkami początkowymi przyjętymi w odniesieniu do pól prędkości i temperatury [5, 17-19] oraz warunkami brzegowymi założonymi na zaznaczonych powierzchniach (rys.1) [9-22]. Warunki początkowe określają stan fizyczny układu (płynu) w pewnej chwili $t=t_0$, obranej za początkową. Na wejściu do krystalizatora zastosowano warunki brzegowe I rodzaju (Dirychleta). W osi symetrii i na płaszczyznach ograniczających rozpatrywany obszar założono warunki brzegowe II rodzaju (Neumana). Odprowadzenie ciepła z powierzchni wlewka przeprowadzono zgodnie z warunkiem III rodzaju (Newtona).

Postawione zadanie rozwiązano metodą elementów skończonych w sformułowaniu reszt ważonych [8-10, 12-19]. Podobszary rozważanego układu zdyskretyzowano siatką czterowęzłowych prostokątnych elementów skończonych. Do aproksymacji powyższych równań zastosowano metodę Petrowa-Galerkina z funkcjami bazowymi przesuwającymi punkty całkowania ("upwind function"), która zapewniła stabilność obliczeń numerycznych przy wymuszonej, ograniczeniami pamięciowymi komputera, ustalonej gęstości siatki elementów skończonych [17-19]. Agregacja równań rozpisanych dla poszczególnych elementów daje globalny układ równań, do którego wprowadza się warunki brzegowe i początkowe. Rozwiązanie takiego układu równań daje poszukiwane wartości funkcji w węzłach siatki elementów skończonych, dyskretyzujących przyjęty do rozważań obszar.



Rys. 2. Rozważany układ i warunki brzegowe przyjęte w rozwiązywanym zadaniu

6. Obliczenia numeryczne

Symulacje komputerowe procesu krzepnięcia stali w krystalizatorze COS przeprowadzono przy wykorzystaniu tzw. modelu złożonego dla następujących wariantów doprowadzenia ciekłego metalu do krystalizatora:

- zalewano metal wylewem umieszczonym centralnie, pionowy wypływ metalu (tzw. zalewanie centralne-pionowe),
- zalewano metal wylewem zanurzonym umieszczonym centralnie, poziomy wypływ metalu (tzw. zalewanie centralne-poziome).

Badano w ten sposób wpływ wariantu zalewania metalu do krystalizatora na rozkład prędkości w fazie ciekłej i kinetykę narastania fazy stałej wlewka ciągłego (tworzenie się naskórka). Wymuszone, przez odpowiednie doprowadzenie metalu do krystalizatora, ruchy ciekłego metalu nie są bez wpływu na jakość wlewka, związaną z rozkładem zanieczyszczeń we wlewku COS, co wykazali autorzy prac [16-18, 24].

6.1. Warunki przeprowadzenia symulacji komputerowych

Obliczenia wykonano dla układu wlewek-krystalizator o wymiarach wewnętrznych krystalizatora 0.25x1x0.7[m] i długości kontrolnej tworzącego się wlewka wynoszącej 2.9[m]. Symulacje numeryczne, odlewania stali na urządzeniu COS, przeprowadzono dla prędkości odlewania $v_0 = 0.01$[m/s]. Dla założonej prędkości przesuwu wlewka (v_0) z bilansu masy obliczono prędkość zalewania ciekłego metalu (v_{in}). Dla obu przypadków zalewania zachowano równość wydatków oraz zastosowano takie same cieplne warunki brzegowe.

Własności termofizyczne stali dla poszczególnych podobszarów rozpatrywanego układu zaczerpnięto z prac [14-19, 21-25]. Założono liniową zmianę gęstości (ρ) i współczynnika przewodzenia ciepła (λ) w przedziale temperatur krzepnięcia stali T_L–T_S. Zmiana współczynnika lepkości dynamicznej (μ) z temperaturą, określana według zależności wykładniczej w zakresie wartości 0.003 – 0.1[Ns/m^2], obowiązywała do wartości 0.9 udziału fazy stałej (Φ) w strefie dwufazowej. Powyżej tej wartości Φ, lepkość nagle wzrasta do dużych wartości. Dla zachowania stabilności obliczeń przyjęto $\mu_S = 10^5$[Ns/m^2]. Uwzględniono w ten sposób, znikomy ruch metalu (nawet jego brak) w pobliżu linii solidusu. Ciepło krzepnięcia było równe L=272[kJ/kg]. Przyjęto następujące wartości temperatury dla cieczy metalicznej: T_p=1850[K], T_L =1800[K], T_S =1760[K] i wody chłodzącej T^*_w =293 [K], T_w =300 [K]. Współczynnik przejmowania ciepła od krystalizatora do otoczenia wynosił α_k=100[W/(m^{2}K)], natomiast współczynnik przejmowania ciepła od wlewka do wody chłodzącej zmieniano po długości wlewka w zakresie wartości α =1100-750[W/(m^{2}K)].

W pierwszym etapie obliczeń przeprowadzono symulacje procesu wypełniania krystalizatora ciekłym metalem, obserwując grubości narosłej fazy stałej na ściankach krystalizatora. W etapie następnym wprawiano wlewek w ruch łącznie z drągiem startowym. Od rozpoczęcia ruchu drąga startowego ku dołowi, do momentu opuszczenia przez drąg startowy przyjętej do rozważań długości kontrolnej, ruch wlewka śledzono we współrzędnych Lagrange'a a ruchy ciekłego metalu we współrzędnych Eulera. Na podstawie analizy prac [10, 22], w których rozważane są różne sposoby opisu ruchów wlewka ciągłego, takie podejście wydaje się być najbardziej poprawne.

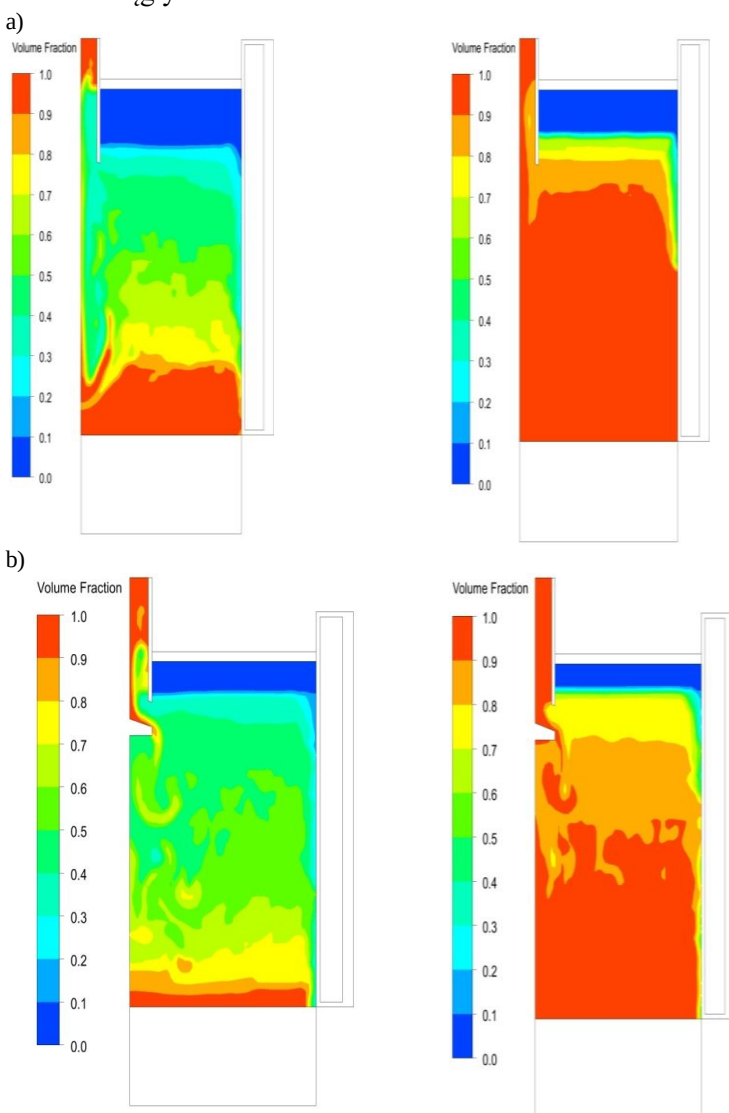
6.2. Wyniki symulacji numerycznych

Wybrane wyniki symulacji numerycznych, procesu wypełniania ciekłym metalem wnęki krystalizatora, w postaci pól prędkości, temperatury i stanu wypełnienia krystalizatora przedstawiono na rysunkach 3-5. Dzięki wykorzystaniu metody śledzenia położenia powierzchni swobodnej VOF, możliwe było zdefiniowanie dwóch faz (powietrza i ciekłego metalu) oraz analizowanie ich aktualnego wzajemnego położenia w procesie wypełniania wnęki formy ciekłym metalem, co pokazano na rysunku 3. Rozkłady wektorów prędkości i pola temperatury w wybranych krokach czasu pokazano, w zależności od wariantu zalewania metalu do krystalizatora, na rysunkach 4 i 5.

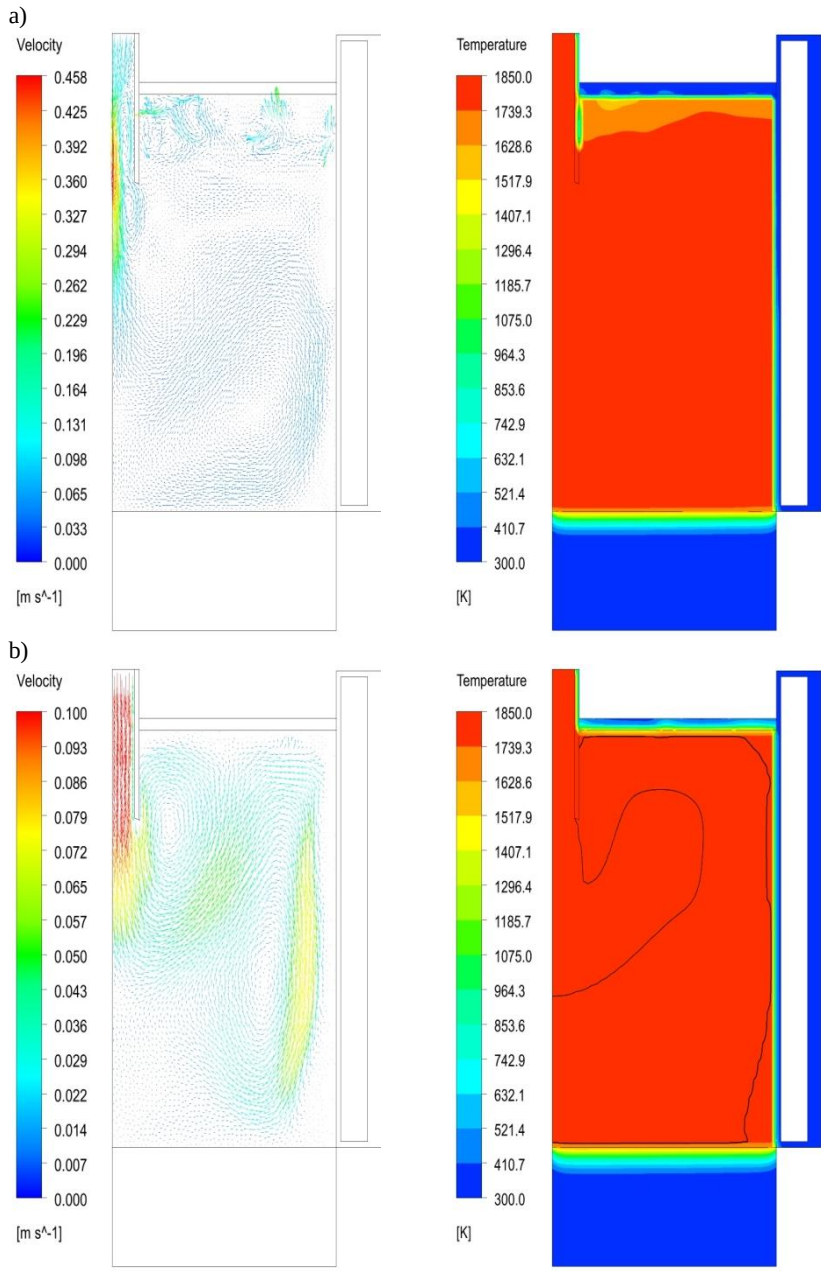
Zauważa się również wyraźne różnice w rozkładzie wektorów prędkości, w zależności od wariantu zalewania metalu do krystalizatora, podczas ruchu wlewka ku dołowi (rys. 6-7). Pokazują one kolejne momenty, gdy czoło ciekłego metal znajdujące

się nad przesuwającym się drągiem startowym mija wyróżnione przekroje (I, II, III) zaznaczone na rysunku 2 i określone, odległością (l) odmierzaną od powierzchni górnej ciekłego metalu w krystalizatorze do danego przekroju, następująco: I - l_1 = 0.9m, II - l_2 = 1.9, III - l_3 = 2.9m. Podane zatem wymiary (l) odpowiadają utworzonej długości wlewka, przypadającej określonej chwili czasu.

Obserwując zmiany pól prędkości i temperatury w fazie ciekłej wlewka można dokonać oceny ilości narosłej fazy stałej w kolejnych krokach czasu a widoczne różnice w charakterze ruchów metalu nie są bez znaczenia dla rozkładu zanieczyszczeń we wlewku ciągłym.

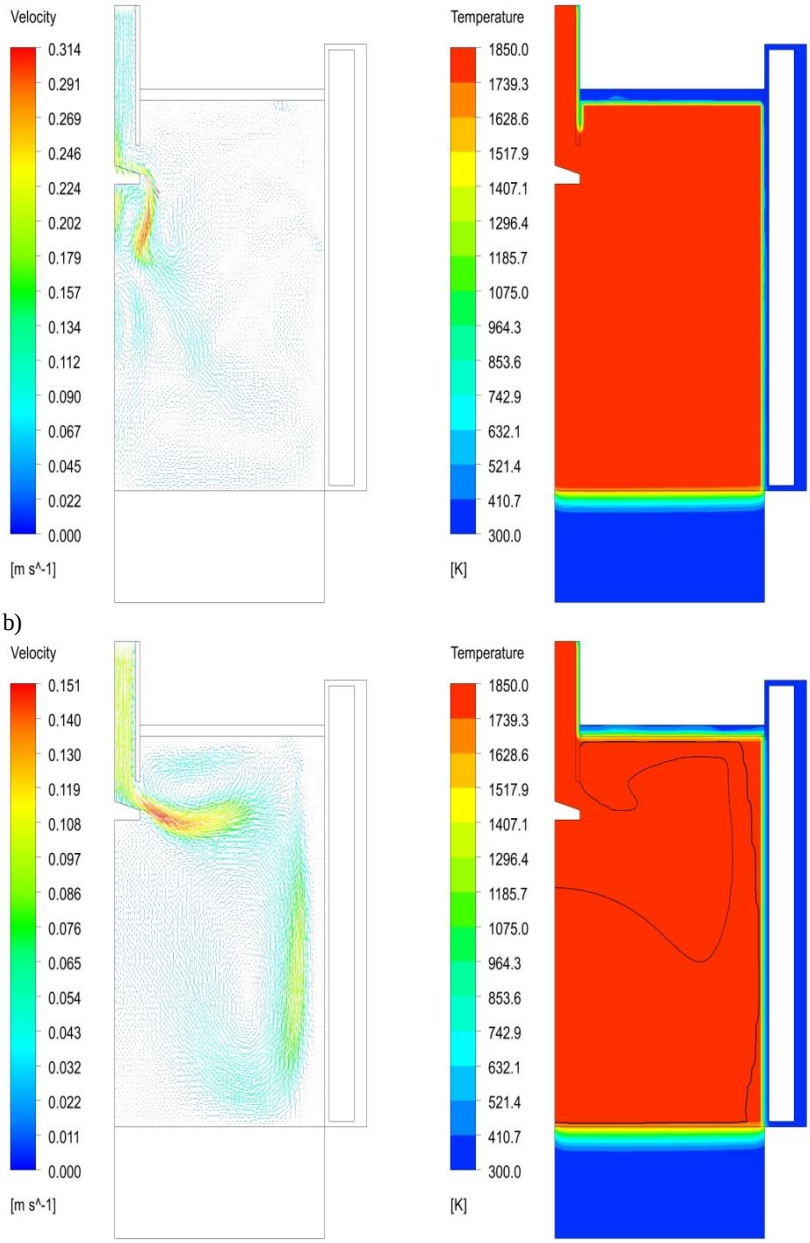


Rys. 3. Stan wypełnienia wnęki krystalizatora po czasie t =40 i 60s: a) zalewanie centralne-pionowe, b) zalewanie centralne-poziome

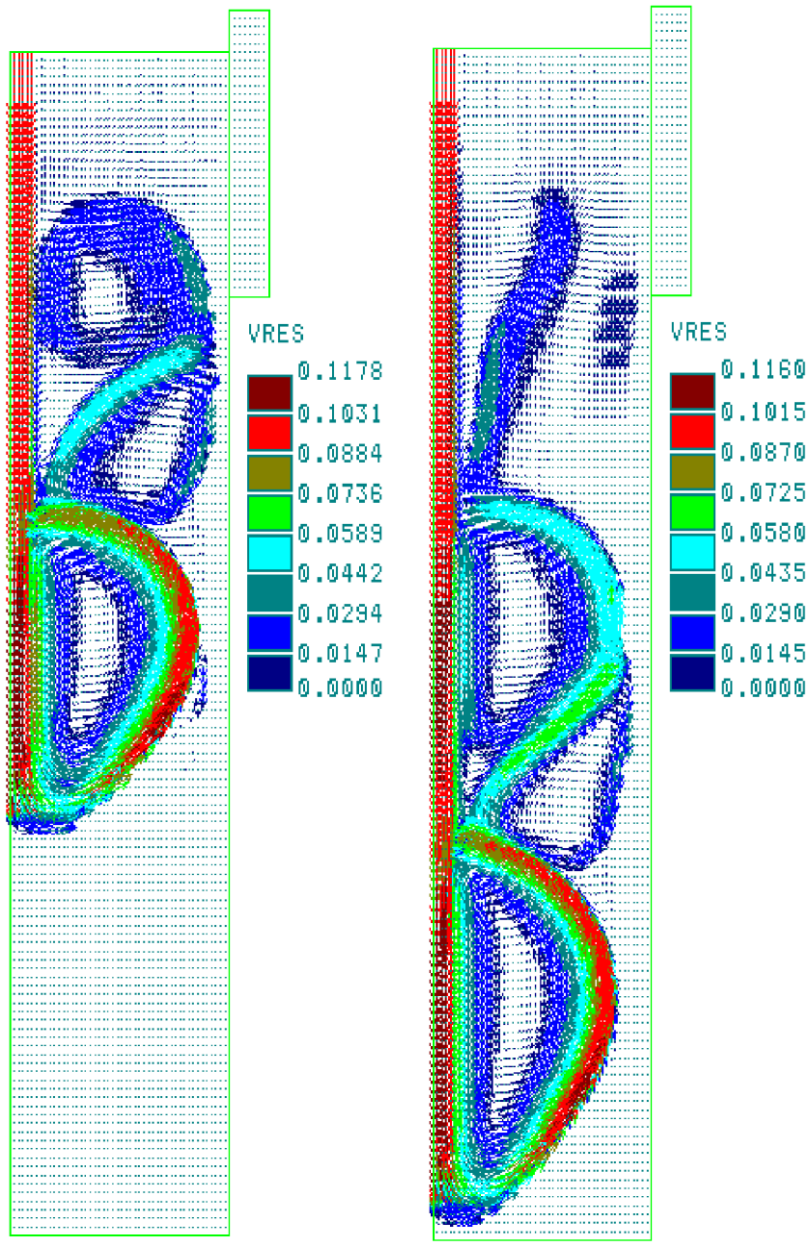


Rys. 4. Wektory prędkości i pole temperatury po czasie: a) *t*=60s, b) *t*=75s, zalewanie centralne-pionowe

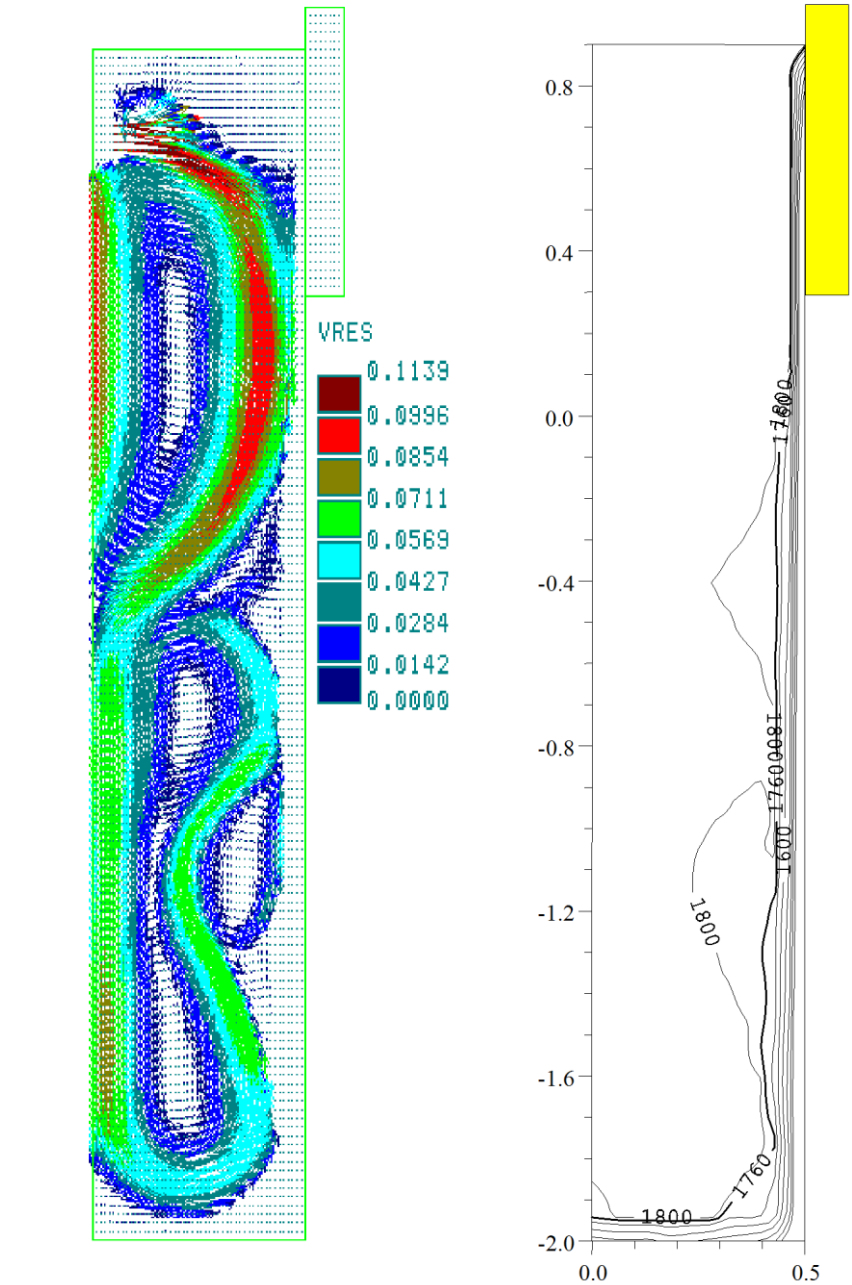
a)



Rys. 5. Wektory prędkości i pole temperatury po czasie: a) *t*=60s, b) *t*=75s, zalewanie centralne-poziome



Rys. 6. Wektory prędkości [m/s] po osiągnięciu przez ciekły metal poziomu: a) II (t=200s), b) III (t=300s), zalewanie centralne-pionowe



Rys. 7. Wektory prędkości [m/s] i pole temperatury [K] po osiągnięciu przez ciekły metal poziomu III (t=300s), zalewanie centralne-poziome

7. Podsumowanie

W pracy podano propozycję opisu krzepnięcia ujmującego wzajemny wpływu zjawisk cieplnych i przepływowych w procesie krzepnięcia wlewka ciągłego odlewania. Obliczenia prowadzono od początku zalewania krystalizatora ciekłym metalem do momentu opuszczenia przez drąg startowy rozważanego obszaru kontrolnego wlewka. Badano wpływ pola prędkości w fazie ciekłej na pole temperatury i narastanie fazy stałej wlewka ciągłego w kolejnych etapach jego ruchu. Na podstawie przeprowadzonych symulacji numerycznych można stwierdzić, że zakrzepła warstwa (naskórek) wykazuje wyraźną nierównomierność grubości na długości wlewka spowodowaną ruchem cieczy, która wymywa jego fazę przejściową (rys.7). W często stosowanych modelach, bez uwzględniania ruchów fazy ciekłej, nie zauważa się tego zjawiska. Faza stała narasta wtedy regularnie na całej długości wlewka. W obrębie krystalizatora zauważa się dość szybkie ustalenie się procesów cieplnych. Grubość naskórka na poziomie dołu krystalizatora ustala się po jego wypełnieniu na około 20[mm] (rys. 4b i 5b), co pokazuje zaznaczona na tych rysunkach linia solidusu. Porównując pola temperatury, otrzymane z symulacji krzepnięcia wlewka ciągłego przy różnych wariantach doprowadzenia ciekłego metalu do krystalizatora, nie zauważono istotnych różnic w grubości narosłej fazy stałej. Występujące natomiast różnice w charakterze ruchów metalu nie są bez znaczenia dla rozkładu zanieczyszczeń we wlewku ciągłym, a zatem wpływają na stan powierzchni i własności wytrzymałościowe wlewka.

Literatura

1. Perzyk M., Waszkiewicz S., Kaczorowski M., Jopkiewicz A. *Odlewnictwo*, Wydawnictwo Naukowo Techniczne, Warszawa 2004.
2. Pater Z. *Podstawy metalurgii i odlewnictwa*, Wydawnictwo Politechniki Lubelskiej, Lublin 2014.
3. Szweycer M., Nagolska D. *Metalurgia i odlewnictwo*, Wydawnictwo Politechniki Poznańskiej, Poznań 2002.
4. Kozakowski S. *Badanie odlewów – technologie odlewnicze, typowe dla nich wady i metody ich ujawniania*, Wydawnictwo Gamma, Warszawa 2001.
5. Kudliński Z. *Technologie odlewania stali*, Wydawnictwo Politechniki Śląskiej, Gliwice 2006.
6. Gzieło A. *System zapewniania jakości wlewków ciągłych płaskich stali węglowej*, Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2006.
7. Sorek A., Ostrowska-Popielska P. *Rola zasypek krystalizatorowych w procesie ciągłego odlewania stali*, Prace IMŻ, 2 (2012), s. 18-22.
8. Real-Ramirez C.A., Miranda-Tello R., Hoyos-Reyes L., Reyes M., Gonzales-Trejo J.I. *Numerical evaluation of submerged entry nozzle for continuous casting of steel*, Journal of Engineering & Materials Sciences, 19 (2012), s. 179-188.
9. Lamut J., Falkus J., Jurjevec B., Knap M. *Influence of inclusions modification of nozzle clogging*, Archives of Metallurgy and Materials, 57 (2012), s. 319-324.
10. Choudhary S.K., Mazumdar D. Mathematical modelling of transport phenomena in continuous casting of steel, ISIJ International, 34 (1994), s. 584-592.
11. Lally B., Biegler L., Henein H., *Finite difference heat-transfer modeling for continuous casting*, Metallurgical Transactions B, 21 (1990), s. 761-770.
12. Huang X., Thomas B.G., Najjar F.M. *Modeling superheat removal during continuous casting of steel slabs*, Metallurgical Transactions B, 23 (1992), s. 339-356.

13. McDavid R.M., Thomas B.G. *Flow and thermal behavior of the top surface flux/powder layers in continuous casting molds*, Metallurgical and Materials Transactions B, 27 (1996), s. 672-685.
14. T. Telejko, Z. Malinowski, M. Rywotycki, *Analysis of heat transfer and fluid flow in continuous steel casting*, Archives of Metallurgy and Materials, 54, (2009), s. 837-844 .
15. Bokota A., Parkitny R. *Model powstawania szczeliny skurczowej pomiędzy krystalizatorem a wlewkiem ciągłym*, Krzepnięcie Metali i Stopów, 30 (1997), s. 13-24.
16. Ho H-Y., Chen Ch.H, Hwang W.S., *Analysis of molten steel flow in slab continuous caster mould*, ISIJ International, 34 (1994), s. 255-264.
17. Bokota, L. Sowa, *Numerical modelling of solidification of the cast slab at the beginning of continuous casting process*, Acta Metallurgica Slovaca, 8 (2002), s. 245-249 .
18. Zhao, B.G. Thomas, S.P. Vanka, R.J. Omalley, *Transient fluid flow and superheat transport in continuous casting of steel slabs*, Metallurgical and Materials Transactions, 36B (2005), s. 801-823 .
19. Sowa L., Bokota A. *Numerical model of thermal and flow phenomena the process growing of the CC slab*, Metallurgy and Materials, 56 (2011), s. 359-366.
20. Louhenkilpi S., Nieminen R. *Real-time simulation of heat transfer in continuous casting*, Metallurgical Transactions B, 24B (1993), s. 685-693.
21. Choudhary S.K., Mazumdar D., Ghosh A. *Mathematical modelling of heat transfer phenomena in continuous casting of steel*, ISIJ International, 33 (1993), s. 764-774.
22. An G., Sun X. *Turbulent fluid flow and heat transfer calculation in mold filling and solidification processes of casting*, Shenyang Research Institute of Foundry, Shenyang 2000.
23. Zhang L., Wang Y. *Transient Fluid Flow Phenomena during Continuous Casting: Part I – Cast Start*, ISIJ International, 50 (2010), s. 1777-1782.
24. Zhang L., Wang Y., Zuo X. *Flow Transport and Inclusion Motion in Steel Continuous-Casting Mold under Submerged Entry Nozzle Clogging Condition*, Metallurgical and Materials Transactions B, 39B (2008), s. 534-550.
25. Piekarska W., Kubiak M., Saternus Z., Domański T., Stan S., Radcenko M., Ivanov S. *Numerical modeling of thermal phenomena in Yb: YAG laser welding process*, Journal of Applied Mathematics and Computional Mechanics, 13 (2014), s. 175-186.

Wykorzystanie symulacji komputerowych do modelowania zjawisk cieplno-przepływowych procesu krzepnięcia wlewka COS

Streszczenie

W pracy przedstawiono model matematyczny i wyniki symulacji numerycznych przepływu ciekłej stali przez wylew zanurzony do krystalizatora ciągłego odlewania z uwzględnieniem procesu jego wypełniania. Zadanie potraktowano kompleksowo rozwiązując go metodą elementów skończonych. Symulacje numeryczne krzepnięcia wlewka ciągłego odlewania, z uwzględnieniem procesu wypełniania krystalizatora, wykonano dla różnych wariantów doprowadzenia ciekłego metalu. Badano w ten sposób wpływ sposobu zalewania krystalizatora na pole prędkości w fazie ciekłej i kinetykę narastania fazy stałej wlewka, które mają istotny wpływ na jakość wlewka ciągłego odlewania stali.

Słowa kluczowe: symulacje numeryczne, proces COS, przepływ ciekłego metalu

Application of computer simulation to modeling the thermal-flow phenomena of solidification process of the slab CSC

Abstract

The mathematical model and numerical simulations of the molten steel flow by the submerged entry nozzle and the filling process of the continuous casting mold cavity are presented in this paper. The problem was treated as a complex and solved by the finite element method. Numerical simulations of the solidification of the continuous cast slab including the filling process of the continuous casting mold cavity was performed for different variants of the pouring of liquid metal. The influences of cases of the continuous casting mould pouring on the velocity fields in liquid phase and the solid phase growth kinetics of the cast slab were estimated, because these magnitudes have essential an influence on high-quality of a continuous steel cast slab.

Keywords: numerical simulation, CSC process, molten metal flow

Anna Małafiejska[1], Michał Małafiejski[2], Krzysztof Ocetkiewicz[3], Krzysztof Pastuszak[4]

Szeregowanie zadań dwuprocesorowych w systemach otwartych

1. Wprowadzenie

Problemy szeregowania zadań należą do głównych gałęzi dziedziny badań operacyjnych. Klasycznymi przykładami zastosowań są układanie harmonogramów napraw pojazdów mechanicznych na masową skalę [1] i układanie planów zajęć dla szkół i uczelni [2].

W problemach szeregowania zadań często dane są zbiory maszyn (procesorów, stacji roboczych), zwykle oznaczane przez M i zbiory zadań J. Zadania mogą składać się z operacji, których zbiór oznaczany jest przez O.

W przypadku szeregowania zadań na maszynach dedykowanych, każde zadanie jest przypisane do konkretnej maszyny, na której może zostać wykonane.

Problemy szeregowania zadania często opisywane są za pomocą notacji trójpolowej, wprowadzonej po raz pierwszy przez Grahama w 1982 roku [3]. W notacji trójpolowej, problem szeregowania zadań reprezentowany jest za pomocą trzech pól, $\alpha \setminus \beta \setminus \gamma$:

- w polu α przedstawiona jest charakterystyka zbioru maszyn,
- pole β opisuje wykonywane zadania,
- pole γ zawiera kryterium optymalizacyjne.

Rozpatrywane w pracy kryteria optymalizacyjne to:

- istnienie harmonogramu, opisywane przez "-" w polu gamma,
- minimalizacja całkowitej długości harmonogramu – $C_{max,}$
- minimalizacja łącznego (średniego) czasu wykonywania zadań lub operacji – ΣC_i.

W problemach szeregowania zadań możliwe jest wprowadzenie dodatkowych ograniczeń, umożliwiających lepsze zamodelowanie rzeczywistych problemów. Występujące ograniczenia możemy podzielić na:

- restrykcje dotyczące instancji np. każde zadanie składa się z dokładnie dwóch operacji,
- restrykcje ograniczające przestrzeń rozwiązań, np. w modelu no-wait przyjęte może być jedynie rozwiązanie, w którym każda maszyna pracuje bez przestojów [4].

[1] ania@animima.org, Katedra Rachunku Prawdopodobieństwa i Biomatematyki, Wydział Fizyki Technicznej i Matematyki Stosowanej, Politechnika Gdańska
[2] michal@animima.org, Katedra Algorytmów i Modelowania Systemów, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
[3] Krzysztof.Ocetkiewicz@eti.pg.edu.pl, Katedra Algorytmów i Modelowania Systemów, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
[4] krzysztof_pastuszak@yahoo.com, Katedra Algorytmów i Modelowania Systemów, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska

1.1. Zadania dwuprocesorowe

Rozpatrzmy sytuację, w której każde zadanie korzysta z dwóch procesorów – pierwszego w trybie odczytu (*read*), a drugiego w trybie zapisu (*write*). Operacja odczytu wymaga jedynie niewielkiego kwantu czasu procesora, możliwe jest zatem równoległe wykonanie wielu operacji odczytu na tej samej maszynie w jednej jednostce czasowej. Wykonanie operacji zapisu zużywa cały czas pracy procesora w danej jednostce czasowej, jest zatem operacją blokującą. Nie jest możliwe równoległe wykonanie innych operacji na maszynie, na której realizowana jest operacja zapisu.

Z uwagi na przyjęty w szeregowaniu chromatycznym model grafowy, zakładamy, że każde zadanie jest symetryczne - jeżeli istnieje zadanie J_j, złożone z operacji zapisu na maszynie M_i i operacji odczytu na maszynie M_{i*}, to istnieje też zadanie J_{j*}, złożone z operacji zapisu na maszynie M_{i*} i operacji odczytu na maszynie M_i.

Zakładamy, że każda operacja zapisu wymaga identycznego czasu przetwarzania. Rozpatrujemy model bez wywłaszczania - nie jest możliwe przerwanie wykonywania zadania.

Potencjalne zastosowania związane są z serwerami i systemami bazodanowymi, np. w przypadku testowania połączeń pomiędzy stacjami roboczymi lub uwierzytelniania kanałów komunikacyjnych.

2. Konstrukcja harmonogramów

W problemach szeregowania zadań celem jest konstrukcja harmonogramu spełniającego wszystkie narzucone restrykcje, przy jednoczesnej optymalizacji wybranej funkcji celu. W każdym poprawnym harmonogramie wszystkie zadania muszą zostać w pełni wykonane.

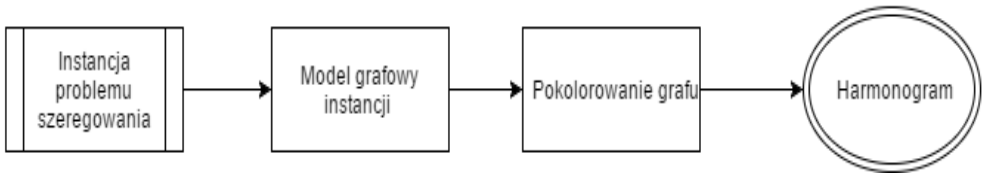
Wiele problemów szeregowania zadań jest trudnych obliczeniowo [4, 5], zwłaszcza po wprowadzeniu dodatkowych restrykcji ograniczających przestrzeń dopuszczalnych rozwiązań. Do konstrukcji legalnych harmonogramów wykorzystywane są wykorzystywane algorytmy aproksymacyjne [6], programowanie dynamiczne [7] i algorytmy sztucznej inteligencji, np. [8].

2.1. Szeregowanie chromatyczne

Przy konstrukcji harmonogramów można stosować metodę szeregowania chromatycznego, umożliwiającą, z uwagi na skuteczne narzędzia (dobrze zbadane modele, efektywne algorytmy), tworzenie dokładnych algorytmów układania harmonogramów [4,9]. Główną ideą szeregowania chromatycznego jest odtworzenie harmonogramu z pokolorowania grafu modelującego instancję problemu szeregowania zadań. W pierwszym kroku, na podstawie danej instancji problemu, konstruowany jest model grafowy. W przypadku szeregowania na maszynach dedykowanych często stosowana jest reprezentacja w postaci grafu dwudzielnego, z partycją A odpowiadającą maszynom, i partycją B, odpowiadającą zadaniom. Przypisanie zadań do maszyn reprezentowane jest za pomocą krawędzi łączących odpowiednie wierzchołki z partycji A i B. Wybór modelu kolorowania zależy od rozpatrywanego problemu szeregowania zadań.

Po konstrukcji prawidłowego pokolorowania otrzymanego grafu, budowany jest harmonogram. Kolory przypisane poszczególnym strukturom w grafie odpowiadają

oknom czasowym, w których wykonane zostaną reprezentowane zadania. Na ogół, kolor k odpowiada oknu czasowemu [k-1, k). Rysunek 1. przedstawia schemat ideowy szeregowania chromatycznego.



Rysunek 1. Schemat ideowy szeregowania chromatycznego

3. Kolorowanie końcówkowe grafów

Niech dany jest graf prosty, G = (V, E). Końcówką nazywamy parę uporządkowaną (v,e), v ∈ V, e ∈ E, gdzie wierzchołek v jest jednym z końców krawędzi e. Zbiór wszystkich końcówek w grafie G oznaczamy jako I(G).

Definicja 1

Końcówki (v,e) i (u,f), v,u ∈ V, e,f ∈ E sąsiadują ze sobą (są sąsiednie), jeżeli jest spełniony jeden z poniższych warunków:

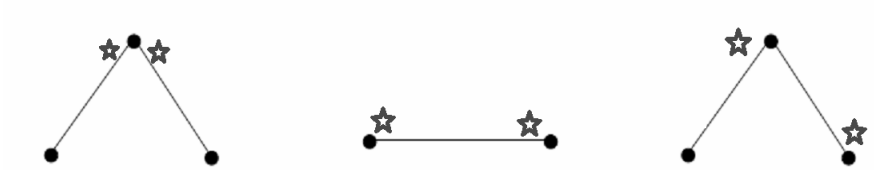
- v = u , e ≠ f;
- e = f, v ≠ u;
- e = {v,w}, f = {w,u} i v ≠ w;

Relacja sąsiedztwa końcówek przedstawiona została na Rysunku 1. Wewnętrzną końcówką względem wierzchołka v nazywamy końcówkę (v, {v,u}). Końcówkę (u, {v,u}) nazywamy zewnętrzną względem wierzchołka v. Końcówka (u, {v,u}) leży na krawędzi {v,u} [10].

Definicja 2

Funkcję c: I(G) → N^+ nazywamy kolorowaniem końcówkowym grafu G, jeżeli dla każdej pary końcówek (v,e) i (u,f) zachodzi c((v,e)) ≠ c((u,f)) [10].

- v = u , e ≠ f;
- e = f, v ≠ u;
- e = {v,w}, f = {w,u} i v ≠ w;



Rysunek 2. Sąsiedztwo końcówek. Sąsiednie końcówki zostały zaznaczone gwiazdkami

Końcówkowe kolorowanie c wykorzystujące k kolorów nazywamy końcówkowym k-kolorowaniem [10]. Końcówkową liczbą chromatyczną grafu G nazywamy

najmniejszą liczbę k, dla której istnieje poprawne końcówkowe k-kolorowanie [10]. Końcówkową liczbę chromatyczną oznaczamy jako χ_i (G).

3.1. Szeregowanie chromatyczne zadań dwuprocesorowych

Niech dana jest instancja problemu szeregowania zadań dwuprocesorowych ze zbiorami maszyn M i zadań J. Aby skonstruować model grafowy instancji, należy skonstruować graf G zawierający |M| wierzchołków. Każdy wierzchołek odpowiada jednej maszynie. Każda para zadań J_{j1} i J_{j2}, gdzie zadanie J_{j1} składa się z operacji odczytu na maszynie M_{i1} i operacji zapisu na maszynie M_{i2}, a zadanie J_{j2} z operacji odczytu na maszynie M_{i2} i operacji zapisu na maszynie M_{i1}, reprezentowana jest przez krawędź łączącą wierzchołki odpowiadające maszynom M_{i1} i M_{i2}. Każde zadanie reprezentowane jest przez pojedynczą końcówkę.

3.2. Szeregowanie zadań dwuprocesorowych z minimalizacją długości harmonogramu

Dla grafów ogólnych, problem kolorowania końcówkowego minimalizującego liczbę użytych kolorów jest trudny [10]. Dolne oszacowanie końcówkowej liczby chromatycznej dla grafów wynosi Δ +1 [10]. Jedyna znana klasa grafów, dla której do konstrukcji legalnego pokolorowania końcówkowego potrzebne jest więcej niż $\Delta + 2$ kolorów, to grafy Paleya, najmniejszy znany graf wymagający więcej niż $\Delta + 2$ kolorów ma 1091 wierzchołków [11]. Górne oszacowanie końcówkowej liczby chromatycznej to $\Delta + O(\log(\Delta))$ [11].

Problem istnienia $\Delta + 1$ – kolorowania końcówkowego jest NP-zupełny dla grafów kubicznych, semi-kubicznych i kubicznych dwudzielnych [11].

Do grafów $\Delta + 1$- końcówkowo kolorowalnych zaliczają się m.in. drzewa [12], cykle zawierające liczbę wierzchołków będącą wielokrotnością 3 [13] i grafy zewnętrznie planarne, dla których $\Delta \geqslant 7$ [14]. Przykładem klasy grafów wymagających co najmniej $\Delta + 2$ kolorów są grafy pełne dwudzielne (z wyjątkiem gwiazd) [12, 15].

4. Szeregowanie zadań dwuprocesorowych z minimalizacją łącznego czasu wykonywania zadań

Problem szeregowania zadań dwuprocesorowych z minimalizacją łącznego czasu wykonywania zadań i rozpatrywanymi w pracy restrykcjami może zostać rozwiązany przy pomocy szeregowania chromatycznego. Instancję problemu szeregowania zadań modelujemy grafowo w sposób opisany w podrozdziale 3.1. Wybrany model kolorowania to sumacyjne kolorowanie końcówkowe grafu, z minimalizacją sumy użytych kolorów. Według wiedzy autorów, model kolorowania końcówkowego z kryterium sumacyjnym nie był wcześniej rozpatrywany w literaturze anglojęzycznej.

Twierdzenie 1

Każdą ścieżkę P można pokolorować sumacyjnie końcówkowo w czasie wielomianowym w sposób optymalny.

Dowód

Optymalne pokolorowanie pojedynczej krawędzi używa zawsze dwóch kolorów - 1 i 2. Rozpatrzmy ścieżkę P_2. Dla P_2 nie istnieje pokolorowanie wykorzystujące tylko dwa kolory. P_2 może zostać pokolorowane sekwencją: (1,2,3,1). Pokolorowanie ścieżki P_k o długości k może zostać rozszerzone do pokolorowania ścieżki P_{k+1} o długości k+1:

- jeżeli końcówka wewnętrzna względem końca ścieżki do którego dołączana jest nowa krawędź miała kolor 1, nowa krawędź kolorowana jest sekwencją (2,3),
- jeżeli końcówka wewnętrzna względem końca ścieżki do którego dołączana jest nowa krawędź miała kolor 2, nowa krawędź kolorowana jest sekwencją (3,2),
- jeżeli końcówka wewnętrzna względem końca ścieżki do którego dołączana jest nowa krawędź miała kolor 3, nowa krawędź kolorowana jest sekwencją (1,2).

Jeżeli liczba krawędzi o poszczególnych kolorach się różni, dokonywana jest permutacja zbioru kolorów (1,2,3) w sposób maksymalizujący liczbę wystąpień koloru 1 i minimalizujący liczbę wystąpień koloru 3..

Twierdzenie 2

Każdy cykl C można pokolorować sumacyjnie końcówkowo w czasie wielomianowym w sposób optymalny.

Dowód

Po rozcięciu przy arbitralnie wybranym wierzchołku, otrzymana ścieżka jest kolorowana w sposób zgodny z Twierdzeniem 1.

Przy ponownym złączeniu cyklu możliwe są trzy przypadki:

- końcówki wewnętrzne względem obu końców ścieżki przyjmują nielegalne pokolorowanie – ten sam kolor lub różne kolory, w sposób niezgodny z sekwencją (1,2,3) – końcówka z kolorem o wyższym numerze zostaje przekolorowana na 4.
- końcówki wewnętrzne względem obu końców ścieżki przyjmują różne kolory, w sposób zgodny z sekwencją (1,2,3) - otrzymane pokolorowanie jest legalne.

4.1. Algorytm 2-przybliżony dla drzew

Problem kolorowania końcówkowego z kryterium sumacyjnym pozostaje otwarty dla drzew. W poniższym podrozdziale przedstawiony zostanie algorytm 2-względnie przybliżonym wraz z dowodem poprawności oszacowania współczynnika aproksymacji.

Niech G=(V, E) jest drzewem. Wybierzmy dowolny wierzchołek jako korzeń i konstruujemy orientację drzewa. W zorientowanym grafie końcówki zewnętrzne względem wierzchołka będziemy nazywać końcówkami wchodzącymi do jego sąsiadów. Zauważmy, że do korzenia nie wchodzi żadna końcówka. Rozpatrzmy potencjalne konflikty pomiędzy kolorami przypisanymi końcówkom. Kolory

końcówek wewnętrznych względem korzenia muszą się różnić, natomiast kolory końcówek zewnętrznych – wchodzących do sąsiadów korzenia – są od siebie niezależne. Analogicznie wygląda sytuacja dla korzeni każdego z drzew w lesie powstałym po usunięciu z grafu poprzedniego korzenia. Ponieważ każde drzewo jest grafem dwudzielnym, możemy podzielić zbiór wierzchołków na partycje A i B. Kolory końcówek wchodzących do wierzchołków z tej samej warstwy są od siebie niezależne.

Obserwacja 1

Konflikty przy wyborze kolorów końcówek wchodzących do wierzchołków odpowiadają konfliktom występującym przy kolorowaniu wierzchołkowym tych wierzchołków.

Obserwacja 2

Konflikty przy wyborze kolorów końcówek wychodzących z wierzchołków (końcówek wewnętrznych względem tych wierzchołków z wyjątkiem końcówki wchodzącej) odpowiadają konfliktom występującym przy kolorowaniu krawędziowym krawędzi, na których leżą te końcówki.

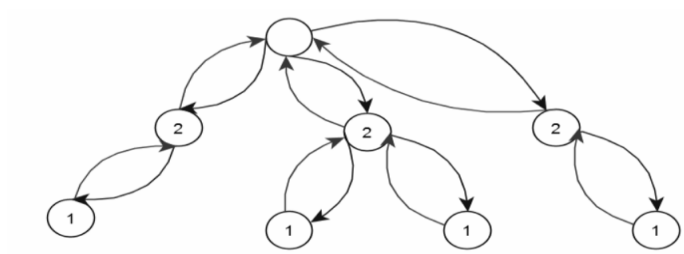
Zauważmy, że użycie rozłącznych zbiorów kolorów do pokolorowania końcówek wchodzących do wierzchołków i końcówek wychodzących z wierzchołków eliminuje konflikty, które mogą wystąpić przy kolorowaniu pomiędzy końcówkami wchodzącymi, i wychodzącymi.

Twierdzenie 3

Konstrukcja optymalnego sumacyjnego 2-kolorowania wierzchołkowego drzewa (nie kolorując korzenia) jest możliwa w czasie wielomianowym.

Dowód

W kolorowaniu używającym tylko 2 kolorów wierzchołki o tym samym kolorze muszą należeć do tej samej partycji. Aby skonstruować takie pokolorowanie, należy rozpatrywać zbiór wierzchołków bez korzenia. Wierzchołki należące do partycji o większej mocy otrzymują kolor 1, kolor 2 zostaje przypisany wierzchołkom z drugiej partycji.



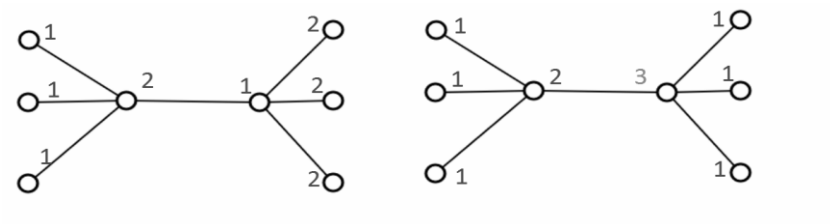
Rysunek 3. Przykład optymalnego sumacyjnego 2-kolorowania wierzchołkowego (bez korzenia). Końcówki reprezentowane są przez łuki

Twierdzenie 4

Optymalne sumacyjne 2-kolorowanie wierzchołkowe drzewa G jest 3/2-względnie przybliżonym sumacyjnym wierzchołkowym kolorowaniem G.

Dowód

Zauważmy, że optymalne sumacyjnego 2-kolorowanie wierzchołkowe może dawać wynik gorszy od optymalnego dla problemu sumacyjnego kolorowania wierzchołkowego bez ograniczeń dotyczących liczby użytych kolorów. Rysunek 4 przedstawia przykład takiego grafu.



Rysunek 4. Przykład grafu, w którym optymalne sumacyjne kolorowanie wierzchołkowe używa większej liczby kolorów niż liczba chromatyczna. Po lewej - optymalne sumacyjne 2-kolorowanie wierzchołkowe (suma 12), po prawej - optymalne sumacyjne kolorowanie wierzchołkowe (suma 11)

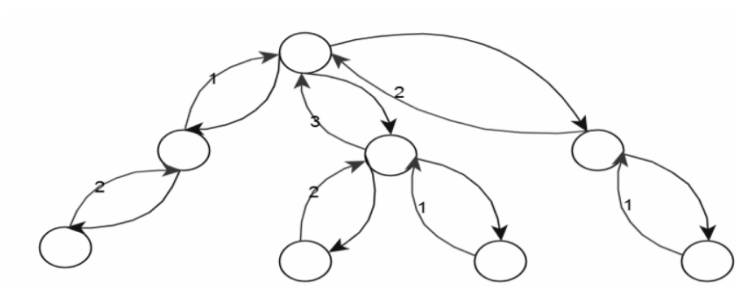
Suma kolorów użytych w optymalnym sumacyjnym 2-kolorowaniu wierzchołkowym drzewa nie przekroczy 3/2 n, gdzie n = |V \ korzeń | - co najmniej połowa wierzchołków (partycja o większej mocy) dostanie kolor 1. Jednocześnie żadne sumacyjne kolorowanie wierzchołkowe grafu nie może użyć sumy kolorów mniejszej niż n. Stąd współczynnik aproksymacji algorytmu dla sumacyjnego kolorowania wierzchołkowego bez ograniczeń dotyczących liczby użytych kolorów można oszacować odgórnie przez 3/2n.

Twierdzenie 5

Optymalne sumacyjne kolorowanie krawędziowe drzewa może zostać skonstruowane w czasie wielomianowym [16].

Dowód

Dowód i algorytm zostały przedstawione w [16].



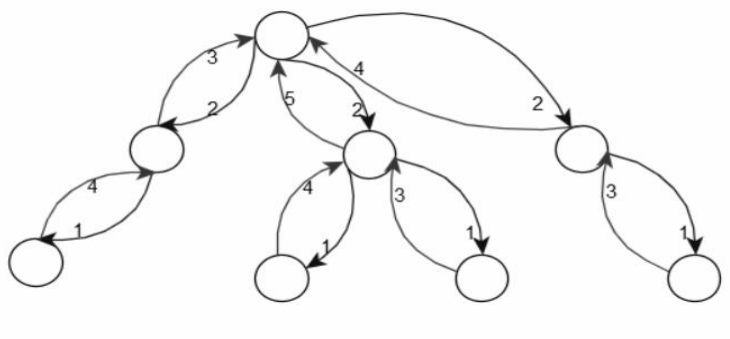
Rysunek 5. Przykład optymalnego sumacyjnego kolorowania krawędziowego. Końcówki reprezentowane są przez łuki

Obserwacja 3

Jeżeli pokolorowania krawędziowe i wierzchołkowe korzystają z rozłącznych zbiorów kolorów, możliwa jest konstrukcja z nich legalnego (niekoniecznie optymatnego) sumacyjnego pokolorowania końcówkowego.

W pierwszym kroku algorytmu konstruowana jest orientacja grafu G. Następnie konstruowane są pokolorowania wierzchołkowe (oprócz korzenia) i krawędziowe G. Wszystkie kolory użyte w pokolorowaniu krawędziowym zostają podniesione o 2 – z zakresu [1, Δ] do [3, Δ+2].

Następnie z pokolorowań krawędziowego i wierzchołkowego konstruowane jest pokolorowanie końcówkowe. Końcówki wewnętrzne (ale nie wchodzące) otrzymują kolory z pokolorowania krawędziowego. Końcówki wchodzące otrzymują kolory z pokolorowania wierzchołkowego.



Rysunek 6. Kolorowanie końcówkowe otrzymane z kolorowań wierzchołkowego i krawędziowego, przedstawionych na Rysunkach 4,5

Otrzymane pokolorowanie jest legalnym, 2-przybliżonym sumacyjnym kolorowaniem końcówkowym.

Współczynnik aproksymacji i analiza złożoności

W poniższych analizach niech m = |E|, n = |V|. Zauważmy, że dla drzew m = n - 1.

Twierdzenie 6

Przedstawiony algorytm sumacyjnego kolorowania końcówkowego jest 2-przybliżony.

Dowód

Zauważmy, że sumę kolorów potrzebnych do pokolorowania końcówkowego można oszacować od dołu przez sumę kolorów użytych w optymalnych kolorowaniach krawędziowym i wierzchołkowym (bez korzenia). Nie jest możliwe użycie mniejszej sumy kolorów – w przeciwnym wypadku przynajmniej jedno z kolorowań krawędziowych i wierzchołkowych nie byłoby optymalne. Zachodzi zatem:

$$OPTSUM\_INCIDENCE \leq OPTSUM\_VERTEX + OPTSUM\_EDGE - 2 \qquad (1)$$

gdzie:

$OPT_{SUM\_INCIDENCE}$ – suma kolorów w optymalnym pokolorowaniu końcówkowym,

OPT_{SUM_VERTEX} – suma kolorów w optymalnym pokolorowaniu wierzchołkowym,

OPT_{SUM_EDGE} – suma kolorów w optymalnym pokolorowaniu krawędziowym.

Na mocy twierdzenia 4, zastosowane w algorytmie 2-kolorowanie wierzchołkowe jest 3/2-przybliżone:

$$n \leq OPTSUM\_VERTEX \leq ALGSUM\_VERTEX \leq 3/2\ n \qquad (2)$$

gdzie:

ALG_{SUM_VERTEX} - suma kolorów w pokolorowaniu wierzchołkowym skonstruowanym przy pomocy prezentowanego algorytmu, OPT_{SUM_VERTEX} – suma kolorów w optymalnym pokolorowaniu wierzchołkowym.

Suma kolorów otrzymanych w użytym w algorytmie kolorowanie krawędziowym wynosi OPT_{SUM_EDGE} + 2 n - 2 (podbicie kolorów o 2). Otrzymujemy zatem :

$$OPTSUM\_EDGE + 2\ n - 2 + 3/2\ n \leq ALGSUM\_INCIDENCE \qquad (3)$$

gdzie:

OPT_{SUM_EDGE} – suma kolorów w optymalnym pokolorowaniu krawędziowym., $ALG_{SUM\_INCIDENCE}$ – suma kolorów w otrzymanym z algorytmu kolorowaniu końcówkowym;

Dzieląc rozwiązanie algorytmu przez sumę kolorów w rozwiązaniu optymalnym i oszacowanie górne algorytmu przez oszacowanie dolne rozwiązania optymalnego otrzymujemy:

$$\frac{ALGSUM\_INCIDENCE}{OPTSUM\_INCIDENCE} \leq \frac{OPTSUM\_EDGE + 2\ n - 2 + 3/2\ n + 1}{OPTSUM\_EDGE + OPTSUM\_VERTEX} \qquad (4)$$

$$\frac{ALGSUM\_INCIDENCE}{OPTSUM\_INCIDENCE} \leq \frac{OPTSUM\_EDGE + 5/2\ n - 1}{OPTSUM\_EDGE + OPTSUM\_VERTEX} \qquad (5)$$

$$\frac{ALG_{SUM\_INCIDENCE}}{OPT_{SUM\_INCIDENCE}} \leq 1 + \frac{5/2\ n - 1}{5/2\ n} \leq 2 \qquad (6)$$

gdzie:

OPT_{SUM_EDGE} - suma kolorów w optymalnym pokolorowaniu krawędziowym,

OPT_{SUM_VERTEX} - suma kolorów w optymalnym pokolorowaniu wierzchołkowym,

$OPT_{SUM\_INCIDENCE}$ - suma kolorów w optymalnym pokolorowaniu końcówkowym,

$ALG_{SUM\_INCIDENCE}$ - suma kolorów w otrzymanym z algorytmu pokolorowaniu końcówkowym;

Algorytm jest zatem 2-przybliżony.

Twierdzenie 7

2-przybliżony algorytm sumacyjnego kolorowania końcówkowego wykonuje się w czasie $O(n^{3/2}\Delta^{7/2}\log(n))$

Dowód

Orientacja grafu może zostać skonstruowana w czasie O(m). Optymalne sumacyjne 2-kolorowanie wierzchołkowe wierzchołków (oprócz korzenia) wykonywane jest w czasie O(n). Najbardziej kosztowny krok algorytmu stanowi konstrukcja optymalnego sumacyjnego Δ-kolorowania krawędziowego. Algorytm przedstawiony w [15] wykonuje się w czasie $O(n^{3/2}\Delta^{7/2}\log(n))$. Połączenie pokolorowania wierzchołkowego i krawędziowego i konstrukcja pokolorowania końcówkowego wykonuje się w czasie O(m).

5. Podsumowanie

W pracy przedstawiono propozycję modelu szeregowania zadań dwuprocesorowych w systemie otwartym. Przedstawiony został sposób budowy modelu grafowego reprezentującego instancję rozważanego problemu. Przedstawiony został model kolorowania końcówkowego grafów z kryterium sumacyjnym. Przedstawione zostały algorytmy dokładne dla prostych klas grafów. Zaprezentowany został wielomianowy algorytm 2-przybliżony dla drzew.

Model kolorowania końcówkowego z kryterium sumacyjnym wymaga przeprowadzenia dalszych badań, ze szczególnym uwzględnieniem określenia klas, dla których staje się trudny obliczeniowo, oraz znalezienia dolnych i górnych oszacowań sumy potrzebnych kolorów.

Literatura

1. Sahni S., Gonzalez T. *Open shop scheduling to minimize finish time,* Journal of the ACM 3 (1976), s. 665-679.
2. Hansen H.M. *Skemalægning med henblik p°a minimering af ventetid* ,(po duńsku, praca magisterska), University of Odense, 1992.
3. Lenstra J.K., Rinnooy Kan A.H.G., Graham R.L., Lawler E.L. *Optimization and approximation in deterministic sequencing and scheduling: a survey,* Research Institute on Discrete Optimization and Systems Applications of the Systems Science Panel of NATO and of the Discrete Optimization Symposium. Elsevier 5 (1979), s. 287-326.
4. Giaro K. *Szeregowanie zadań metodami kolorowania grafów,* (rozprawa habilitacyjna), Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska, 2002.
5. Anand E., Panneerselvame R. *Literature review of open shop scheduling problems,* Intelligent Information Management 7 (2015), s. 33-52,
6. Chen Y., Zhang A., Chen G., Dong J., *Approximation algorithms for parallel open shop scheduling,* Information Processing Letters 113 (2013), s. 220-224.

7. Baptiste P., *On minimizing the weighted number of late jobs in unit execution time open-shops*, European Journal of Operational Research 149 (2003), s. 344-354.
8. Andresen M., Bräsel H., Mörig M., Tusch J., Werner F., Willenius P., *Simulated annealing and genetic algorithms for minimizing mean flow time in an open shop*, Mathematical and Computer Modelling 48 (2008), s. 1279-1293.
9. Nadolski A., Kubale M., *Chromatic scheduling in a cyclic open shop*, European Journal of Operational Research 164 (2005), s. 585-591.
10. Małafiejska A. *Modele kolorowania końcówkowego*, Raport WETI nr 3/201.
11. Małafiejska A. *Kolorowanie końcówkowe grafów*, (rozprawa doktorska), Wydział Matematyki, Informatyki i Fizyki, Uniwersytet Gdański 2016.
12. Brualdi R.A. Massey J.Q.*Incidence and strong edge coloring*, Discrete Mathematics., 122 (1993), s. 51-58.
13. Chen D.L., Liu X.K., Wang S.D. *The incidence coloring number of graph and the coloring conjecture, Math. Econom*, (People's Republic of China), 15 (1998), s. 47-51.
14. Lih K.W., Wang W.F.*Coloring the square of an outerplanar graph*, Taiwanese J. Math., 10 (2006), s. 1015-1023.
15. Shiu W.C., Sun P.K.*Invalid proofs on incidence coloring*, Discrete Mathematics, 308 (2008), s. 6575-6580.
16. Giaro K., Kubale M.,*Edge-chromatic sum of trees and bounded cyclicity graphs*, Information Processing Letters 75 (2000), s. 65-69.

Szeregowanie zadań dwuprocesorowych w systemach otwartych

Streszczenie
W pracy rozważany jest problem szeregowania zadań dwuoperacyjnych w systemie otwartym (open-shop), z kryterium minimalizacji długości harmonogramu oraz sumy czasów zakończenia wszystkich zadań. Zakładając jednostkowe czasy wykonywania operacji można stosować efektywne metody chromatyczne rozwiązywania problemu, poprzez sprowadzenie go do modelu grafowego oraz zastosowanie w nim wybranego modelu kolorowania, które pozwala uzyskać optymalny harmonogram. W kontekście przetwarzania zadań w obliczeniach bazodanowych oraz komunikacji między serwerami, zaproponowany został szczególny model systemu otwartego, w którym operacjom przypisujemy dwa dedykowane procesory, o asymetrii wykorzystania: jeden procesor przetwarza operację w trybie wyłączności (write), drugi jest wykorzystywany w trybie współdzielenia (read). Przedstawiony został wielomianowy algorytm 2-przybliżony dla sumacyjnego kolorowania końcówkowego drzew oraz wielomianowe algorytmy optymalne dla sumacyjnego kolorowania końcówkowego prostych klas grafów.
Słowa kluczowe: szeregowanie chromatyczne, sumacyjne kolorowanie końcówkowe, szeregowanie zadań, kolorowanie końcówkowe

Scheduling of bi–processor jobs in open systems

Abstract
The problem of scheduling bi-processor jobs in open systems (open shop) is considered in this paper, with the optimization functions of minimization of makespan of resultant schedule and minimization of total (mean) completion time of all jobs. Under the assumption that each job has unit execution time, effective chromatic methods may be used to construct a feasible schedule. Novel job scheduling model was proposed, with potential applications in the database systems and servers, i.e., in testing connections and in authentication of communication channels. In this model, each operation requires two processors - one in exclusive access mode (write), and one in concurrent access mode (read).
Polynomial time 2-approximation algorithm for sum incidence coloring of trees is presented. Polynomial time optimal algorithms for sum incidence coloring of simple classes of graphs are presented.
Keywords: task scheduling, incidence coloring, sum incidence coloring, chromatic scheduling

Stanisław Skulimowski[1], Michał Dobrowolski[2]

Przegląd interfejsów do zarządzania bazami danych

1. Wprowadzenie

Dynamiczny rozwój technologiczny oraz szybki przyrost ilości przepływu informacji w globalnej sieci komunikacyjnej Internet, wywołały potrzebę usprawnienia sposobu gromadzenia, przetwarzania i przesyłu danych. Obecnie każda aplikacja lub system informatyczny współpracuje z różnymi, zewnętrznymi źródłami danych w postaci plików tekstowych, plików xml, czy baz danych [1]. Aby ta współpraca była możliwa konieczne jest zagwarantowanie łącznika między danymi a aplikacją. Jest to realizowane przez tworzenie interfejsów bazodanowych rozumianych jako technologie dostępu do danych. Wraz z rozwojem technologii bazodanowych powstawały nowe, lepsze i szybsze sposoby łączenia z systemami baz danych. Jednocześnie, ze względu na nierównomierny rozkład technologii i metodyk, wiele stosowanych obecnie rozwiązań różni się od siebie pod względem podejścia, architektury oraz funkcjami. Wybór interfejsu do nawiązywania komunikacji z źródłami danych wymaga uprzedniego stwierdzenia: jakie technologie są dostępne oraz jakie technologie dominują w danym systemie informatycznym, jakie jest przeznaczenie aplikacji korzystającej z określonych źródeł danych, oraz tego jak zbudowane są same zbiory danych.

W artykule przedstawiono kilka wybranych interfejsów bazodanowych opisując ich historię powstania, posiadane funkcjonalności, architekturę oraz cechy szczególne. Poza podsumowaniem przeglądu technologicznego, celem pracy było porównanie wybranych interfejsów do zarządzania bazami danych pod względem wydajności. W artykule zamieszczono także wyniki prowadzonych testów, z użyciem lokalnych i zewnętrznych źródeł danych.

2. Wybrane interfejsy bazodanowe

Istnieje wiele technologii umożliwiających zdalny i bezpośredni dostęp do zasobów danych (ang. data access technologies). Spośród nich, autorzy zdecydowali się wybrać te najpopularniejsze i najczęściej stosowane, tj.:

- ODBC, którego pierwsza wersja pojawiła się w 1992 roku,
- OLE DB opracowany przez firmę Microsoft w roku 1996,
- JDBC opracowany w 1997 roku przez firmę Sun Microsystems,
- ADO.NET opracowany prze firmę Microsoft w 2002 roku.

2.1. ODBC

ODBC (ang. The Open Database Connectivity) jest interfejsem pozwalającym aplikacjom łączyć się z systemami zarządzającymi bazami danych (DBMS – ang. database management systems) wykorzystujących strukturalny język zapytań (SQL –

[1] s.skulimowski@pollub.pl, Instytut Informatyki, Wydział Elektrotechniki i Informatyki, Politechnika Lubelska, www.pollub.pl
2 michal.dobrowolski1@pollub.edu.pl, Politechnika Lubelska, www.pollub.pl

ang. structured query language) jako standard dostępu do danych. ODBC umożliwia pełną interoperacyjność. Oznacza to, że aplikacja może uzyskać dostęp do różnych systemów zarządzania bazami danych bez ograniczeń dostępu i implementacji. Pozwala to programistom na tworzenie aplikacji bez konieczności wyboru konkretnego systemu zarządzania bazami danych oraz języka. Użytkownicy programu mogą wybrać, z jakiego DBMS będą korzystać poprzez dobranie odpowiedniego sterownika [2].

ODBC umożliwia wydzielenie z aplikacji modułu do pobierania danych z bazy. Oznacza to, że zmiana środowiska aplikacji nie wymaga ponownej kompilacji kodu źródłowego, co w znaczący sposób przekłada się na wzrost wydajności pracy nad projektem informatycznym.

Ponieważ dostępne są różne metody komunikacji, protokoły oraz systemy zarządzania DB, rozwiązania ODBC umożliwiają wielu technologiom na korzystanie z niego poprzez właściwe definiowanie własnego interfejsu, stosując sterowniki baz danych. Sterowniki zawierają biblioteki współdzielone, pozwalające aplikacji na uzyskanie dostępu do określonych źródeł danych za pośrednictwem zdefiniowanych metody komunikacji.

Interfejs ODBC określa:

- Biblioteki funkcji wywołań, które pozwalają aplikacji na połączenie z DBMS, wykonywanie poleceń oraz odzyskiwanie,
- Specyfikację SQL SQLsyntax based on the X/Open and SQL Access Group (SAG) SQL CAE,
- Zestaw standardowych kodów błędów,
- Standardowy sposób łączenia i logowania do DBMS,
- Standardową reprezentację typów danych.

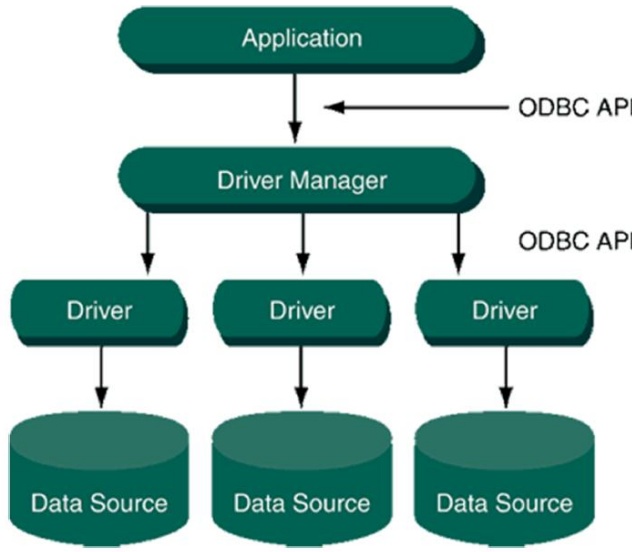
Interfejs ODBC jest elastyczny, ponieważ:

- Łańcuchy znaków zawierające polecenia SQL mogą być włączane w kodzie źródłowym,
- Te same obiekty mogą być użyte w różnych systemach DBMS,
- Wartości danych mogą być wysyłane oraz odzyskiwane w formacie wygodnym dla aplikacji.

Interfejs ODBC zapewnia dwa typy wywołań funkcji:

- Fundamentalne funkcje oparte na specyfikacji X/Open and SQL Access Group Call Level,
- Funkcje rozszerzone wspierające dodatkowe funkcjonalności takie jak przewijany kursor oraz asynchroniczne procesowanie.

Architektura interfejsu ODBC składa się czterech warstw (rys. 1.): Aplikacji, Driver Managera, Sterowników (ang. Drivers) oraz z Źródeł danych. Aplikacja odwołuje się do funkcji ODBC i wykonuje zapytania SQL. Driver Manager zapewnia informacje dla aplikacji(np. liczbę dostępnych źródeł danych) i odpowiedzialny jest za załadowanie odpowiedniego sterownika dla aplikacji. Sterownik procesuje wywołania funkcji ODBC i zarządza wymianą danych między aplikacją a konkretnym źródłem danych. Ostatni element, źródła danych, obejmuje dane i zawiązany z nim silnik baz danych.



Rysunek 1. Architektura ODBC [3]

Aby połączyć się z źródłem danych, aplikacja używa ODBC API. Aplikacja przeprowadza takie operacje jak pobranie danych czy wykonania zapytań SQL, a następnie rozłącza się. Driver manager decyduje jaki sterownik ma zostać załadowany oraz zarządza komunikacją. Po wczytaniu odpowiedniego sterownika, następuje implementacja funkcji ODBC dla określonej bazy danych. Tak zaprojektowana architektura umożliwia dostęp do różnych źródeł danych, w różnych lokalizacjach korzystając z tych samych funkcji.

2.2. JDBC

Większość systemów baz danych nie ma zbyt wiele wspólnego. Łączy je jedynie podobny cel i przede wszystkim zgodne języki zapytań. Ponadto, każda baza danych posiada swoje własne API, które uprzednio należy poznać, aby móc tworzyć programy korzystające z danych zawartych w tych bazach. Dlatego napisanie kodu, który byłby wykorzystany dla każdego z systemów baz danych stanowi pewne wyzwanie. Dodatkowo, ODBC jest technologią stworzoną i stosowaną w środowiskach powstałych w oparciu o języki C i C++. Wraz ze wzrostem popularności języka Java, problem zależności od systemu operacyjnego stał się na tyle poważny, że konieczne było opracowanie nowego standardu tworzenia połączeń bazodanowych. Przepisanie ODBC w Javie nie miało sensu, z uwagi na różnice językowe (konstrukcyjne) między Javą a C. ODBC, w oczach twórców Javy, było zbyt skomplikowane. Nowo powstały interfejs miał być znacznie prostszy w użyciu, a jednocześnie bardziej funkcjonalny [4].

JDBC (ang. Java Database Connectivity) jest produktem firmy Oracle, stanowiącym niezależny od dostawców silników bazodanowych standard dostępowy, który można wykorzystywać w aplikacjach tworzonych w języku Java [5]. Zapewnia łącz-

ność pomiędzy programami stworzonymi w języku Java oraz różnymi źródłami danych w tym baz danych SQL wielu producentów.

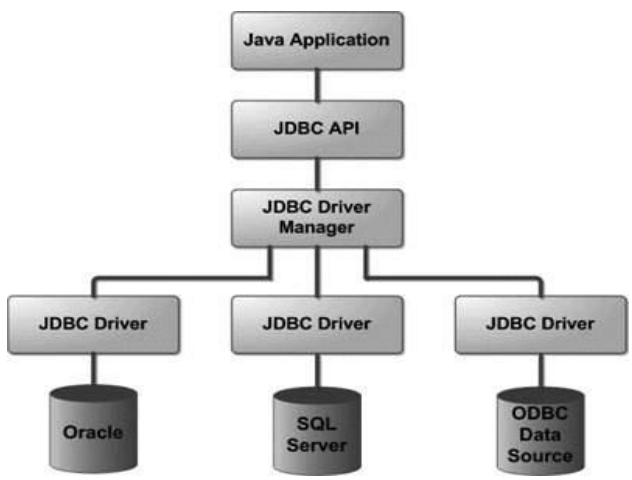
JDBC to specyfikacja, która dostarcza kompletny zestaw interfejsów. Odpowiednio zaimplementowane pozwalają na dostęp do dowolnej bazy danych. Dodatkowo wykorzystywany jest, jako podstawa dla aplikacji czy interfejsów wyższego rzędu, które umożliwiają dostęp do baz danych z wyższego poziomu.

Biblioteki JDBC zawierają API umożliwiające: zestawienie połączenia z bazą danych, tworzenie wyrażeń SQL, wykonywanie zapytań SQL oraz wyświetlanie i modyfikowanie rezultatów. Sterowniki JDBC mogą być wykorzystywane do tworzenia połączeń z bazą danych, z poziomu: aplikacji, appletów, servletów, czy server-pages. JDBC cieszy się dużym wsparciem ze strony firm IT [7].

JDBC jest niezależnym od platformy interfejsem pomiędzy bazą danych a Javą. JDBC udostępnia standardowy zestaw możliwości dostępu do baz danych. Architektura JDBC składa się z dwóch warstw JDBC API i JDBC Driver API (rys. 2.).

JDBC API określa zbiór interfejsów, które hermetyzują główne funkcjonalności bazy danych, wliczając w to uruchomione zapytania, procesowane wyniki oraz określenia informacji o konfiguracji.

JDBC Driver API stanowi zbiór klas, które implementują interfejsy poszczególnych systemów baz danych. Implementacja zależy bezpośrednio od systemu zarządzania bazą danych, z którą będzie się łączyć aplikacja wykorzystująca sterownik.



Rysunek 2. Architektura JDBC [6]

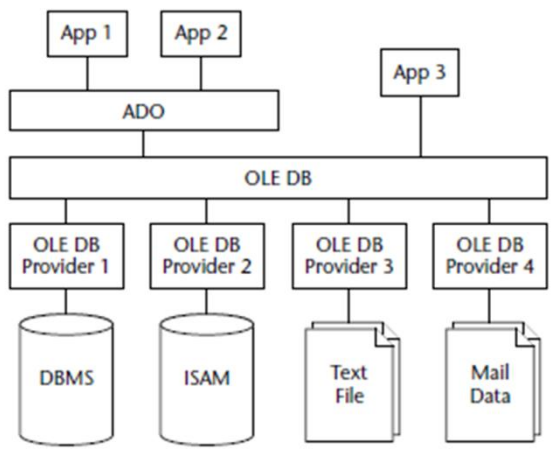
2.3. OLE DB

OLE DB (ang. Object Linking and Embedding Database) jest to produkt firmy Microsoft stanowiący alternatywę dla ODBC, który dodatkowo rozszerza jego funkcję o dostęp do nierelacyjnych źródeł danych [8] a także może być wykorzystywany w procesie eksploracji danych (ang. Data mining) [9]. OLE DB składa się z zestawu interfejsów COM (ang. Component Object Model), które udostępniają dane z różnych

źródeł [10]. OLE DB zapewnia wsparcie dla wielu systemów zarządzania bazami danych. Interfejs ten jest elementem składowym biblioteki MDAC (ang. Microsoft Data Access Components), będącej zbiorem komponentów pozwalającym programistom w jednolity i kompleksowy sposób realizować dostęp do danych. Wspomniany interfejs bazodanowy działa w oparciu o komponenty COM z dobrze zdefiniowanymi interfejsami. Jedną z wad architektury OLE DB jest brak agregatów serwerowych.

W porównaniu z interfejsami otwartymi typu ODBC, OLE DB jest rozwiązaniem własnościowym i nieprzenośnym, przeznaczonym jedynie dla oprogramowania w środowisku Microsoft Windows. Lista firm wspierających technologię OLE DB liczy kilkadziesiąt pozycji [11].

Pod względem architektury (rys. 3.), OLE DB jest trudniejsze w implementacji. Dlatego stworzono technologie wyższego poziomu tj. ADO, które wprawdzie wykorzystują OLE.DB do łączenia się z bazami danych, jednak ich implementacja jest zdecydowanie uproszczona.

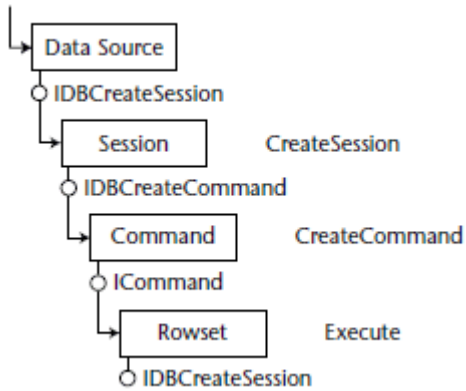


Rysunek 3. Architektura JDBC [12]

ADO (ang. ActiveX Data Objects) jest warstwą pośrednią, przez którą można uzyskać dostęp baz danych, oraz do plików Excela, plików tekstowych, plików Lotusa, HTML i wielu innych źródeł danych [13]. Technologia ADO została zaadoptowana także przez inne firmy, dzięki czemu można aktualnie korzystać z ADO np. w Borland Delph. Idea, która przyczyniła się do powstania ADO, zakładała dostęp do bazy danych bez znajomości jej wewnętrznej struktury.

OLE DB dzieli się na dwie warstwy odbiorców (ang. consumers) i dostawców (ang. providers). Odbiorcę stanowią aplikacje, które korzystają z danych z określonego źródła, natomiast dostawca jest częścią oprogramowania, która odpowiedzialna jest za prezentację danych do odbiorcy (rys. 4.).

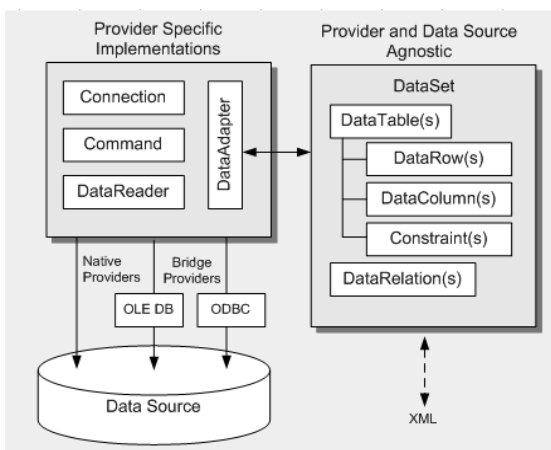
Odpowiednikiem dostawcy dla ODBC są sterowniki (ang. drivers) Każdemu ze źródeł danych odpowiada inny provider. Interfejsy OLE DB implementowane są przez grupę obiektów. Między innymi są to obiekty źródeł danych, sesji, komend i wykonywalny obiekt Rowset.



Rysunek 4. Obiekty implementujące interfejsy OLE DB [12]

2.4. ADO.NET

ADO.NET (ang. ActiveX Data Objects for .NET) to obiektowo zorientowany zestaw bibliotek, które pozwalają nam wchodzić w interakcję ze źródłami danych z poziomu aplikacji [14]. Najczęściej są nimi baza danych, ale podobnie jak w przypadki OLE DB, mogą to być nierelacyjne źródła danych takie jak pliki, maile, arkusz kalkulacyjny excel, pliki xml itp. ADO.NET jest ewolucją ADO, która zapewnia lepszą współpracę i skalowalną platformę dostępu do danych. Zapewnia on bogaty zestaw komponentów do tworzenia aplikacji rozproszonych, udostępniania danych. Stanowi integralną część platformy .NET Framework. ADO.NET obsługuje różne potrzeby rozwojowe [15], w tym tworzenie obiektów biznesowych.



Rysunek 5. Architektura ADO.NET [16]

ADO.NET został zaprojektowany w celu dostosowania się do nowego modelu programowania, czyli odłączonej architektury danych, ścisłej integracji z XML, wspólnej reprezentacji danych z możliwością łączenia danych z wielu źródeł. ADO.NET

wspiera obsługę takich źródeł danych jak: SQL Server, SQL Server Compact 4.0, OLE DB, ODBC czy źródła danych Oracle.

Architekturę ADO.NET możemy podzielić na dwie warstwy (rys. 5.): warstwę mechanizmów dostępu do danych oraz warstwę przechowywania danych.

Warstwy te wyposażone są w zestawy obiektów do pracy z danymi. Do głównych obiektów należą:

- SqlConnection – ustanawia połączenie z określonego źródła danych,
- SqlCommand – Wykonuje polecenie źródeł danych. Obiekt command używa obiektu connection do określenia bazy danych, na której ma zostać wykonane polecenie,
- SqlDataReader – Obiekt ten zawiera rezultat wykonania instrukcji z obiektu command. Dostęp do danych można uzyskać tylko w sposób sekwencyjny,
- SqlDataAdapter – Zapewnia bezpołączeniowy dostęp do danych. DataAdapter wypełnia obiekt DataSet odczytanymi danymi i zapisuje je w pojedynczych pakietach. DataAdapter zawiera referencję do obiektu połączenia i automatycznie otwiera i zamyka połączenie po odczycie lub zapisie danych,
- DataSet – Reprezentuje bezpołączeniowy zestaw danych, jest niezależny od dostawcy. Nie jest przywiązany do źródeł danych które mogłyby zostać użyte do wypełnienia go danymi,
- ADO.NET wspiera dwa typy warstw: połączeniową i bezpołączeniową [17].

Pierwsza wymaga rzeczywistego połączenia z bazą danych, które musi pozostać otwarte podczas interakcji. Używane jest zazwyczaj dla krótkich transakcji takich jak wywołań procedur składowanych, uruchamiania aktualizacji schematów czy przy transakcjach zapisu. Warstwa bezpołączeniowa wypełnia danymi obiekt DataSet, który stanowi reprezentację danych offline. Wykorzystując tą warstwę, połączenie otwierane jest tylko podczas pobrania danych do uzupełnienia obiektu DataSet.

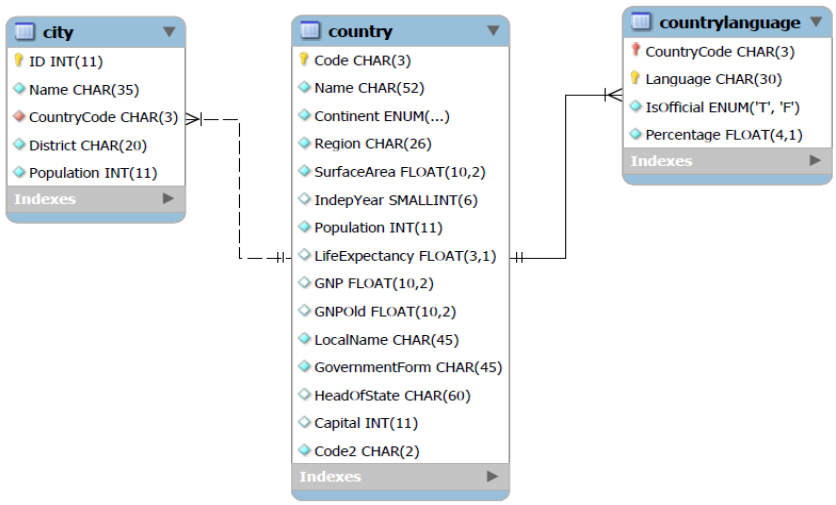
3. Testy wydajnościowe interfejsów połączeniowych

Do przeprowadzenie testów wydajnościowych niezbędne było przygotowanie odpowiedniego środowiska. Środowisko to składało się z bazy danych wystawionej na serwerze MySql. Tak jak zaprezentowano na rysunku 6 baza ta składa się z trzech relacyjnie połączonych tabel.

Tabela city zawierająca nazwy miast oraz kod państwa do którego należy, country zawierająca dane o danym państwie oraz countrylanguage zawierająca język którym posługuje się w danym państwie. Tabele powiązane są za pomocą kodu państwa. Dodatkowo należało napisać programy, które za pomocą określonego sterownika, łączyłyby się do bazy danych oraz wykonywałyby pomiar czasu każdego z testów.

Na każdym z interfejsów przeprowadzono cztery testy.

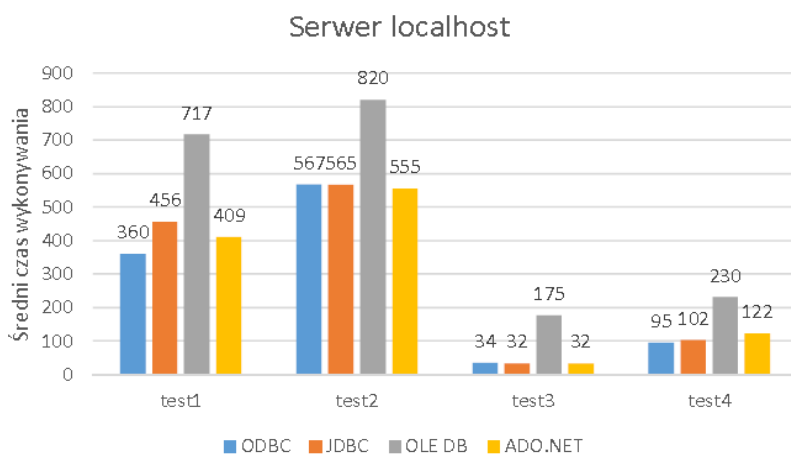
- Test pierwszy polegał na wykonaniu zapytania SQL wyświetlenia wszystkich rekordów z tabeli city,
- Drugi test polegał na złączeniu dwóch tabel i wyświetleniu wyniku o podanym kryterium,
- Trzeci test polegał na dodaniu pewnej liczby rekordów do tabeli city,
- Czwarty test polegał na usunięciu nowo dodanych rekordów.



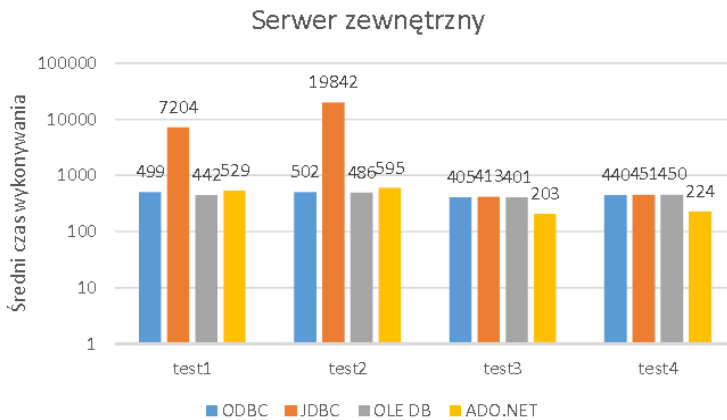
Rysunek 6. Struktury bazy danych wykorzystanej do testów [opracowanie własne]

Każdy z testów został powtórzony i zmierzony 50 razy, a otrzymane czasy zostały uśrednione. Testy zostały przeprowadzone na serwerze lokalnym oraz na serwerze zdalnym. Wyniki testów przeprowadzonych na serwerze lokalnym zaprezentowano na wykresie 1.

Aby zbadać wydajność poszczególnych interfejsów, niniejsza baza danych została utworzona na zewnętrznym serwisie Amazon Elastic Compute Cloud. Jest ona dostępna pod adresem 52.33.22.14:3306. Otrzymane wyniki zostały zaprezentowane na wykresie 2.



Wykres 1. Porównanie wyników testów przeprowadzonych na serwerze lokalnym [opracowanie własne]



Wykres 2. Porównanie wyników testów przeprowadzonych na serwerze zewnętrznym [opracowanie własne]

Z przeprowadzonych testów wynika, że przy odwoływaniu się do danych lokalnych, najgorzej spisuje się interfejs OLE DB. Pozostałe technologie uzyskały zbliżone rezultaty. ODBC uzyskiwał minimalnie najkrótsze czasy realizacji testów.

W przypadku testów prowadzonych z użyciem zewnętrznej bazy danych, najwolniejszą spośród testowanych technologii okazała się JDBC, która w teście pierwszym oraz drugim wykazała bardzo duże opóźnienia. Wyniki osiągnięte przez ADO.NET, OLE DB i ODBC są bardzo zbliżone i nie pozwalają na jednoznaczne określenie najlepszego spośród interfejsu pod względem szybkości realizacji zapytań.

4. Podsumowanie

Każdy z przedstawionych interfejsów może być z powodzeniem wykorzystywany w środowisku przewidzianym przez dostawcę technologii. Należy przy tym pamiętać, że każde bardziej skomplikowane zadanie, wymaga dostosowania wybranej technologii i że podstawowa konfiguracja może nie wystarczyć do wykorzystania pełni potencjału interfejsów bazodanowych.

Klasyfikacja interfejsów pod względem wydajności jest zależna między innymi od lokalizacji serwera i wykorzystywanego łącza, co udało się dowieść na podstawie przeprowadzonych testów.

Dalsze prace nad testowaniem interfejsów bazodanowych będą zakładały uściślenie formy prowadzenia testów oraz uwzględnienie takich technologii JPA (ang. Java Persistence API).

Uwagi ogólne

Autorzy zdają sobie sprawę, że przeprowadzone testy wydajnościowe można rozszerzyć o dodatkowe interfejsy, warianty prowadzonych operacji, a także można je wzbogacić przez dokładny opis lokalizacji serwera docelowego, prędkości posiadanego łącza internetowego. Testy mogą być także prowadzone w oparciu o różne środowiska uruchomieniowej oraz mogą bazować na większej ilości powtórzeń.

Literatura

1. Goodson J., Steward R.A., The Data Access Handbook: Achieving Optimal Database Application Performance and Scalability, 2009.
2. *Microsoft Open Database Connectivity (ODBC)*, Microsoft, https://msdn.microsoft.com/en-us/library/ms710252
3. Agarwal V., *ODBC Driver Development*, DR.Dobb's, http://www.drdobbs.com/windows/odbc-driver-development/184416434, (2002).
4. He G.I., *Application of design pattern in the JDBC programming*, Computer Science & Education (ICCSE), (2013).
5. *JDBC Overview*, Oracle, http://www.oracle.com/technetwork/java/overview-141217.html.
6. *JDBC – Introduction*, tutorials point – simply easy learning, https://www.tutorialspoint.com/jdbc/jdbc-introduction.htm
7. *Industry Support*, Oracle, http://www.oracle.com/technetwork/java/index-136695.html
8. Blakeley J.A., *Universal data access with OLE DB*, Compcon '97 Proceedings, (1997)
9. Tang Z., Maclennan J., Kim P. P., *Building data mining solutions with OLE DB for DM and XML for analysis*, ACM SIGMOD Record, 2(2005, s. 80-85.
10. *Microsoft OLE DB*, Microsoft, https://msdn.microsoft.com/en-us/library/ms722784
11. *OLE DB Vendors*, datamystic, http://www.datamystic.com/datapipe/oledb_vendors
12. *Introducting Ole Db*, wisdomjobs.com, https://www.wisdomjobs.com/e-university/data-mining-tutorial-199/introducing-ole-db-1899.html
13. Chappell D.A., Understanding ActiveX and OLE, (1996).
14. *ADO.NET*, Microsoft, https://msdn.microsoft.com/en-us/library/e80y5yhx
15. Adya A., Blakeley J.A., Melnik S., *Muralidhar S., Anatomy of the ADO.NET entity framework*, SIGMOD '07 Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 2007, s. 877-888.
16. *Chapter 12 – Improving ADO.NET Performance*, Microsoft, https://msdn.microsoft.com/en-us/library/ff647768.aspx, (2004)
17. Riordan R.M., Riorsan R., *Microsoft Ado.Net Step by Step*, Microsoft Ado.Net Step by Step, (2002).

Przegląd interfejsów do zarządzania bazami danych

Streszczenie

Następstwem rozwoju globalnej sieci informacyjnej jest tworzenie bardziej wydajnych systemów pośredniczących pomiędzy dostawcami danych i zbiorami danych a użytkownikami wykorzystującymi te dane do dalszej analizy. Opracowanie architektury oddzielającej warstwę analizy od zarządzania analizy doprowadziło do powstania różnych interfejsów służących do komunikacji aplikacji z źródłami danych. Artykuł zawiera opis funkcjonalności i architektury kilku wybranych interfejsów bazodanowych.

Słowa kluczowe: technologie dostępu do danych, ODBC, JDBC, ADO.NET, OLE DB

Overview of database management interfaces

Abstract

The consequence of the global information network development is creation of more efficient intermediary systems between data providers and data sets, and users using that data for further analysis. Developing an architecture that separates the analysis layer from the management of analysis has led to emerging different data access technologies. This article describes the functionality and architecture of some selected database interfaces.

Keywords: data access technologies, ODBC, JDBC, ADO.NET, OLE DB

Jakub Pizoń[1], Tomasz Cieplak[2], Łukasz Kański[3]

Modele dojrzałości i potencjał rozwiązań Internetu rzeczy

1. Wstęp

Nowoczesny świat charakteryzuje się tym, że każdego dnia tworzone są nowe, niesamowite rozwiązania. Co kilka tygodni pojawiają się nowe wzory, prototypy i rozwiązania.

W rzeczywistości tylko część nowych produktów zostaje przyjęta i wykorzystana przez osoby w ich codziennym życiu. To czy technologia lub koncepcja będą realizmwane, w znacznej mierze zależy od jego wartości dodanej, funkcjonalności, stabilności, ale także od statusu badań potwierdzonego liczbą opisywanych prac naukowych.

Nowe technologie dzięki globalizacji wpływają na ludzi, firmy, kraje, a czasami na cały świat. Jednym z takich zjawisk jest teraz i nazywa się Internet rzeczy (ang. Internet of things, IoT). Z tego powodu coraz więcej osób koncentruje się na tym, czym jest IoT [16]. Dlatego ważne jest dla nich, czy nowy paradygmat to po prostu koncepcja czy już wydajna technologia, której nie trzeba dostosować. Ma to szczególne znaczenie zarówno dla firm, które poszukują odpowiedniej technologii, która wzbudzi rozwój i innowacje, a przede wszystkim zwiększy wartość dodaną wytwarzanych produktów.

Wdrażanie nowych technologii zależy od różnych czynników. Przy czym w oczach managerów szczególną rolę grają czynniki finansowe, takie jak stopa zwrotu z inwerstycji. Jeśli menedżerowie zainwestują w nie korzystne technologie, może to spowodować zagrożenie dla firmy. Tym samym to jakie technologie są wdrażane stanowi kluczowy element inwestycji.

Z drugiej strony ważna jest też perspektywa użytkownika. Ważne jest czy dana technologia jest bezpieczna i ma realny wpływ na ich pracę oraz na życie codzienne. Dlatego też, firmy tak samo jak użytkownicy z rezerwą przyjmują i kupują nowe technologie. Zwłaszcza gdy proponowane rozwiązanie nie jest powszechne na rynku wokół firmy/użytkownika. Szczególnie z punktu widzenia zarządzania i rozwiązań technicznych konieczne jest zidentyfikowanie gotowości technologii jak i są rozwiązania, które zapewniają odpowiednio do poziomu jego dojrzałości.

Niniejszy artykuł dotyczy Internetu rzeczy, przedstawia ona dyskusję na temat różnych modeli dojrzałości IoT, która wskazuje kluczowe elementy proponowanej dojrzałości IoT.

[1] jakub.pizon@pollub.edu.pl, Doktorant na Wydziale Mechanicznym Politechniki Lubelskiej, Wydział Mechaniczny, Politechnika Lubelska, www.pollub.pl
[2] t.cieplak@pollub.pl, Katedra Organizacji Przedsiębiorstwa, Wydział Zarządzania, Politechnika Lubelska, www.pollub.pl
[3] kansil@o2.pl, CA Consulting SA

2. Internet rzeczy (IoT)

Kluczową ideą rozwiązań Internetu rzeczy jest uzyskanie informacji o otaczającym środowisku, w celu jej zrozumienia, kontrolowania i wpływania na nią. Cechą charakterystyczną i kluczową dla Internetu rzeczy jest to, że korzysta z urządzeń o niskim zużyciu energii, ograniczonej pamięci i baterii oraz niewielkiej mocy wysyłania informacji [6].

W 2005 r., kiedy Międzynarodowe Stowarzyszenie Telekomunikacji (ITU) opublikowało pierwsze sprawozdanie w tej sprawie IoT [3].

ITU wskazało, że: „Nowy wymiar został dodany do świata technologii informacyjnych i komunikacyjnych (ICT): z jakiekolwiek miejsca, niezależnie od łączności dla każdego, będziemy dostęp do każdej rzeczy. Liczba połączeń będzie tworzyć zupełnie nową dynamiczną sieć sieci - Internet rzeczy". Przy czym istotny jest fakt, że nie chodzi tylko o tag'owanie RFID, ale także oznaczanie dużej liczby różnych obiektów – jednoznacznie adresowalnych, które stanowią podstawową strukturę IoT [3].

2.1. Stos technologii

Typowy stos technologii rozwiązania IoT składa się z trzech ogólnych warstw

Rysunek 1. Stos technologii internetowych rzeczy [16]

- Czujniki – osadzone w urządzeniu lub środowisku fizycznym. Rola czujników polega na przechwytywaniu ważnych zdarzeń lub surowych danych,
- Sieć i przetwarzanie – głównym zadaniem tej warstwy jest przechwytywanie danych czujników i odpowiednio reagowanie na nie,
- Platforma usług – jest poziomem, który odpowiada za agregaty i analizę danych. Poziom platformy usług jest dedykowany dla użytkowników końcowych rozwiązań IoT [16].

Prezentowany stos technologiczny może być realizowany w wielu różnych scenariuszach zarówno przemysłowych, jak i typowych – związanych z życiem

codziennym ludzi. Wartość dodana rozwiązania IoT jest budowana przez wszystkie warstwy z poziomu fizycznego czujnika poziomu do poziomu usług cyfrowych, gdzie użytkownik korzysta z zaawansowanych funkcji.

W zależności od producenta poszczególne stosy IoT są odpowiednio rozbudowane, aby obejmować nowe podsieci lub interfejsy tych usług. Te miniaturowe urządzenia zwane węzłami łączącymi tworzą bezprzewodowe sieci czujników [6].

Od linii produkcyjnych i magazynowych do dostaw detalicznych, regałów sklepowych, opieki zdrowotnej, inteligentnych miast – IoT przekształca procesy biznesowe dzięki zapewnieniu dokładniejszej i aktualnej widoczności przepływu materiałów i produktów. Niemniej jednak, w celu zapewnienia pełnej wizji IoT istotne znaczenie ma sprawne, bezpieczne, skalowalne i zorientowane na rynek, na sprzęt i przechowywanie danych. Tutaj niezbędna jest chmura obliczeniowa (ang. cloud computing, CC), który zapewnia wiarygodne usługi przetwarzania oparte na technologiach wirtualizacji pamięci masowej. CC działa jako odbiorca danych z wszechobecnych czujników. Stanowi wymiar analizy i interpretacji danych. A także jako wymiar, który dostarcza użytkownikowi dane w formie wizualizacji. Cloud computing stanowi doskonałe rozwiązanie dla obsługi ogromnych strumieni danych i przetwarzania ich dla nieprzygotowanej liczby urządzeń IoT i ludzi w czasie rzeczywistym [3].

2.2. Adaptacja technologii

Wraz z rozwojem IoT wzrasta, liczba firm i podmiotów zainteresowanych jego rozwiązaniami. W związku z tym przedmiotem zainteresowania staje się analiza kosztów IoT. Dzieje się to, ze względu na potencjalne, ale niepewne korzyści i wysokie koszty inwestycji w IoT. Firmy muszą dokładnie ocenić każdą otwieraną przez IoT możliwość tak, aby zapewnić, że ich zasoby (budżet, kapitał ludzki) będą wykorzystywane w sposób rozsądny [4].

Dlatego dzisiaj IoT ma status wzrastającej innowacji i dlatego konieczne jest zdefiniowanie ram, które będą wskazywały kierunek rozwoju IoT. Ramy, które zawierają elementy, które zwiększają liczbę i przekonują ludzi do korzystania z tego rodzaju technologii. Elementy, które skrócą czas dyfuzji IoT wśród użytkowników od innowatorów do większości użytkowników i w ten sam sposób skrócą czas dojrzałości IoT. Ramy, które spełniają te wymogi, określane są jako model dojrzałości.

3. Poziom dorosłości

Zagadnienie poziomu dojrzałości w literaturze reprezentowana jest przez koncepcję modelu dojrzałości. Model dojrzałości jest narzędziem biznesowym służącym do oceny ludzi/kultury, procesów/struktur oraz obiektów/technologii [7]. Istnieją dwa główne podejścia do projektowania modeli dojrzałości – góra-dół, bądź dół-góra. Podejście od góry do dołu zakłada ustaloną liczbę etapów lub poziomów zapadalności, które zostały określone w pierwszej kolejności, które są związane z charakterystyką (zazwyczaj w formie konkretnych pozycji oceny), oraz które wspierają początkowe założenia dotyczące rozwoju dojrzałości [1]. W podejściu oddolnym określane są różne cechy charakterystyczne lub pozycje oceny, a następnie skupione w drugim

etapie na poziomach dojrzałości, aby uzyskać bardziej ogólny pogląd na różne etapy ewolucji dojrzałości [5].

W wyniku tych dwóch podejść powstało wiele różnych wzorców [9-14, 18, 20]. Z tego powodu można znaleźć wiele różnych modeli identyfikujących różne aspekty technologii lub organizacji. Dla przykładu jednym z najbardziej znanych jest integracja modelu dojrzałości zdolności (ang. Capability Maturity Model Integration, CMMI), który jest programem doskonalenia i oceny postępów w procesie. Model został opracowany w celu poprawy użyteczności modeli dojrzałości – łącząc wiele różnych modeli w jedną strukturę. CMMI ma pięć poziomów: początkowy, zarządzany, definiowany, ilościowo zarządzany, optymalizujący, które wskazują stan procesów firmy. Modele CMMI i zasadniczo modele dojrzałości stanowią w większości ścieżkę ewolucji a nawet plan poprawy dla przedsiębiorstwa [11]. Z drugiej strony dają także menedżerom i użytkownikom szkielet wskazujący szereg kryteriów klasyfikacji, które pozwalają użytkownikowi na określenie, jaki etap rozwoju jest reprezentowany przez rozwiązanie, technologię, proces lub firmę. Ponadto, służą jako ramy do wyboru właściwej ścieżki przyszłego rozwoju lub jako wzorzec korzystania z różnych doświadczeń innych firm – benchmark. Tak więc modele dojrzałości działają jako wskazówki pokazujące czego należy się uczyć, jaką strukturę wprowadzać tak by akcelerować rozwój organizacji. Co więcej, prawdziwy wynik oceny modelu dojrzałości nie jest reprezentowany przez badany poziom, lecz przez wykaz rzeczy, które należy podjąć w celu jego poprawy i zwiększenia jego potencjału [13].

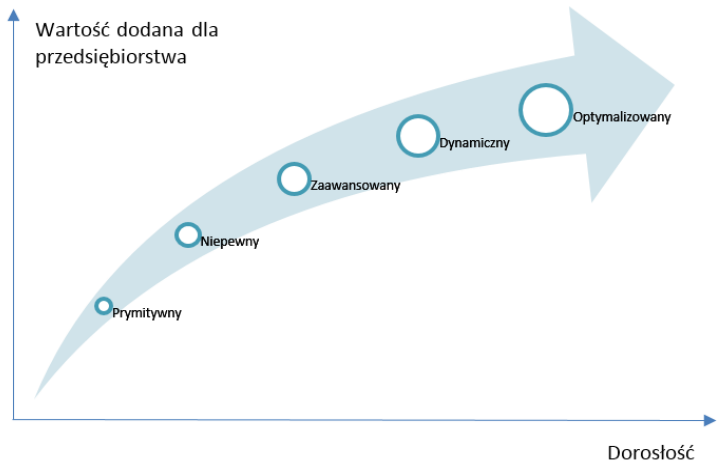
Warto wspomnieć, że każdy model dojrzałości, podobnie jak każdy model, jest uproszczeniem, które jest w jakiś sposób złe, ale oczekiwane. Czasami nawet surowy model może pomóc w odkryciu tego, co stanowi następny krok dla organizacji, technologii, czy też grupy rozwiązań.

W związku z tym podjęto dyskusję, która odpowiada na pytania dotyczące sposobu zorganizowania modeli dojrzałości dla rozwiązań wykorzystujących IoT i określonych przez nie kryteriów. Z tego powodu przedstawiono poniżej kilka modeli.

3.1. Modele dojrzałości IoT

Ze względu na intensywny rozwój IoT niniejszy rozdział wskazuje kilka różnych modeli dojrzałości. Modele podkreślają inne elementy i dotyczą wielu aspektów i obszarów IoT takich jak analityka, adaptacja czy przemysłowy IoT. Modele prezentowane są szczegółowo przez autorów w ramach cytowanych publikacji [9-14, 14, 18-20].

Pierwszy zidentyfikowany model proponuje Tony Show (Rysunek 2). Autor swój opis dojrzałości IoT nazywa IoT Maturity Model (IoTMM). Model jest jakościową metodą mierzenia wzrostu i zwiększania wpływu funkcji IoT w środowisku informatycznym z punktu widzenia perspektyw gospodarczych i technologicznych. Zawiera zestaw kryteriów, parametrów i czynników, które mogą być użyte do opisu i pomiaru skuteczności przyjęcia i wdrożenia programu IoT. Wykorzystanie IoTMM pozwala firmie na ocenę metod, procesów i operacji, w oparciu o wyraźny zestaw obiektywnych poziomów odniesienia dostosowanych do standardów branżowych i najlepszych praktyk. Moment spełnienia kryteriów danego etapu jest mierzony według dopasowania do określonego poziomu [10, 12].



Rysunek 2. Model dojrzałości IoT Tony Show [10]

Model osłabia, że wraz ze wzrostem dojrzałości obserwuje się wzrost wartości dodanej rozwiązania dla firmy. Definiuje pięć poziomów:

- Prymitywny – początkowy etap rozłącznych niezorganizowanych działań: izolowane czujniki, izolowane aplikacje M2M, ograniczone funkcje,
- Wstępne – eksperymenty ad hoc dotyczące prób i błędów z pewnym poziomem łączności: połączone urządzenia, jednosteki międzykanałowe, lekkie protokoły
- Zaawansowane – kompleksowe framework'i i cykle życia efektywnego zarządzania wdrożeniami i usługami: usługi zarządzane, zdalne zarządzanie, niezawodna jakość usług,
- Dynamiczne – wyrafinowane analizy i konsekwentne operacje za pomocą dyscyplin architektury i najlepszych praktyk: inteligentna decyzja, analiza kontekstowa, gromadzenie wiedzy i spostrzeżeń.
- Zoptymalizowana – platforma konwergentna i jednolity stos techniczny z powtarzalnym procesem i kodyfikacją opartą na zasadach: współdziałanie z innymi dyscyplinami, unifikowane inteligentne rozwiązania [17].

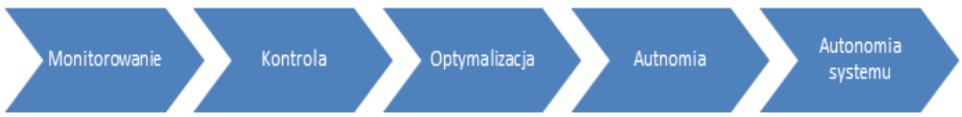
Inne podejście do struktury modelu dojrzałości jest przedstawione w badaniach nad przyjęciem "Internetu rzeczy" wśród duńskich firm przygotowanych przez firmę Ericsson we współpracy z Deloitte i Digital [20]. Opisany model zakłada, że IoT przynosi wartość zarówno z wewnętrznego wzrostu wydajności, jak np.: w łańcuchu dostaw i produkcji, jak i ofertowaniu.

Proponowany model (Rysunek 3.) składa się z pięciu etapów dojrzałości:

- Monitorowanie – używanie czujników, urządzeń obsługujących protokół IoT umożliwia znaczne monitorowanie stanu produktu, środowiska zewnętrznego i używania produktu,
- Sterowanie – funkcje produktu reagują na określone zmiany w jego stanie lub otoczeniu (jeśli X występuje, wykonano Y), używając algorytmów i oprogramowania,

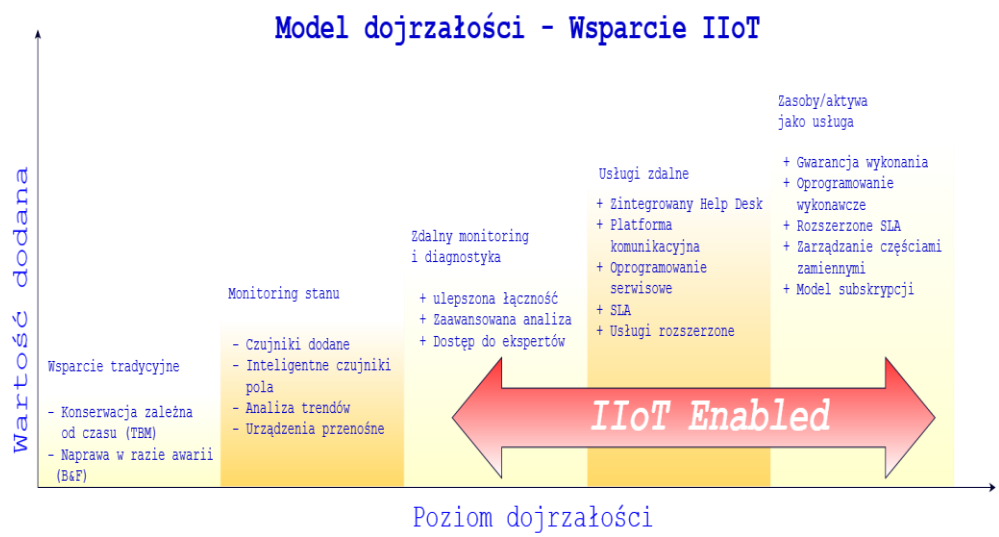
- Optymalizacja – analityka umożliwia produktowi ciągłe i automatyczne optymalizację wydajności,
- Autonomia – praca bez ingerencji człowieka, ciągle dostosowująca się do danych dotyczących środowiska i preferencji użytkowników,

Autonomia systemu – zdolność produktu do ciągłego dialogu z innymi powiązanymi rzeczami, wpływająca na funkcję obu [20].



Rysunek 4. Model dojrzałości IoT firmy Ericsson [20]

Przykładem modelu dojrzałości, który koncentruje się na dokładnym odcinku rozwiązań IoT jest model dojarzałości przemysłowego internetu rzeczy (ang. Industrial Internet of Things, IIoT) prezentowany przez ARC Adivisory Group [14]. Model ten (Rysunek 4.) prezentuje, w jaki sposób rozwiązania IIoT tworzą nowe możliwości lub rozszerzają usługi, gdzie IIoT stymulują rozwój każdego aspektu na drodze do określonego poziomu dorosłości. Dlatego też, IIoT może redefiniować oczekiwania firm dotyczące wydajności aktywów. Ponadto IIoT umożliwia nowe modele biznesowe i potencjalnie przekształca relację dostawcy/klienta [14].



Rysunek 4. Model dojrzałości uaktywnienia usług IIoT[14]

Istnieje wiele propozycji różnych modeli – niestety większość z nich nie została przetestowana i odpowiednio opisana. Często modele są tworzone odpowiednio do potrzeb użytkowników i organizacji. Tym samym można oczekiwać dalszego rozwoju modelów dorosłości.

To co łączy wskazane modele to fakt, że stopień dorosłości Internetu rzeczy wyraża się w liczbie rozwiązań występujących w określonych kategoriach. Ponadto, co jest zgodne z wszystkimi przedstawionymi modelami to fakt, że obecnie największa grupa rozwiązań może być klasyfikowana do trzech pierwszych stopni każdego modelu. Największą grupą rozwiązań są te, które monitorują stan, agregują i prezentują dane. Mniejsze liczba rozwiązań znajduje się na wyższych poziomach gdzie wykorzystywana jest analiza danych i metody sztucznej inteligencji. Jest więc oczywiste, że obecny stopień dojrzałości rozwiązań IoT znajduje się na średnim poziomie, który ma swoją reprezentację mniej lub bardziej na środku ścieżki wskazanej przez prezentowane modele dojrzałości. Dokładna determinacja wymaga opracowania ogólnego modelu dojrzałości i dalszych badań IoT.

4. Potencjał rozwiązań

Potencjał rozwiązań technologii IoT jest trudny do wskazania z powodu szumu jaki został wytworzony w wokół niego. Dlatego na potrzeby tego badania przyszły potencjał IoT będzie charakteryzował się w dwóch wymiarach. Jednym z aspektów są obecne trendy, a drugi – kluczowe wyzwania, które wymagają dalszego rozwoju w ramach tej technologii.

4.1. Trendy

Internet of Things został zidentyfikowany jako jedna z nowych technologii informatycznych, jak odnotowano w cyklu Hype Cycle Gartnera [15]. Hype Cycle to sposób na przedstawienie powstawania, przyjęcia, dojrzałości i wpływu na zastosowania konkretnych technologii. Przewiduje się, że IoT zajmie 5-10 lat w celu przyjęcia na rynek.

Popularność różnych paradygmatów różni się w zależności od czasu. Niemniej jednak popularność wyszukiwania frazy hasła „Internetu rzeczy", w ciągu ostatnich 10 lat, stale rośnie.[19]. Opierając się na danych i badaniach rynku, tendencja ta będzie prawdopodobnie kontynuowana [7].

Podsumowując, tendencje napędzające potencjał IoT są reprezentowane w obszarach:

- Przetwarzanie i udostępnianie danych – możliwość przetwarzania wielu zewnętrznych strumieni danych w celu wzbogacenia zawartości lokalnej. Umożliwiłoby to także użytkownikom publikowanie niektórych swoich strumieni do osób trzecich,
- Wsparcie dla programistów – możliwość tworzenia aplikacji na nowe produkty, które docierają do szerokiego grona klientów,
- Tworzenie ekosystemu – zwiększanie świadomości o nowych nurtach i możliwości tworzenia nowych modeli biznesowych,
- Rynki i rozliczenia – zdolność do wyszukiwania/poszukiwania danych i aplikacji oraz sprzedaży/nabycia praw do ich wykorzystania [8].

4.2. Wyzwania

W tej sekcji omówiono techniczne i menedżerskie wyzwania związane z zarządzaniem danymi, eksploracją danych, prywatnością, bezpieczeństwem i chaosem opisywanym w literaturze. Wyzwania te są głównie związane z rozwojem IoT przez przedsiębiorstwa. Zidentyfikowane wyzwania związane są głównie z wybuchem danych generowanych przez maszyny IoT.

- Wyzwanie w zarządzaniu danymi - czujniki i urządzenia IoT generują ogromne ilości danych, które muszą być przetwarzane i przechowywane. Niewiele przedsiębiorstw mogłoby zainwestować w wystarczającą ilość danych, aby przechowywać wszystkie dane IoT zebrane w sieci. W rezultacie potrzebne będą bardziej rozproszone centra danych, aby dostarczać coraz większej liczby danych urządzenia IoT.
- Wyzwanie związane z wydaniem danych - ze względu na to, że dane są dostępne do przetwarzania i analizy, użycie narzędzi do wyszukiwania danych staje się koniecznością. Dane składają się nie tylko z tradycyjnych danych dyskretnych, ale także z transmisji danych. Narzędzia do wyszukiwania danych mogą wywoływać procesy korygujące, ale muszą być dostosowane do zaawansowanych narzędzi do kopiowania danych w celu kopiowania danych strumieniowych z sieci czujników i danych obrazu i obrazu, brakuje właściwych analityków danych.
- Wyzwanie dotyczące prywatności - jak w przypadku inteligentnych urządzeń medycznych i inteligentnych służb ratowniczych w samochodzie, urządzenia IoT mogą dostarczyć ogromną ilość danych dotyczących lokalizacji i ruchów użytkowników, ich stanów zdrowotnych i preferencji zakupowych użytkowników, co może powodować znaczne obawy dotyczące prywatności Które muszą być uwzględnione w rozwiązaniach IOT.
- Wyzwanie związane z bezpieczeństwem - wraz z rosnącą liczbą i różnorodnością powiązanych urządzeń są wprowadzane do sieci IoT, eskaluje się potencjalne zagrożenie dla bezpieczeństwa. Brak bezpieczeństwa i prywatności spowoduje opór wobec adopcji IoT przez osoby fizyczne i prawne. Wyzwania związane z bezpieczeństwem informacji mogą zostać rozwiązane przez przeszkolenie programistów w firmowych rozwiązaniach zabezpieczających (np. Systemach zapobiegania włamaniom, łączeniach) w produktach i zachęcanie użytkowników do korzystania z wbudowanych w ich urządzenia funkcji bezpieczeństwa IoT.
- Wyzwanie chaosu – ewolucja technologii IoT (np. czujniki, technologie bezprzewodowe) jest w hiper przyspieszonym cyklu innowacji znacznie szybszym niż typowy cykl innowacji dotyczących produktów konsumenckich. Nadal istnieją konkurencyjne standardy, niewystarczające bezpieczeństwo, kwestie prywatności, złożona komunikacja i proliferacja numerów słabo testowanych urządzeń. Jeśli nie wyrejestrowano się ostrożnie, urządzenia wielofunkcyjne i aplikacje współpracujące mogą przekształcić nasze życie w chaos [6].
- IoT i sieci społeczne - Obiekty mogą stać się "przyjaciółmi" i mogą tworzyć grupy społeczne autonomicznie, z korzyścią dla człowieka, ale bez ich interwencji.

Rozwiązania IoT ze względu na niewielki rozmiar, koszt i rozwój technologii przetwarzania – cloud computing – umożliwiają pozyskiwanie danych z obszarów niedostępnych. To pozwala by sukcesywnie powiększać wymiar wiedzy jawnej w stosunku do ukrytej zaszytej w szczegółach procesów i kapitale intelektualnym pracowników [2].

5. Wnioski

Internet rzeczy to wschodzący paradygmat mający wpływ na wiele domen życia, dostarczając nowe, ewoluujące dane i wymagane zasoby obliczeniowe do tworzenia rewolucyjnych rozwiązań [8].

Dlatego istnieje silna potrzeba budowy modeli w celu wsparcia przedsiębiorstw i użytkowników w zakresie wyboru odpowiednich technologii. Niemniej jednak, każdy model jest używany do zrozumienia aktualnej sytuacji i wnioskowania. Dlatego kluczowe jest by zweryfikować to czy model jest odpowiedni do specjalnych okoliczności. Z tego powodu, jeśli wybrany model nie reprezentuje odpowiedniej ścieżki, to nie oznacza złego modelu, ale oznacza to, że nie jest odpowiedni dla tej sytuacji, produktu lub firmy. Dlatego przed wdrożeniem modelu w ramach procedur firmy istnieje obowiązek szczególnej uwagi w ocenie sprawności modelu.

Między innymi dlatego w niniejszym artykule przedstawia się kilka przykładów modeli jako dowód na to, że obecny stopień dojrzałości rozwiązań IoT znajduje się na średnim poziomie. Ma to swoją reprezentację mniej więcej na środku ścieżki wskazanej przez prezentowane modele dojrzałości. Dokładna determinacja wymaga opracowania ogólnego modelu dojrzałości i dalszych badań nad rozwiązaniami IoT.

Literatura

1. Becker, J., Knackstedt, R., Pöppelbuß, J., *Developing Maturity Models for IT Management - A Procedure Model and its Application*. Business & Information Systems Engineering 1(3), 213-222.
2. Borgia E., *The Internet of Things vision: Key features, applications and open issues, Computer Communications*, Volume 54, 1 December 2014, Pages 1-31.
3. Gubbi J., Buyya R., Marusic S., Palaniswami M., *Internet of Things (IoT): A vision, architectural elements, and future directions, Future Generation Computer Systems*, Volume 29, Issue 7, September 2013, Pages 1645-1660.
4. Hsu C., Lin J., *An empirical examination of consumer adoption of Internet of Things services: Network externalities and concern for information privacy perspectives, Computers in Human Behavior*, Volume 62, September 2016, Pages 516-527.
5. Lahrmann G., Marx F., Mettler T., Winter R.,Wortmann F., *Inductive design of maturity models: applying the Rasch algorithm for design science research, Proceedings of the 6th international conference on Service-oriented perspectives in design science research*, Springer-Verlag 2011, Pages 176-191.
6. Lee I., Lee K., *The Internet of Things (IoT): Applications, investments, and challenges for enterprises, Business Horizons*, Volume 58, Issue 4, July–August 2015, Pages 431-440.
7. Mettler T., *Maturity assessment models: a design science research approach,International* Journal of Society Systems Science 2011 3:1-2, Pages 81-98.
8. Mineraud J., Mazhelis O., Su X., Tarkoma S., *A gap analysis of Internet-of-Things platforms,* Computer Communications, March 2016.

9. http://blogs.microsoft.com/iot/2015/05/07/the-microsoft-vision-for-iot-and-what-it-could-mean-for-retail/.
10. http://cloudonomic.blogspot.com/2015/02/internet-of-things-maturity-model.html.
11. http://cmmiinstitute.com/.
12. http://java.sys-con.com/node/3293684..
13. http://martinfowler.com/bliki/MaturityModel.html
14. http://www.arcweb.com/events/arc-industry-forum-orlando/arcindustryforum2015presentations/GGorbach-ARC-IIoT%20The%20New%20Frontier.pdf.
15. http://www.gartner.com/newsroom/id/3114217.
16. http://www.pwc.com/us/en/technology-forecast/2013/issue1/features/technology-guide-customers-toward-goals.html.
17. http://www.techmahindra.com/sites/blogs/IOT_SOLUTIONS_MATURITY_MODEL.aspx.
18. http://www.vitria.com/wp-content/uploads/2016/02/Analytics-Maturity-Model-for-IoT-2pger-v02.pdf.
19. https://www.google.com/trends/explore?q=%22internet%20of%20things%22.
20. https://www2.deloitte.com/content/dam/Deloitte/dk/Documents/strategy/IoT-Report.pdf.

Modele dojrzałości i potencjał rozwiązań Internetu rzeczy

Streszczenie
Internet przedmiotów (IoT) jest staje się zjawiskiem, które ma coraz większy wpływ na globalny przemysł. Ze względu na rosnące zainteresowanie IoT ważne jest, aby mieć narzędzie, które wskaże poziom dorosłości rozwiązań IoT. Ma to szczególne znaczenie zarówno dla firm, które poszukują odpowiedniej technologii, które będą stymulować rozwój firmy jak i dla innowatorów poszukujących unikalnej wartości dodanej dla swoich pomysłów. Z tego powodu ten artykuł przedstawia dyskusję w kwestii modeli dorosłości (ang. maturnity model) oraz wskazuje potencjał IoT w wymiarze nadchodzących wyzwań jak i trendów.
Słowa kluczowe: Internet rzeczy, inżynieria produkcji, modele dojrzałości

Maturnity models and potential solutions Internet of things

Abstract
Internet of things (IoT) paradigm is phenomena that has increasing influence on global industry. Because of growing attention to IoT it is important to have framework that indicates level of maturnity used and proposed by IoT solutions. This is particularly significant both for companies that look for the right technology that will stimulate company development and innovators that looks for unique added value to their ideas. Therefore this article discuss issue of maturnity models and indicate future potential in dimension challenges and trends concerning IoT.
Keywords: Internet of things, .production engineering, maturnity models paper

Jakub Pizoń[1], Tomasz Cieplak[2]

Analiza i monitorowanie danych produkcyjnych z wykorzystaniem strumieniowego przetwarzania danych

1. Wstęp

Nowoczesne systemy zarządzania i monitorowania produkcji działają w określonych pozycjach i obszarach operacyjnych. Dzięki temu tworzą dedykowaną i hermetycznie zamkniętą architekturę systemów - wdrożone bez możliwości wymiany danych w otwartym formacie. W wyniku tego powstaje wiele niezintegrowanych źródeł danych. To sprawia, że trudne lub nawet niemożliwe jest przeanalizowanie i identyfikacja kluczowych czynników wydajności – bądź wychwycenia zmian – w czasie rzeczywistym. Jest to poważne ryzyko, które występuję pomimo tego, że wykorzystywane są zaawansowane systemy informatyczne wspierające procesy produkcyjne.

Ponadto, biorąc pod uwagę globalną perspektywę, konieczne jest zwrócenie uwagi na fakt, że w sektorze produkcji, w której firmy mają podobne maszyny, technologie produkcyjne i zasoby ludzkie, konkurencja przyjmuje zupełnie nowy wymiar. Kluczowy staje się sam proces organizacji produkcji i jego optymalizacja, właśnie to stanowi element przewagi nad innymi firmami.

Dlatego też, istnieje potrzeba analizy bieżących danych, aby wybrać najważniejsze atrybuty zgodnie z logiką produkcji w celu zapewnienia pełnej informacji zarządczej. To może stanowić podstawę skutecznych decyzji, które podnoszą jakość i obniżają koszty jakości. To otwiera bardzo dużą perspektywę dla wdrożenia ciągłego monitorowania i analizy strumieni danych produkcyjnych przy użyciu strumieniowego przetwarzania danych.

Dlatego też celem tego badania jest zaproponowanie modelu systemu sterowania procesem produkcyjnym przydatnego do monitorowania procesu produkcyjnego przy użyciu technologii Stream Computing.

Hipoteza badawcza określona w tym artykule koncentruje się na odpowiedziach na pytania, czy stosowany jest system produkcyjny oparty na zintegrowanych danych produkcyjnych w czasie rzeczywistym, umożliwia znaczne zmniejszenie czasu obecności składników produktu w środowisku produkcyjnym. Niniejszy artykuł stanowi fragment wstępnej analizy prototypu systemu ciągłego monitorowania i analizy strumieni danych produkcyjnych z wykorzystaniem strumieniowego przetwarzania danych.

2. Rola technologii informacyjnych w przemyśle

Cyberfizyczne systemy (ang. Cyber-Physical Systems, CPS) stanowią ścisłą integrację warstwy obliczeń i procesów fizycznych. Najczęściej występują w postaci

[1] jakub.pizon@pollub.edu.pl, Doktorant na Wydziale Mechanicznym Politechniki Lubelskiej, Wydział Mechaniczny, Politechnika Lubelska, www.pollub.pl
[2] t.cieplak@pollub.pl, Katedra Organizacji Przedsiębiorstwa, Wydział Zarządzania, Politechnika Lubelska, www.pollub.pl

wbudowanych komputerów oraz sieci monitorowania i kontroli procesów fizycznych działających w sprzężeniu zwrotnym, gdzie procesy fizyczne wpływają na obliczenia i na odwrót [4].

Cyberfizyczne systemy są identyfikowane z przejawem czwartej rewolucji przemysłowej, która ma miejsce w czasach nowożytnych [5]. Tym samym fizyczny zakres stosowania technologii jest ściśle związany z zastosowaniem urządzeń czujników i urządzeń pomiarowych wykorzystywanych szeroko w rozwiązaniach Internetu rzeczy (ang. Internet of things, IoT). Dlatego też kluczem do implementacji takich rozwiązań w środowisku produkcyjnym jest relacja CPS do powszechnie znanych zintegrowanych rozwiązań do zarządzania produkcją (ang. Computer Integrated Manufacturing, CIM).

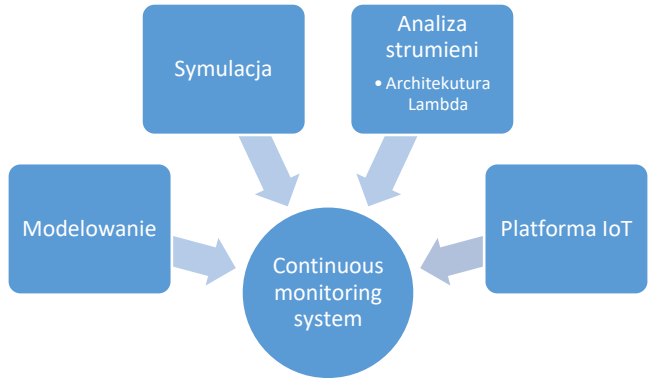
W rzeczywistości obecnie rozwiązania CIM są szeroko stosowane w przemyśle motoryzacyjnym, lotniczym, kosmicznym i stoczniowym. Warto wspomnieć, że rozwiązania te oferują przedsiębiorstwom pionową metodologię integracji systemów informacyjnych na różnych poziomach operacyjnych i zarządzania. W systemach CIM, automatyzacja produkcji odbywa się za pośrednictwem połączonych systemów na poziomie urządzenia – tworząc w ten sposób – lokalny system sieciowy. Taki system wbudowany w sieć umożliwia rozwiązanie tylko niektórych typowych problemów w konwencjonalnych systemach produkcji. Niemniej jednak ten system jest zamknięty, a jego funkcje i obowiązki są rozmieszczone w hierarchicznej strukturze, która jest mniej elastyczna w radzeniu sobie z dynamicznymi zmianami zewnętrznymi.

Z drugiej strony koncepcja CPS jest szeroko stosowana w przemyśle lotniczym, obronnym, energetyce, przemyśle transportowym (np.: inteligentnych pojazdach), przemyśle medycznym (np.: urządzeniach medycznych), budownictwie (np.: inteligentne domach). W tradycyjnym środowisku produkcyjnym, gdzie systemy automatyki, sterowania procesami i automatyki przemysłowej są uniwersalne, CPS może wspierać integrację pionową i horyzontalną systemów informatycznych, która integruje cały łańcuch dostaw lub nawet wszystkie branże branży. Daje to szeroką perspektywę do produkcji rozwiązań z dwuwymiarową integracją, zarówno w pionie, jak i w poziomie poziomym. Daje to podstawę dla budowy nowoczesnych systemów z pogranicza systemów produkcji i zarządzania, znanego jako systemy cyber fizyczne zarządzania produkcją (ang. Cyber-physical production systems systemy produkcji, CPPS) [8].

CPPS składają się z autonomicznych komponentów i podsystemów, które współpracują w sposób zależny od sytuacji na wszystkich etapach produkcji, od procesów realizowanych za pomocą urządzeń do produkcji i sieci logistycznych. Pojęcie systemów CPPS otwiera perspektywę architektoniczną dla zastosowań budowlanych w różnych dziedzinach. CPS i CM są siłami napędowymi Czwartej Rewolucji przemysłowej.

3. Metody i narzędzia

Poniższa sekcja opisuje metody stosowane w celu opisu modelu ciągłego systemu monitorowania produkcji.



Rysunek 1. Metody proponowane w modelu ciągłego systemu monitorowania produkcji

3.1. Modelowanie

Jedną z kluczowych metod wykorzystywanych do zweryfikowania proponowanego modelu jest modelowanie. Modelowanie jest definiowane jako próbę przedstawienia pewnego zjawiska lub własności, które staramy się zrozumieć lub zbadać w odniesieniu do innych zjawisk, które już się rozumieją. Dlatego ważne jest, aby podkreślić, że modele skomponowane nie modelują dokładnego obiektu, a modelują proces, który ma w nim miejsce.

Z tego powodu modelowanie wykorzystywane jest do odkrywania nowej wiedzy na temat procesów reprezentowanych przez uproszczony system, który odzwierciedla tylko pewne cechy procesu.

Zastosowanie modelowania można podzielić na dwie podstawowe grupy:

- Opis danych i odkrywanie ważnych relacji, regularności i wzorców,
- Predefiniowane wartości zmiennych wyjściowych.

W przypadku niniejszego artykułu modelowanie zastosowano w celu sprawdzenia postawionych tez. Po to by opracować logiczny model zintegrowanego systemu kontroli produkcji, który implementuje mechanizmy dynamicznej analizy strumieni danych (ang. Stream computing) w odniesieniu do danych historycznych uzyskanych z maszyn produkcyjnych firmy produkcyjnej.

Przygotowany model logiczny będzie używany do przygotowania systemu CPS przy użyciu języka programowania wysokiego poziomu.

3.2. Symulacja

Drugą metodą stosowaną do wdrożenia systemu prototypowego jest symulacja komputerowa. Głównym celem symulacji jest zweryfikowanie funkcjonalności modelu na podstawie generowanych danych.

Wdrożony system będzie testowany w środowisku symulacji przy użyciu danych testowych i rzeczywistych danych produkcyjnych. Ostatecznie system zostanie wdrożony z wykorzystaniem rozwiązań platformy Internetu oraz wybranych mikrokontrolerów i czujników.

3.3. Analiza strumieni

Kolejną metodą zaplanowaną w następnym teście jest analiza strumieni danych. Wybór takiego podejścia jest określony przez cechy środowiska produkcyjnego. Należy wziąć pod uwagę, że w przypadku zarządzania produkcją w czasie rzeczywistym kluczowa jest możliwość szybkiej analizy przepływów danych nazywanych strumieniami danych. Te strumienie danych zawierają istotne informacje, które mogą być ujawnione w wyniku wyszukiwania danych, przetwarzania lub innych metod analitycznych [2].

Analizy te można również przeprowadzić z udziałem standardowych systemów zarządzania produkcją, jednak najważniejsza jest szybkość przetwarzania danych. Tradycyjne systemy z relacyjnymi bazami danych wykorzystują przetwarzanie wsadowe. Główną ideą tego tradycyjnego podejścia do obsługi i analizy danych jest utrzymywanie danych w bazie danych i wykonywanie z nim zapytań. Istnieją jednak scenariusze, w których nie wszystkie dane mogą być przechowywane z powodu dużej ilości danych, lub gdy proces przetwarzania w czasie rzeczywistym jest potrzebny do szybkiego reagowania. Strumienie danych umożliwia przetwarzanie dużych ilości przejściowych danych w czasie rzeczywistym. W porównaniu z tradycyjnymi systemami zarządzania bazami danych, w których są wysyłane zapytania do danych, w strumieniowym przetwarzaniu dane są stosowane do ciągle wykonanych zapytań [2, 6].

W związku z tym konieczne jest użycie odpowiedniej zapytań do ciągłego przepływu danych produkcyjnych, które w nieregularny sposób występują w różnych formach. Potrzebne jest jeszcze to, aby móc odpowiednio zidentyfikować zdarzenia w celu podjęcia działań korygujących, naprawczych lub optymalizacji.

Dlatego też w celu zaprojektowania systemu przesyłania strumieniowego za pomocą architektury Lambda [3]. Architektura Lambda jest klasyfikowana metodami Big Data. Architektura Lambda jest przeznaczona do gromadzenia i analizowania danych w czasie rzeczywistym i historycznym. Architektura Lambda jest modelem ogólnego przeznaczenia do wykonywania dowolnej funkcji na dowolnym zestawie danych, która ma funkcję, która zwraca wyniki z niewielkim opóźnieniem. Zwykle wykorzystywane do analizy logów – zestawy danych uzyskane z systemów VPN, portali internetowych lub serwisów społecznościowych (Twitter, Facebook) i internetowych rozwiązań – porównywanych danych historycznych z bieżących danych [3]

System produkcyjny, podobnie jak inne systemy zmienia czas na wybielanie. Wprowadzane są nowe produkty, pojawiają się pewne opóźnienia dostawcy. Odtąd operatorzy zwykle nie mają żadnej kontroli nad tempem, w którym tworzone są zdarzenia. Tutaj platformy strumieniowe muszą być w stanie dostosować. Liczba zdarzeń w okresach o małej aktywności może znacznie różnić się od okresów szczytowych [2]. Jest to ściśle związane z kwestią elastyczności obliczeniowej. Elastyczność w obliczeniach definiuje dynamiczne zmiany zasobów komputerowych w celu radzenia sobie ze zmianami środowiskowymi, np. Różnymi obciążeniami pracą, zmianami cen zasobów informatycznych i zmieniającymi się wymaganiami użytkowników dotyczącymi jakości usług

Wczesne formy adaptacji obejmują strategie oparte na pojedynczych zdarzeniach lub strumieniach zdarzeń (np. Wyrzucanie obciążenia), a także bardziej zgrubną reorganizację platformy przetwarzania (np. Zoptymalizowane położenie operatora). Pojawienie się przetwarzania danych w chmurze popularyzowało podejście dosto-

sowane do zasobów, co pozwala na elastyczne przydzielanie i udostępnianie zasobów na żądanie odpowiednio do przepływu danych produkcyjnych. Dlatego też dla opisywanego analityki strumieniowego badania będą wdrażane w środowisku przetwarzania danych w chmurze przy użyciu proponowanych metod.

3.4. Platforma IoT

Internetem rzeczy (IoT) jest ostatnim elementem służącym do składania badań. Platforma IoT jest podstawą do wdrożenia systemu w rzeczywistym scenariuszu. Typowy IoT implementuje standardowy stos parapetu IoT składa się z trzech podstawowych warstw:

Czujniki - osadzone w urządzeniu lub środowisku fizycznym. Rola czujników polega na przechwytywaniu ważnych zdarzeń lub surowych danych,

Sieć i przetwarzanie – głównym zadaniem tej warstwy jest przechwytywanie danych czujników i odpowiednio reagujących na nie,

Platforma usług – jest poziomem, który odpowiada za agregowanie i analizy danych. Poziom platformy usług jest stale ustawiony dla użytkowników końcowych rozwiązań IoT [9].

Rysunek 2. Metody proponowania modelu ciągłego systemu monitorowania produkcji [9]

Prezentowany stos technologiczny może być realizowany w wielu odmiennych scenariuszach zarówno przemysłowych, jak i typowych – związanych z życiem codziennym ludzi. Opracowanie tego prostego przykładu zostanie przedstawione poniżej – czyli jazda na rowerze. Jak wspomniano, platforma IoT ma przetwarzać i analizować dane ze stosem technologii. W tym kontekście codziennej sytuacji – tutaj jazda na rowerze – warunek czujnika odpowiada za pomiar prędkości, współrzędne lokalizacji i inne wybrane parametry. Następnie na poziomie sieci z wykorzystaniem protokołów Bluetooth wysyłane są do smartfona, a następnie dane o połączeniach z Internetem są na ostatnim poziomie – to jest platforma usług – która analizuje dane.

Z drugiej strony, na czujnikach rowerów otrzymano dane o poziomie mikrosterownika (czujników), które są przetwarzane i wyświetlane – w tym parametrze jako szybkość, odległość przejechana, średnie wyniki w porównaniu do wcześniej ustalonych celów. Jednak w warstwie serwisowej zestaw różnych usług jest dostarczany

dokładnie dla użytkownika. Dlatego przy korzystaniu z platformy zewnętrznej użytkownik uzyskuje zyski z funkcjonalności w czasie rzeczywistym, które wspierają użytkownika poprzez wizualizację w czasie rzeczywistym przejechanej trasy, dostarczają informacji zwrotnych na temat aktualnych postępów, pokazuje postępy w sekcjach ("jazda pod górę", "jazda zjazdem") i wskazuje obszary dla ulepszenia.

Ten przykład pokazuje, że nawet niewielka ilość danych dotyczących cyfr może być używana do wydawania wielu usług, które są atrakcyjne dla użytkownika końcowego.

W ten sposób wartość dodana urządzeń używanych z czujnikami jest zbudowana we wszystkich warstwach od poziomu fizycznego czujnika poziomu do poziomu usług cyfrowych, gdzie użytkownik korzysta z zaawansowanych możliwości. Tak więc rozwiązania IoT są powszechnie nazywane także takimi, jak systemy cyber-fizyczne.

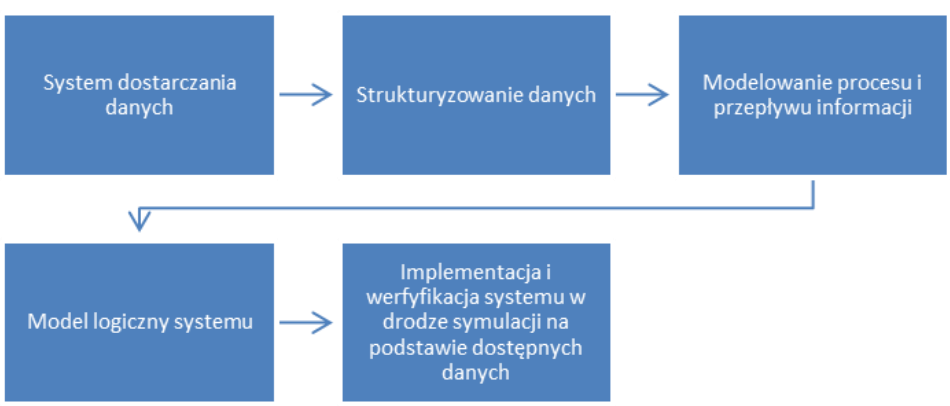
W zależności od producenta rozwiązań technologicznych IoT, poszczególne stosy są odpowiednio rozbudowane, aby obejmować nowe podsieci lub interfejsy tych usług [1]. Niniejsze opracowanie dotyczące potrzeb przeprowadzonego przeszukiwania będzie wykorzystywać platformę IBM Bluemix, która umożliwi przetwarzanie danych w czasie rzeczywistym.

4. Metodologia

W celu sprawdzenia postawionych tez i osiągnięcia zamierzonego celu zostanie opracowany logiczny model zintegrowanego systemu kontroli produkcji, który implementuje mechanizmy dynamicznej analizy strumieni danych w odniesieniu do danych historycznych uzyskanych z maszyn produkcyjnych firmy produkcyjnej.

Przygotowany model logiczny będzie używany do programowania systemu CPS przy użyciu języka programowania wysokiego poziomu. Wdrożony system zostanie przetestowany w środowisku symulacji przy użyciu danych testowych i faktycznej produkcji. Ostatecznie system zostanie wdrożony z wykorzystaniem rozwiązań IBM Bluemix oraz wybranych mikrokontrolerów i czujników.

Projekt zostanie przeprowadzony z wykorzystaniem metodologii rozwoju oprogramowania wodno-spadkowego w odniesieniu do rzeczywistych danych produkcyjnych wybranej firmy. Projekt będzie realizowany w następujących etapach:



Rysunek 3. Metodologia wykorzystywana do zaproponowania modelu strumieniowego przetwarzania danych

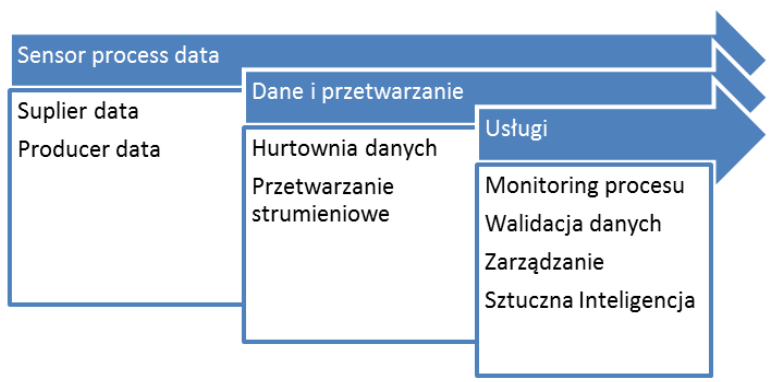
5. Model strumieniowego przetwarzania danych

Architektura logiczna systemów ciągłego monitorowania i analizy danych charakteryzuje się strukturą warstw, która zarządza przetwarzaniem danych produkcyjnych. Dlatego w strukturze można wyróżnić trzy poziomy

- pozyskiwanie danych,
- przechowywanie i przetwarzanie.
- analiza zarządzania danymi.

Na poziomie gromadzenia danych funkcje zbierania danych przez interfejsy danych zarówno od producenta, jak i dostawcy są wykonywane. Dane uzyskane z danych producenta są zarówno procesem sprawności, jak i danymi produkcyjnymi. Dane biznesowe są reprezentowane jako dane systemów zarządzania zasobami (ERP – Enterprise Resource Planning). Choć dane produkcyjne pochodzą bezpośrednio z maszyn sterowanych numerycznie, które wykonują proces produkcyjny. Wykorzystując dane z czujników lub elektronicznych znaczników (sygnalizatorów RFID) możliwe jest monitorowanie przepływu produktów wewnątrz firmy [7].

Z drugiej strony dane pochodzące od dostawców są w zasadzie opisane w szczegółach dotyczących zamówień lub danych produkcyjnych z przewidywaniem czasu wykonania zamówienia. Następny poziom wdrażany jest na poziomie systemów przetwarzania danych, które są postrzegane jako zestaw danych, które działają w pełni lub częściowo w obrębie rozwiązania cloud computing. Dlatego też można osadzić dane procesowe w ramach własnego producenta infrastruktury lub umieścić je w rozwiązaniu cloud. Najważniejsze jest to, że każda usługa korzystająca z danych w czasie rzeczywistym działa w ramach rozwiązania cloud computing [7].



Rysunek 4. Architektura logiczna systemu strumieniowego przetwarzania danych

Głównym zadaniem ostatniego poziomu monitorowania danych i analizy procesu produkcji jest interfejs zarządzania umożliwiający wizualizację bieżącej analizy danych – przy użyciu zaimplementowanych usług internetowych – znanych jako aplikacje. Interfejs zarządzania może być dostosowany i dostęp do funkcji może być ograniczony w zależności od potrzeb planistów lub menedżerów. Dlatego powinien on być dostępny w innej formie lub w systemie i być dostępny z urządzeń przenośnych i przeglądarki internetowej. Omówiono ogólne elementy logiki i poziomy logicznej architektury wskazanej na Rysunku 4.

6. Wnioski

Kluczowe oczekiwane rezultaty wdrożenia takiego modelu architektury stanowią:

- Próbę wykrycia ukrytych informacji w danych produkcyjnych,
- Dynamiczne dostosowywanie wydajności systemu produkcyjnego do zmiennej wydajności poszczególnych maszyn,
- Ograniczenie czasu przebywania produktu w systemie produkcyjnym,
- Ograniczenie kosztów i zwiększenie konkurencyjności.

Osiągnięcie oczekiwanych rezultatów pozytywnie wpłynie na rozwój nauki, identyfikując nową wiedzę na temat zakresu bieżących wymagań dotyczących rozwiązań systemów zarządzania produkcją wykorzystujących Internet w ramach systemów cyber-fizycznych.

Opracowana metodologia projektowania dostarczy naukowych odpowiedzi na pytanie, jakie rozwiązania i aplikacje, jakie algorytmy, czy metody analizy danych mają potencjał do opracowania i mogą być wykorzystane przy planowaniu nowoczesnych systemów zarządzania produkcją. Możliwości dają architekturę zaprojektowaną w ten sposób, jakie są ich zalety i wady.

Ponadto projekt przyczyni się do rozwoju komputerowych systemów sterowania, które są podstawą nowoczesnych metod produkcji – bezpośrednio odpowiadających potrzebom ludzi, przemysłu i całych społeczeństw.

Lepsze zarządzanie procesem produkcji oznacza niskie koszty jakości związane z przestojami, braki, a także oszczędza materiał i czas - co ma swoją reprezentację w cenie końcowego dostarczanego produktu. Tańsze produkty o odpowiedniej jakości stanowią przewagę konkurencyjną iw perspektywie rozwoju gospodarczego stanowi podstawę dobrobytu społeczeństwa.

Podsumowując wskazany temat badawczy, według literatury, jest intensywnie rozwijającym się obszarem i stwarza wiele pytań. W związku z tym konieczne jest przeprowadzenie większej liczby badań w celu przetestowania proponowanej architektury logicznej w obliczu postrzeganych korzyści.

Literatura

1. Cieplak T.; Malec M. *Practical usage of cloud computing in computer integrated manufacturing, New methods in intustrial engineering and production management*, Politechnika Lubelska, 2012, Pages 7-18.
2. Hummer W. Satzger B.,Dustdar S., *Elastic stream processing in the Cloud, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Volume 3, 2013, Pages 333 – 345.
3. Kreps J. I Heart Logs: Event Data, Stream Processing, and Data Integration, O'Reilly Media, 2015
4. Lee E. Cyber *Physical Systems: Design Challenges, Electrical Engineering and Computer Sciences University of California at Berkele*, January 200
5. Lee J. *Bagheri B., Kao H., A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems, Manufacturing Letters*, Volume 3, January 2015, Pages 18-23.
6. Marz N., Warren J., Big Data – *Principles and best practices of scalable realtime data systems*, Manning Publications, 2015.

7. Pizoń J., Lipski J. *Manufacturing Process Support Using Artificial Intelligence,. Applied Mechanics and Materials*, Vol. 791, September, 2015, s. 89-95.
8. Yu C., Xu X., Lu Y., *Computer-Integrated Manufacturing, Cyber-Physical Systems and Cloud Manufacturing – Concepts and relationships*, Manufacturing Letters, Volume 6, October 2015, Pages 5-9.
9. The Thing Stack: *Technologies that guide customers to their goals*, [online],: http://www.pwc.com/us/en/technology-forecast/2013/issue1/features/technology-guide-customers-toward-goals.html.

Analiza i monitorowanie danych produkcyjnych z wykorzystaniem strumieniowego przetwarzania danych

Streszczenie
Współczesne systemy zarządzania produkcją oraz monitorowania procesów technologicznych działają w ramach konkretnych stanowisk operacyjnych – tworząc najczęściej dedykowane hermetycznie zamknięte architektury systemów informatycznych – implementowane bez możliwości wymiany danych w otwartym formacie. Duża liczba niezintegrowanych źródeł danych, bądź ich brak w kluczowych obszarach utrudnia, a nawet niemożliwa identyfikację przez planistę w czasie rzeczywistym istotnych zmian. Tym samym podejmuje on decyzje na podstawie nie pełnej informacji zarządczej, co jest to źródłem ryzyka podejmowania decyzji powodujących zwiększenie kosztów, jakości. Co z kolei w przypadku konkurencji w danym sektorze działalności wytwórczej, gdzie firmy dysponują podobnym parkiem maszynowym, technologiami produkcji jak i zasobami ludzki, może istotnie stanowić o przewadze, jednych przedsiębiorstw nad innymi. W związku z tym istnieje silna potrzeba analizy bieżących danych, aby wybrać najważniejsze atrybuty zgodnie z logiką procesu produkcji w celu zapewnienia pełnej informacji zarządczej. Analityka danych rzeczywistych może być bazą dla skutecznych decyzji, które zwiększają jakość i obniżają koszty jakości. To otwiera bardzo duże perspektywy dla wdrożeń ciągłego monitorowania i analizy strumieni danych produkcyjnych z wykorzystaniem metod przetwarzaniu strumieni. Dlatego też celem tego badania jest zaproponowanie modelu systemu sterowania procesami produkcyjnymi przydatnych do monitorowania procesu produkcyjnego z wykorzystaniem Stream Computing.
Słowa kluczowe: internet rzeczy, inżynieria produkcji, systemy cyber-fizyczne

Analysis and monitoring of production data with use of stream computing

Abstract
Modern production management and monitoring systems operate within specific operational positions and areas. Therefore they create dedicated and hermetically sealed architecture of systems - implemented without the possibility of exchanging data in an open format. As a result of that, a large number of non-integrated data sources is produced. That makes them difficult or even impossible to analyse and identify key factors of performance – to notice significant changes – in real time. That is a cause of great risk that companies needs to cover, even though they use advance IT software to support manufacture processes. Therefore there is a strong need to analyse ongoing data to pick the most important attributes in accordance with the manufacture logic to provide complete management information. That can be base for effective decisions that increase quality and lower the costs of quality. This opens up a very large perspective for deployments of continuous monitoring and analysis of streams of production data with use of stream computing. Therefore, the aim of this research is to propose a model of production process control system useful for monitoring the production process with use of Stream Computing.
Keywords: Internet of things, .production engineering cyber-physical systems, stream computing

Robert Kozakiewicz[1], Robert Lewoń[2], Michał Małafiejski[3]

Równowaga strategiczna dla zbiorów defensywnych w drzewach

1. Wstęp

Celem niniejszej pracy jest wprowadzenie pojęcia równowagi strategicznej dla zbiorów defensywnych oraz przedstawienie wyników badań nad tym zagadnieniem: dowodu na równoważność problemu równowagi strategicznej dla zbiorów defensywnych i problemu doskonałej równowagi strategicznej dla zbiorów defensywnych oraz algorytmu udowadniającego łatwość problemu równowagi strategicznej dla zbiorów defensywnych w grafach będących drzewami.

Struktury bezpieczne, do których zalicza się równowaga strategiczna dla zbiorów defensywnych, znajdują zastosowanie chociażby w takich dziedzinach jak przetwarzanie odporne na błędy [1] czy społeczności internetowe [2].

1.1. Definicja problemu

W niniejszej pracy będą rozważane jedynie proste, spójne, skończone, niepuste grafy. Będą używane standardowe oznaczenia z teorii grafów. Dla grafu G przez $V(G)$ oznaczany będzie zbiór wierzchołków grafu G, a przez $E(G)$ zbiór krawędzi grafu G. Przez $\deg v$ oznaczany będzie stopień wierzchołka v, czyli liczba jego sąsiadów, a przez Δ oznaczany będzie największy ze stopni wierzchołków danego grafu. Dla każdego wierzchołka $v \in V(G)$ zbiory $N_G(v) = \{u: u \in V(G) \wedge \{u, v\} \in E(G)\}$ oraz $N_G[v] = N_G(v) \cup \{v\}$ to odpowiednio *sąsiedztwo otwarte* i *domknięte* wierzchołka v w grafie G. Podobnie dla podzbioru $X \subseteq V(G)$ zbiory $N_G(X) = \bigcup_{v \in X} N_G(v)$ oraz $N_G[X] = N_G(X) \cup X$ to odpowiednio *sąsiedztwo otwarte* i *domknięte* zbioru X w grafie G. Dla podzbioru $S \subseteq V(G)$ oraz jego niepustego podzbioru $X \subseteq S$ zdefiniowany będzie predykat $SEC_{G,S}(X)$, który jest spełniony wtedy i tylko wtedy, gdy $|N_G[X] \cap S| \geq |N_G[X] \backslash S|$. W niniejszej pracy będzie używane skrócone oznaczenie $SEC(X)$ zamiast $SEC_{G,S}(X)$, gdy zbiór S oraz graf G będą jednoznacznie wynikać z kontekstu.

W niniejszej pracy będzie rozważany problem równowagi strategicznej dla zbiorów defensywnych. Pojęcie równowagi strategicznej zostało zaproponowane w [3] w odniesieniu do innej struktury bezpiecznej – koalicji. Pojęcie koalicji zostało wprowadzone w [4] oraz [5]. Pojęcie zbioru defensywnego zostało zaproponowane w [6]. Niepusty podzbiór wierzchołków $S \subseteq V(G)$ jest *zbiorem defensywnym* jeśli dla każdego wierzchołka $v \in S$ zachodzi $SEC(v)$ lub istnieje wierzchołek $u \in N_G(v)$ taki, że

[1] kozakiewicz.robert@gmail.com, Katedra Algorytmów i Modelowania Systemów, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
[2] lewon.robert@gmail.com, Katedra Algorytmów i Modelowania Systemów, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
[3] michal@animima.org, Katedra Algorytmów i Modelowania Systemów, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska

$SEC(\{u,v\})$. Zbiór defensywny jest globalny jeśli jest również zbiorem dominującym w grafie G, czyli $N_G[S] = V(G)$.

Dla podzbiorów wierzchołków grafu G: N, S oraz I, gdzie N oraz S są globalnymi zbiorami defensywnymi, para $(\{N,S\},I)$ nazywana jest *równowagą strategiczną dla zbiorów defensywnych*. Jeśli zbiór I jest pusty, równowaga strategiczna nazywana jest *doskonałą* i oznaczana jako $\{N,S\}$.

Zbiory N oraz S mogą być interpretowane jako dwie strony konfliktu, w którym pożądane jest, aby żadna ze stron nie uzyskała nad drugą przewagi. Oba te zbiory mają zapewnione bezpieczeństwo przez właściwość zbioru defensywnego oraz dominują cały graf. Natomiast zbiór I to wierzchołki, które nie wybrały żadnej ze strony konfliktu, jednak potencjalnie mogą dołączyć do każdej z grup. Dlatego wierzchołki ze zbiorów N i S muszą być gotowe na najgorszy scenariusz, czyli atak wierzchołków z przeciwnego zbioru oraz z I.

Problem istnienia równowagi strategicznej dla zbiorów defensywnych w grafie nazywany jest *problemem równowagi strategicznej dla zbiorów defensywnych* (*DSSB*). Analogiczne, problem istnienia doskonałej równowagi strategicznej dla zbiorów defensywnych w grafie nazywany jest *problemem doskonałej równowagi strategicznej dla zbiorów defensywnych* (*DSPSB*).

1.2. Powiązane rezultaty badań

W pracy [3] autorzy zajmowali się problemem równowagi strategicznej dla koalicji. Udowodnili oni równoważność pomiędzy problemem równowagi strategicznej a problemem doskonałej równowagi strategicznej dla koalicji oraz zaproponowali wielomianowy algorytm konstruujący doskonałą równowagę strategiczną dla drzew. W niniejszej pracy wykazana zostanie równoważność problemu równowagi strategicznej oraz problemu doskonałej równowagi strategicznej dla zbiorów defensywnych. Zaproponowany zostanie również wielomianowy algorytm konstruujący doskonałą równowagę strategiczną dla zbiorów defensywnych w drzewach.

2. Równowaga strategiczna a doskonała równowaga strategiczna

W tym rozdziale pokazana zostanie równoważność między problemem równowagi strategicznej dla zbiorów defensywnych a problemem doskonałej równowagi strategicznej dla zbiorów defensywnych.

Dla uproszczenia oznaczeń, dla danego grafu G oraz podzbioru jego wierzchołków $A \subseteq V(G)$ zdefiniowane zostaną symbole $Av = N_G[v] \cap A$ oraz $A\{u,v\} = N_G[\{u,v\}] \cap A$.

Lemat 1: Dla dowolnych $A \subseteq B \subseteq V(G)$, dla każdego $v \in V(G)$ zachodzi $|Bv| \geq |Av|$oraz dla każdego $u \in V(G)$ takiego, że $\{u,v\} \in E(G)$ zachodzi $|B\{u,v\}| \geq |A\{u,v\}|$.

Dowód: Ponieważ $A \subseteq B$, zachodzi $|Bv| = |B \cap N_G[v]| \geq |A \cap N_G[v]| = |Av|$ oraz $|B\{u,v\}| = |B \cap N_G[\{u,v\}]| \geq |A \cap N_G[\{u,v\}]| = |A\{u,v\}|$.■

Lemat 2: Dla każdych dwóch (globalnych) zbiorów defensywnych A i B w grafie G, $A \cup B$ jest (globalnym) zbiorem defensywnym w G.

Dowód: Niech $v \in A \cup B$ oraz bez straty ogólności przyjęte zostanie założenie, że $v \in A$. Z lematu 1 wiadomo, że $|(A \cup B)v| \geq |Av|$ oraz $|(A \cup B)\{u,v\}| \geq |A\{u,v\}|$ dla każdego $u \in V(G)$. Z lematu 1 wiadomo również, że $|(V(G)\backslash A)v| \geq |(V(G)\backslash(A \cup B))v|$ oraz $|(V(G)\backslash A)\{u,v\}| \geq |(V(G)\backslash(A \cup B))\{u,v\}|$ dla każdego $u \in V(G)$, zatem $A \cup B$ jest zbiorem defensywnym. Jeśli A i B są zbiorami dominującymi w G, ich suma w oczywisty sposób również jest zbiorem dominującym w G.■

Lemat 3: Niech $N \subseteq V(G)$ będzie zbiorem defensywnym. Dla każdego wierzchołka $v \in V(G)\backslash N$, jeśli $|Nv| > |(V(G)\backslash N)v|$ oraz dla każdego wierzchołka $u \in V(G)\backslash N$, takiego że $\{u,v\} \in E(G)$ zachodzi $|N\{u,v\}| > |(V(G)\backslash N)\{u,v\}|$, to nie istnieje zbiór defensywny $S \subseteq V(G)\backslash N$ zawierający wierzchołek v, a $N \cup \{v\}$ jest globalnym zbiorem defensywnym.

Dowód: Niech $v \in V(G)$, $|Nv| > |(V(G)\backslash N)v|$ oraz dla każdego $u \in V(G)\backslash N$, takiego że $\{u,v\} \in E(G)$ zachodzi $|N\{u,v\}| > |(V\ (G)\backslash N)\{u,v\}|$. Przyjęte zostanie założenie, że istnieje zbiór defensywny $S \subseteq V(G)\backslash N$, taki że $v \in S$. Z definicji zbioru defensywnego, dla każdego $v \in S$ zachodzi $|Sv| \geq |(V(G)\backslash S)v|$ lub istnieje wierzchołek $u \in S$, taki że $\{u,v\} \in E(G)$ oraz $|S\{u,v\}| \geq |(V(G)\backslash S)\{u,v\}|$. Z lematu 1 wiadomo, że $|(V(G)\backslash N)v| \geq |Sv|$ oraz $|(V(G)\backslash N)\{u,v\}| \geq |S\{u,v\}|$ dla każdego $u \in V(G)$, takiego że $\{u,v\} \in E(G)$. Ponieważ $N \subseteq V(G)\backslash S$, z lematu 1 wiadomo, że $|(V(G)\backslash S)v| \geq |Nv|$ oraz $|(V(G)\backslash S)\{u,v\}| \geq |N\{u,v\}|$ dla każdego $u \in V(G)$, takiego że $\{u,v\} \in E(G)$. Musi zatem zachodzić jedna z dwóch nierówności: $|(V(G)\backslash S)v| \geq |Nv| > |(V(G)\backslash N)v| \geq |Sv|$ (sprzeczność) lub $|(V(G)\backslash S)\{u,v\}| \geq |N\{u,v\}| > |(V(G)\backslash N)\{u,v\}| \geq |S\{u,v\}|$ dla pewnego $u \in S$ (również sprzeczność).

Niech $N' = N \cup \{v\}$. Ponieważ N jest zbiorem dominującym, N' również jest zbiorem dominującym. Ponieważ N jest globalnym zbiorem defensywnym, wiadomo, że $|Nv| \geq |(V(G)\backslash N)v|$ lub istnieje $u \in N$, taki że $\{u,v\} \in E(G)$ oraz $|N\{u,v\}| \geq |(V(G)\backslash N)\{u,v\}|$. Z lematu 1 wiadomo, że dla każdego $t \in N'$ zachodzi $|N't| \geq |Nt| \geq |(V(G)\backslash N)t| \geq |(V(G)\backslash N')t|$ lub $|N'\{s,t\}| \geq |N\{s,t\}| \geq |(V(G)\backslash N)\{s,t\}| \geq |(V(G)\backslash N')\{s,t\}|$ dla pewnego $s \in N'$ takiego, że $\{s,t\} \in E(G)$, więc N' jest globalnym zbiorem defensywnym.■

Lemat 4: Niech $N \subsetneq V(G)$ będzie globalnym zbiorem defensywnym. Zostanie zdefiniowany symbol $U = \{v \in V(G)\backslash N: |Nv| > |(V(G)\backslash N)v| \wedge$ $\forall_{u \in N_G(v)\backslash N} |N\{u,v\}| > |(V(G)\backslash N)\{u,v\}|\}$. Zachodzi:

1. $N' = N \cup U$ jest globalnym zbiorem defensywnym;
2. Jeśli $U = \emptyset$, to $S = V(G)\backslash N$ jest zbiorem defensywnym;
3. Dla każdego zbioru defensywnego $S \subseteq V(G)\backslash N$ zachodzi $S \subseteq V(G)\backslash N'$.

Dowód:

(1) Ponieważ $N \cup U = \bigcup_{u \in U} N \cup \{u\}$, z lematu 2 oraz 3 wiadomo, że N' jest globalnym zbiorem defensywnym;

(2) Jeśli $U = \emptyset$, to z definicji U, dla każdego $v \in V(G)\backslash N$ zachodzi $|(V(G)\backslash N)v| \geq |Nv|$ lub $|(V(G)\backslash N)\{u,v\}| \geq |N\{u,v\}|$ dla pewnego $u \in N_G(v)\backslash N$, więc $S = V(G)\backslash N$ jest zbiorem defensywnym;

(3) Niech $u \in U \cap S$, gdzie $S \subseteq V(G)\backslash N$ jest zbiorem defensywnym. Zgodnie z lematem 3 zachodzi sprzeczność.■

Lemat 5: Niech $N \subsetneq V(G)$ będzie zbiorem defensywnym. Zostaną zdefiniowane symbole $N_0 = N,\ U_0 = \emptyset,\ U_i = \{v \in V(G)\backslash N_{i-1} : |N_{i-1}v| > |(V(G)\backslash N_{i-1})v| \wedge \forall_{u \in N_G(v)\backslash N_{i-1}} |N_{i-1}\{u,v\}| > |(V(G)\backslash N_{i-1})\{u,v\}|\}$ oraz $N_i = N_{i-1} \cup U_i$ dla $i \geq 1$. Istnieje $1 \leq k \leq n$, gdzie $n = |V(G)|$, takie, że $U_k = \emptyset$ oraz zachodzi:

1. $N_k = N_{k-1} = N_0 \cup U_1 \cup \cdots \cup U_{k-1}$ jest globalnym zbiorem defensywnym;
2. $S = V(G)\backslash N_{k-1}$ jest zbiorem defensywnym lub zbiorem pustym;
3. Dla każdego zbioru defensywnego $S^* \subseteq V(G)\backslash N$ zachodzi $S^* \subseteq S = V(G)\backslash N_{k-1}$.

Dowód: Jeśli $U_i \neq \emptyset$ dla jakiegoś $i \geq 1$, to dla każdego $j = 1, \cdots, i$ zachodzi $|N_j| = |N_{j-1}| + |U_j| > |N_{j-1}|$, zatem zachodzi $|N_i| > |N| + i$, zatem istnieje $1 \leq k \leq n$, takie że $U_k = \emptyset$. Tezy (1-3) zachodzą przez indukcję z lematu 4 (1-3).

Twierdzenie 1: Problem równowagi strategicznej dla zbiorów defensywnych jest równoważny problemowi doskonałej równowagi strategicznej dla zbiorów defensywnych.

Dowód: Niech $(\{N, S\}, I)$ będzie równowagą strategiczną dla zbiorów defensywnych w grafie G. Rozważana będzie para $\{N', S'\}$, gdzie $N' = N_{k-1}$, a $S' = V(G)\backslash N_{k-1}$. W oczywisty sposób, $N \subseteq N'$. Z lematu 5 (1) wiadomo, że N' jest globalnym zbiorem defensywnym. Z lematu 5 (3) wiadomo, że $S \subseteq S' = V(G)\backslash N_{k-1}$, a z lematu 6 (2) wiadomo, że $S' = V(G)\backslash N_{k-1}$ jest zbiorem defensywnym. Ponieważ S jest zbiorem dominującym, S' również jest zbiorem dominującym, zatem S' jest globalnym zbiorem defensywnym. Zatem para $\{N', S'\}$ jest doskonałą równowagą strategiczną dla zbiorów defensywnych.

3. Wielomianowy algorytm dla problemu *DSPSB w* drzewach

W tym rozdziale zaproponowany zostanie algorytm programowania dynamicznego dla problemu *DSPSB* działający w czasie $O(n\,\Delta \log \Delta)$.

W celu konstrukcji doskonałej równowagi strategicznej dla zbiorów defensywnych (lub podania negatywnej odpowiedzi na problem *DSPSB*), używana będzie technika *bottom-up*. Najpierw wszystkie krawędzie będą kierowane aby utworzyć drzewo *in-tree* z korzeniem, czyli wybierany będzie dowolny liść r i wszystkie krawędzie drzewa T orientowane będą w kierunku r. Zaczynając od liści, do każdego wierzchołka drzewa przypisywane będą dwa bity informacyjne oraz zerojedynkowa tablica o rozmiarze proporcjonalnym do liczby sąsiednich wierzchołków. Jeśli istnieje doskonała równowaga strategiczna, będzie ona konstruowana za pomocą przypisanych struktur zaczynając od korzenia r.

Przez T_v oznaczane będzie poddrzewo drzewa T zakorzenione w v składające sięze wszystkich zorientowanych krawędzi prowadzących do v. Dla każdego wierzchołka $v \in V(T)\backslash\{r\}$ istnieje dokładnie jedna krawędź skierowana wychodząca z v prowadząca do r, oznaczamy ją jako $e_v = \{v, r_v\}$. Dla każdego wierzchołka $v \in V(T)\backslash\{r\}$ zdefiniowany zostanie symbol T_v^* jako drzewo T_v z dodaną krawędzią e_v, czyli $T_v^* = (V(T_v) \cup \{r_v\}, E(T_v) \cup \{e_v\})$ zakorzenione w wierzchołku r_v. Przez $C(v)$ oznaczany będzie zbiór wszystkich dzieci wierzchołka v, czyli $C(r)$ to zbiór

wszystkich wierzchołków sąsiednich do r, a dla każdego $v \in V(T)\backslash\{r\}$ zbiór $C(v)$ to wszystkie wierzchołki sąsiadujące z v inne niż r_v. Zdefiniowane zostanie drzewo T_v^l uzyskane z T_v^* przez dołączenie $l \geq 0$ dodatkowych wierzchołków $L_l = \{u_1, \cdots, u_l\}$ do wierzchołka r_v. Zauważmy, że $T_v^0 = T_v^*$.

Dla drzewa T zakorzenionego w r, *doskonała A-prawie równowaga strategiczna dla zbiorów defensywnych* to podział $V(T)$ na dwa zbiory, takie że jeden z nich jest globalnym zbiorem defensywnym, a drugi jest zbiorem defensywnym oprócz wierzchołków ze zbioru A. Wierzchołki ze zbioru A nie muszą być zdominowane przez drugi zbiór defensywny, nie muszą też spełniać warunków zbioru defensywnego.

Niech $v \in V(T)\backslash\{r\}$, $p_v = \deg v - 1$,$q_v = \deg r_v - 1$. Niech L_v będzie zbiorem wszystkich wierzchołków sąsiadujących z r_v (łącznie z rodzicem wierzchołka r_v jeśli r_v nie jest korzeniem) innych niż v. Zauważmy, że $q_v = |L|$. Wprowadzone zostaną następujące oznaczenia:

1. $s_v = 1$ wtedy i tylko wtedy, gdy istnieje doskonała $\{r_v\}$-prawie równowaga strategiczna dla zbiorów defensywnych w poddrzewie T_v^*, taka że v i r_v są w tym samym zbiorze defensywnym oraz $SEC(v)$ lub $SEC(\{u, v\})$, gdzie $u \in C(v)$;
2. $d_v = 1$ wtedy i tylko wtedy, gdy istnieje doskonała równowaga strategiczna dla zbiorów defensywnych w poddrzewie T_v^*, taka że v i r_v są w różnych zbiorach defensywnych oraz $SEC(v)$ lub $SEC(\{u, v\})$, gdzie $u \in C(v)$;
3. $se_v[k] = 1$, gdzie $0 \leq k \leq q_v - 1$ wtedy i tylko wtedy, gdy istnieje doskonała $(\{r_v\} \cup L_v)$-prawie równowaga strategiczna dla zbiorów defensywnych w poddrzewie T_v^q, taka że v i r_v są w tym samym zbiorze defensywnym, $SEC(\{v, r_v\})$ oraz dokładnie k wierzchołków ze zbioru L_v są w tym samym zbiorze defensywnym co v i r_v.

Łatwo zauważyć, że:

Lemat 6: Niech $v \in V(T)\backslash\{r\}$ i $k < q_v - 1$. Jeśli $se_v[k] = 1$, to $se_v[k+1] = 1$.

Dla każdego liścia $l \in V(T)\backslash\{r\}$ zachodzi $s_l = 0$, $d_l = 1$ oraz $s[k] = 0$ dla każdego $k \in \{0, \cdots, q_l - 1\}$. Dla każdego $v \in V(T)$, $i, j, o \in \{0, 1\}$ oraz $k \in \{0, \cdots, q_v - 1\}$ zdefiniowany zostanie symbol $C_v^{ijok} = \{w \in C(v) : s_w = i \wedge d_w = j \wedge se_w[k] = o\}$.

Twierdzenie 2: Dla każdego $v \in V(T) - \{r\}$ niebędącego liściem, którego wszystkie dzieci mają przypisane wartości s, d oraz $se_v[k]$ dla każdego $k \in \{0, \ldots, p_v - 1\}$, zachodzi $s_v = 1$ wtedy i tylko wtedy, gdy istnieje $h \in \{0, \ldots, p_v - 1\}$, dla którego spełnione są wszystkie z następujących warunków:

1. $C_v^{000h} = \emptyset$;
2. $C_v^{010h} \cup C_v^{011h} \cup C_v^{110h} \cup C_v^{111h} \neq \emptyset$;
3. $\left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| - x = h$, gdzie $x \in \{0, 1\}$ oraz $x = 1$ wtedy i tylko wtedy, gdy $C_v^{010h} = \emptyset$.

oraz spełniony jest przynajmniej jeden z następujących warunków:

1. $2 + |C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}| \geq |C_v^{010h}|$;
2. $C_v^{010h} \cup C_v^{110h} \neq \emptyset$ oraz $C_v^{001h} \cup C_v^{011h} \cup C_v^{101h} \cup C_v^{111h} \neq \emptyset$;
3. $|C_v^{001h} \cup C_v^{011h} \cup C_v^{101h} \cup C_v^{111h}| \geq 2$.

Dowód: ($\Rightarrow$) Niech $\{N, S\}$ będzie doskonałą $\{r_v\}$-prawie równowagą strategiczną w poddrzewie T_v^*, a h liczbą wierzchołków $c \in C(v)$ należących do tego samego zbioru defensywnego co v. Zostanie poczynione założenie, bez straty ogólności, że $v, r_v \in N$. Wierzchołek v nie jest liściem, więc $C(v) \neq \emptyset$. Dla każdego $c \in C(v)$ rozważane będzie poddrzewo $T_c^{q_c}$. Para $\{V(T_c^{q_c}) \cap N, V(T_c^{q_c}) \cap S\}$ jest doskonałą $(\{r_c\} \cup L_c)$-prawie równowagą strategiczną w $T_c^{q_c}$. Zatem co najmniej jedna z wartości s_c, d_c, $se_c[h]$ musi mieć wartość 1, zatem $C_v^{000h} = \emptyset$. Zbiór S dominuje wszystkie wierzchołki oprócz r_v, zatem istnieje wierzchołek $c \in C(v)$, dla którego $d_c = 1$, zatem $C_v^{010h} \cup C_v^{011h} \cup C_v^{110h} \cup C_v^{111h} \neq \emptyset$.

Ponieważ h wierzchołków $c \in C(v)$ należy do N, gdzie $h \in \{0, \cdots, p_v - 1\}$, są dwa przypadki:

1. jeśli $C_v^{010h} \neq \emptyset$, to wierzchołek v jest zdominowany, zatem dokładnie h wierzchołków $c \in C(v)$ musi mieć wartość $s_c = 1$ lub $se_c[h] = 1$, zatem $\left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| = h$;
2. jeśli $C_v^{010h} = \emptyset$, to wierzchołek v musi być zdominowany przez jeden z wierzchołków $c \in C(v)$, dla których $s_c = 1$ lub $se_c[h] = 1$, więc musi być dokładnie $h + 1$ takich wierzchołków, zatem $\left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| = h + 1$.

Z definicji zbioru defensywnego, wierzchołek v musi spełniać jeden z dwóch warunków:

1. spełniony predykat $SEC(v)$ – w tym wypadku musi zachodzić $2 + |C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}| \geq |C_v^{010h}|$;
2. spełniony predykat $SEC(\{u, v\})$, gdzie $u \in C(v)$ – w tym wypadku są dwie możliwości:
   a. jeśli $C_v^{010h} \cup C_v^{110h} \neq \emptyset$, to wierzchołek v może zostać zdominowany jednym z wierzchołków $c \in C_v^{010h} \cup C_v^{110h}$, więc wystarczy że istnieje jeden wierzchołek $c \in C(v)$, dla którego $se_c[h] = 1$, zatem $C_v^{001h} \cup C_v^{011h} \cup C_v^{101h} \cup C_v^{111h} \neq \emptyset$,
   b. jeśli $C_v^{010h} \cup C_v^{110h} = \emptyset$, to trzeba zdominować wierzchołek v jednym z dzieci, dla których $se_c[h] = 1$, zatem $|C_v^{001h} \cup C_v^{011h} \cup C_v^{101h} \cup C_v^{111h}| \geq 2$.

($\Leftarrow$) Zachodzi $C_v^{000h} = \emptyset$, dla jakiegoś $h \in \{0, \cdots, p_v - 1\}$, zatem dla każdego wierzchołka $c \in C(v)$ istnieje doskonała $(\{r_c\} \cup L_c)$-prawie równowaga strategiczna w poddrzewie $T_c^{q_c}$. Doskonałą $(\{r_c\} \cup L_c)$-prawie równowaga strategiczna $\{N, S\}$ tworzona jest w następujący sposób:

1. $v, r_v \in N$;
2. $C_v^{010h} \subseteq S$;
3. $C_v^{100h} \cup C_v^{101h} \cup C_v^{001h} \subseteq N$;
4. jeśli $C_v^{010h} \neq \emptyset$, to $C_v^{011h} \cup C_v^{110h} \cup C_v^{111h} \subseteq N$;
5. jeśli $C_v^{010h} = \emptyset$, to $\left|\left(C_v^{011h} \cup C_v^{110h} \cup C_v^{111h}\right) \cap S\right| = 1$ oraz $\left|\left(C_v^{011h} \cup Cv110h \cup Cv111h \cap N = Cv011h \cup Cv110h \cup Cv111h - 1\right.\right.$;
6. dla każdego wierzchołka $c \in C(v) \cap N$, który nie jest liściem powtarzane są powyższe kroki;

7. dla każdego wierzchołka $c \in C(v) \cap S$, który nie jest liściem wykonywane są kroki z dowodu twierdzenia 3 z odwróconymi zbiorami N oraz S.

Twierdzenie 3: Dla każdego $v \in V(T)\backslash\{r\}$ niebędącego liściem, którego wszystkie dzieci mają przypisane wartości s, d oraz $se[k]$ dla wszystkich $k \in \{0, \cdots, p_v - 1\}$ zachodzi $d_v = 1$ wtedy i tylko wtedy, gdy istnieje $h \in \{1, \cdots, p_v\}$, dla którego spełnione są wszystkie z następujących warunków:

1. $C_v^{000(h-1)} = \emptyset$;
2. $\left|C_v^{001(h-1)} \cup C_v^{011(h-1)} \cup C_v^{100(h-1)} \cup C_v^{101(h-1)} \cup C_v^{110(h-1)} \cup C_v^{111(h-1)}\right| = h$.

oraz spełniony jest jeden z następujących warunków:

1. $\left|C_v^{001(h-1)} \cup C_v^{011(h-1)} \cup C_v^{100(h-1)} \cup C_v^{101(h-1)} \cup C_v^{110(h-1)} \cup C_v^{111(h-1)}\right| \geq C_v^{010(h-1)}$;
2. $C_v^{001(h-1)} \cup C_v^{011(h-1)} \cup C_v^{101(h-1)} \cup C_v^{111(h-1)} \neq \emptyset$.

Dowód: ($\Rightarrow$) Niech $\{N, S\}$ będzie doskonałą równowagą strategiczną w poddrzewie T_v^*, a h liczbą wierzchołków $c \in C(v)$ należących do tego samego zbioru defensywnego co v. Zostanie poczynione założenie, bez straty ogólności, że $v \in N$ oraz $r_v \in S$. Wierzchołek v nie jest liściem, więc $C(v) \neq \emptyset$. Dla każdego $c \in C(v)$ rozważane będzie poddrzewo $T_c^{q_c}$. Para $\{V(T_c^{q_c}) \cap N, V(T_c^{q_c}) \cap S\}$ jest dosko-nałą $(\{r_c\} \cup L_c)$ – prawie równowagą strategiczną w $T_c^{q_c}$. Zatem co najmniej jedna z wartości s_c, d_c, $se_c[h-1]$ musi mieć wartość 1, zatem $C_v^{000(h-1)} = \emptyset$.

Ponieważ h wierzchołków $c \in C(v)$ należy do N, gdzie $h \in \{1, \cdots, p_v\}$, dokładnie h wierzchołków $c \in C(v)$ musi mieć wartość $s_c = 1$ lub $se_c[h-1] = 1$, zatem $\left|C_v^{001(h-1)} \cup C_v^{011(h-1)} \cup C_v^{100(h-1)} \cup C_v^{101(h-1)} \cup C_v^{110(h-1)} \cup C_v^{111(h-1)}\right| = h$.

Z definicji zbioru defensywnego, wierzchołek v musi spełniać jeden z dwóch warunków:

1. spełniony predykat $SEC(v)$: w tym wypadku musi zachodzić $\left|C_v^{001(h-1)} \cup Cv011(h-1) \cup Cv100(h-1) \cup Cv101(h-1) \cup Cv110(h-1) \cup Cv111(h-1) \geq Cv010(h-1)\right|$;
2. spełniony predykat $SEC(\{u, v\})$, gdzie $u \in C(v)$: w tym wypadku musi istnieć $u \in C(v)$, dla którego $se_u[h-1] = 1$, a więc $C_v^{001(h-1)} \cup C_v^{011(h-1)} \cup C_v^{101(h-1)} \cup C_v^{111(h-1)} \neq \emptyset$.

($\Leftarrow$) Zachodzi $C_v^{000h} = \emptyset$, dla jakiegoś $h \in \{1, \cdots, p_v\}$, zatem dla każdego wierzchołka $c \in C(v)$ istnieje doskonała $(\{r_c\} \cup L_c)$-prawie równowaga strategiczna w poddrzewie $T_c^{q_c}$. Doskonała $(\{r_c\} \cup L_c)$-prawie równowaga strategiczna $\{N, S\}$ tworzona jest w następujący sposób:

1. $v \in N, r_v \in S$;
2. $C_v^{010(h-1)} \subseteq S$;
3. $C_v^{001(h-1)} \cup C_v^{011(h-1)} \cup C_v^{100(h-1)} \cup C_v^{101(h-1)} \cup C_v^{110(h-1)} \cup C_v^{111(h-1)} \subseteq N$;

4. dla każdego wierzchołka $c \in C(v) \cap S$ powtarzane są powyższe kroki z odwróconymi zbiorami N oraz S;
5. dla każdego wierzchołka $c \in C(v) \cap N$ wykonywane są kroki z dowodu twierdzenia 2.

Twierdzenie 4: Dla każdego $v \in V(T)\backslash\{r\}$ niebędącego liściem, którego wszystkie dzieci mają przypisane wartości s, d oraz $se[k]$ dla wszystkich $k \in \{0, \cdots, p_v - 1\}$ zachodzi $se_v[k] = 1$ wtedy i tylko wtedy, gdy istnieje $h \in \{0, \cdots, p_v - 1\}$, dla którego spełnione są wszystkie z następujących warunków:

1. $C_v^{000h} = \emptyset$;
2. $C_v^{010h} \cup C_v^{011h} \cup C_v^{110h} \cup C_v^{111h} \neq \emptyset$;
3. $\left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| - x = h$, gdzie $x \in \{0, 1\}$ oraz $x = 1$ wtedy i tylko wtedy, gdy $C_v^{010h} = \emptyset$;
4. $2 + 2k - q + \left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| - x \geq \max\{1, \left|C_v^{010h}\right|\}$, gdzie $x \in \{0, 1\}$ oraz $x = 1$ wtedy i tylko wtedy, gdy $C_v^{010h} = \emptyset$; gdzie $k \in \{0, \cdots, q_v - 1\}$.

Dowód: ($\Rightarrow$) Niech $\{N, S\}$ będzie doskonałą $(\{r_v\} \cup L_v)$-prawie równowagą strategiczną w poddrzewie T_v^q, a h liczbą wierzchołków $c \in C(v)$ należących do tego samego zbioru defensywnego co v. Zostanie poczynione założenie, bez straty ogólności, że $v, r_v \in N$. Wierzchołek v nie jest liściem, więc $C(v) \neq \emptyset$. Dla każdego $c \in C(v)$ rozważane będzie poddrzewo $T_c^{q_c}$. Para $\{V(T_c^{q_c}) \cap N, V(T_c^{q_c}) \cap S\}$ jest doskonałą $(\{r_c\} \cup L_c)$-prawie równowagą strategiczną w $T_c^{q_c}$. Zatem co najmniej jedna z wartości s_c, d_c, $se_c[h]$ musi mieć wartość 1, zatem $C_v^{000h} = \emptyset$. Zbiór S dominuje wszystkie wierzchołki oprócz tych ze zbioru $\{r_v\} \cup L_v$, zatem istnieje wierzchołek $c \in C(v)$, dla którego $d_c = 1$, zatem $C_v^{010h} \cup C_v^{011h} \cup C_v^{110h} \cup C_v^{111h} \neq \emptyset$.

Ponieważ h wierzchołków $c \in C(v)$ należy do N, gdzie $h \in \{0, \cdots, p_v - 1\}$, są dwa przypadki:

1. jeśli $C_v^{010h} \neq \emptyset$, to wierzchołek v jest zdominowany, zatem dokładnie h wierzchołków $c \in C(v)$ musi mieć wartość $s_c = 1$ lub $se_c[h] = 1$, zatem $\left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| = h$;
2. jeśli $C_v^{010h} = \emptyset$, to wierzchołek v musi być zdominowany przez jeden z wierzchołków $c \in C(v)$, dla których $s_c = 1$ lub $se_c[h] = 1$, więc musi być dokładnie $h + 1$ takich wierzchołków, zatem $\left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| = h + 1$;
3. Wiadomo, że krawędź $\{v, r_v\}$ ma zapewnione bezpieczeństwo. Rozpatrywane będą dwie możliwości:
   a. $C_v^{010h} \neq \emptyset$: w tym przypadku wierzchołek v jest zdominowany, więc musi być jedynie spełniony warunek bezpieczeństwa dla krawędzi $\{v, r_v\}$: $2 + 2k - q + \left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| \geq \left|C_v^{010h}\right|$,
   b. $C_v^{010h} = \emptyset$: w tym przypadku wierzchołek v musi być zdominowany przez jeden wierzchołek $c \in C(v)$, dla którego $s_c = 1$ lub $se_c[h] = 1$, zatem $2 + 2k - q + \left|C_v^{001h} \cup C_v^{011h} \cup C_v^{100h} \cup C_v^{101h} \cup C_v^{110h} \cup C_v^{111h}\right| \geq 2$.

($\Leftarrow$) Zachodzi $C_v^{000h} = \emptyset$ dla jakiegoś $h \in \{0, \cdots, p_v - 1\}$, zatem dla każdego wierzchołka $c \in C(v)$ istnieje doskonała $(\{r_c\} \cup L_c)$-prawie równowaga strategiczna w poddrzewie $T_c^{q_c}$. Doskonała $(\{r_c\} \cup L_c)$-prawie równowaga strategiczna $\{N, S\}$ jest tworzona w następujący sposób:

1. $v, r_v \in N$;
2. $C_v^{010h} \subseteq S$;
3. $C_v^{100h} \cup C_v^{101h} \cup C_v^{001h} \subseteq N$;
4. jeśli $C_v^{010h} \neq \emptyset$, to $C_v^{011h} \cup C_v^{110h} \cup C_v^{111h} \subseteq N$;
5. jeśli $C_v^{010h} = \emptyset$, to $\left|\left(C_v^{011h} \cup C_v^{110h} \cup C_v^{111h}\right) \cap S\right| = 1$ oraz $\left|\left(C_v^{011h} \cup C_v^{110h} \cup C_v^{111h}\right) \cap N\right| = \left|C_v^{011h} \cup C_v^{110h} \cup C_v^{111h}\right| - 1$;
6. dla każdego wierzchołka $c \in C(v) \cap N$ powtarzane są powyższe kroki;
7. dla każdego wierzchołka $c \in C(v) \cap S$ wykonywane są kroki z dowodu twierdzenia 3 z odwróconymi zbiorami N oraz S.

Twierdzenie 5: Doskonała równowaga strategiczna dla zbiorów defensywnych w drzewie T zakorzenionym w r, gdzie s to jedyny wierzchołek sąsiadujący z r, istnieje wtedy i tylko wtedy, gdy $s \in C_v^{0100} \cup C_v^{0110} \cup C_v^{1100} \cup C_v^{1110}$.

Dowód: ($\Rightarrow$) Niech $\{N, S\}$ będzie doskonałą równowagą strategiczną w drzewie T z korzeniem w r, a s jedynym wierzchołkiem sąsiadującym z r. Zostanie poczynione założenie, bez straty ogólności, że $r \in N$. Zbiór S dominuje wierzchołek r, zatem $s \in S$, zatem $d_s = 1$, zatem $s \in C_v^{0100} \cup C_v^{0110} \cup C_v^{1100} \cup C_v^{1110}$.

($\Leftarrow$) Zachodzi $s \in C_v^{0100} \cup C_v^{0110} \cup C_v^{1100} \cup C_v^{1110}$, zatem istnieje doskonała równowaga strategiczna w $T = T_s^{q_s}$. Doskonała równowaga strategiczna $\{N, S\}$ jest tworzona w następujący sposób:

1. $r \in N$;
2. $s \in S$;
3. dla wierzchołka s wykonywane są kroki z dowodu twierdzenia 3 z odwróconymi zbiorami N oraz S.

Tworząc tablicę s, trzeba obliczyć liczności zbiorów C_v^{ijoh}, gdzie $i, j, o \in \{0, 1\}$ dla każdego $h \in \{0, \cdots, p_v - 1\}$, co można zrobić w czasie $O((\deg v)^2)$ oraz wyznaczyć wartość $se_v[k]$ dla każdego $k \in \{0, \cdots, p_v - 1\}$ na podstawie liczności zbiorów C_v^{ijoh}, co również można zrobić w czasie $O((\deg v)^2)$. Korzystając z lematu 6, można ograni-czyć czas potrzebny do wyznaczania liczności zbiorów C_v^{ijoh} do $O(\deg v \log \deg v)$ oraz czas potrzebny do wyznaczenia wszystkich wartości $se_v[k]$ do $O(\log^2 \deg v)$ poprzez zastosowanie przeszukiwania binarnego. Czynnikiem determinującym złożoność jest więc obliczenie liczności zbiorów C_v^{ijoh} wykonywane w czasie $O(\deg v \log \deg v)$, zatem złożoność obliczeniowa całego algorytmu wynosi $O(n \Delta \log \Delta)$.

4. Podsumowanie i dalsze kierunki prac

W niniejszej pracy wprowadzone zostało pojęcie równowagi strategicznej dla zbiorów defensywnych, udowodniona została równoważność problemów istnienia równowagi strategicznej dla zbiorów defensywnych oraz istnienia doskonałej równowagi

strategicznej dla zbiorów defensywnych oraz zaproponowany został wielomianowy algorytm dla problemu doskonałej równowagi strategicznej dla zbiorów defensywnych w drzewach.

Problemem otwartym zostaje generowanie wszystkich możliwych doskonałych równowag strategicznych w danym drzewie. Pozwoliłoby to scharakteryzować cechy podziałów będących doskonałymi równowagami strategicznymi, takie jak proporcje między licznością obu zbiorów.

Literatura

1. Peleg D., *Local majorities coalitions and monopolies in graphs: A review*, Theoretical Computer Science, 282 (2002), s. 213-257
2. Flake G. W., Lawrence S., Giles C. L., *Efficient identification of web communities*, Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2000), s. 150-160
3. Lewoń R., Małafiejska A., Małafiejski M., *Strategic balance in graphs*, Discrete Mathematics, 339 (2014), s. 1837-1847
4. Haynes T. W., Hedetniemi S. T., Henning M. A., *Global defensive alliances*, Proc. of. the 17th Int. Symposium on Computer Information Science, (2002), s. 303-307
5. Hedetniemi S. M., Hedetniemi S. T., Kristiansen P., *Introduction to alliances in graphs, Proc. of. the 17th Int.* Symposium on Computer Information Science, (2002), s. 308-312
6. Lewoń R., Małafiejska A., Małafiejski M., *Global defensive sets in graphs*, Discrete Mathematics, 339 (2014), s. 1861-1870

Równowaga strategiczna dla zbiorów defensywnych w drzewach

W pracy rozważany jest problem defensywnej równowagi strategicznej dla zbiorów defensywnych w drzewach (spójnych grafach acyklicznych), który polega na znalezieniu
dwóch rozłącznych globalnych zbiorów defensywnych. Zagadnienie to znajduje zastosowanie w modelowaniu problemów komunikacyjnych w sieciach.

Dla danego grafu G podzbiór jego wierzchołków S jest zbiorem defensywnym, jeśli dla każdego wierzchołka v należącego do S spełniony jest warunek koalicyjny, tzn. w jego domkniętym sąsiedztwie jest co najmniej tyle samo wierzchołków z S co spoza S lub istnieje wierzchołek u sąsiedni do v taki, że w domkniętym sąsiedztwie krawędzi $\{v, u\}$ jest co najmniej tyle samo wierzchołków z S co spoza S.

W pracy pokazana została równoważność między problemami równowagi strategicznej oraz doskonałej równowagi strategicznej dla zbiorów defensywnych w drzewach oraz przedstawiony został wielomianowy algorytm rozwiązujący problem równowagi strategicznej dla zbiorów defensywnych w drzewach.

Słowa kluczowe: doskonała równowaga strategiczna, zbiory defensywne

Strategic balance for defensive sets in trees

In the paper we study the strategic balance problem for defensive sets in trees (connected acyclic graphs), which is determining if there exist two disjoint global defensive sets for a given tree. This problem is applicable in network communication modeling.

For a given graph G, a subset S contained in $V(G)$ is a defensive set if and only if for each vertex $v \in S$ there are at least as many vertices from the closed neighbourhood of v in S as in $V(G)\backslash S$ or there exists a neighbour u of v such that $u \in S$ and there are at least as many vertices from the closed neighbourhood of $\{u, v\}$ in S as in $V(G)\backslash S$.

In the paper we proved equivalence of strategic balance and perfect strategic balance problems for defensive sets and we proposed a polynomial time algorithm solving the perfect strategic balance problem for defensive sets in trees.

Keywords: perfect strategic balance, defensive sets

Żaklin Maria Grądz[1]

Wybrane metody obrazowania procesu spalania

1. Wprowadzenie

Zanieczyszczenie środowiska naturalnego spowodowane jest w głównej mierze spalaniem paliw kopalnych. W celu ochrony środowiska wprowadzane są rożnego rodzaju rozwiązania prawne, które obostrzają dopuszczalne normy określające ilość emitowanych do atmosfery gazów spalinowych i pyłów. W Polsce energetyka zawodowa opiera się na przetwarzaniu paliw kopalnych. Ich spalanie związane jest z emisją do atmosfery związków chemicznych tj. CO2, SO2, NOx oraz pyłów [1]. Sytuacja ta wymaga wprowadzania nowych technik analizy procesu spalania, do których można zaliczyć spalanie niskoemisyjne. Niskoemisyjne techniki spalania pozwalają na spełnienie wymagań stawianych przez Ochronę Środowiska, a tym samym dają możliwość ograniczenia kosztów [2-3].

Proces spalania powinien zachodzić w optymalnych warunkach, które pozwolą zachować jego efektywność oraz ograniczą ilość emitowanych do atmosfery gazów spalinowych i pyłów. Wpływ na efektywność procesu spalania ma wiele czynników. W celu ich poznania należy dobrać odpowiednią metodę diagnostyczną. Szczególną rolę odgrywają tu metody optyczne zaliczane do grupy metod nieinwazyjnych. W układach monitorujących wykorzystujących czujniki światłowodowe pozyskiwanie i przetwarzanie danych odbywa się w czasie rzeczywistym. W przemyśle występują kotły o różnych konstrukcjach, a co się z tym wiąże z różnymi układami palników [4-6]. Układy monitorujące oparte na czujnikach optycznych powinny obejmować każdy z palników w komorze spalania. Celem takiego zastosowania czujników jest uzyskanie jak najlepszych parametrów diagnostycznych. Ze względu na trudne warunki panujące w komorze paleniskowej, zazwyczaj w optycznych układach monitorujących wykorzystuje się światłowody ze specjalnym pokryciem. Przykładem wykorzystywanym do nieinwazyjnego pozyskiwania danych o procesie spalania jest światłowodowy układ do pomiaru zmian intensywności świecenia płonienia powstały w Instytucie Elektroniki i Technik Informacyjnych (była Katedra Elektroniki Politechniki Lubelskiej) [7-9].

Wykorzystanie metod optycznych w celach diagnostyki procesu spalania opiera się na analizie informacji zawartych w widmie płomienia oraz intensywności jego świecenia. W przypadku analizy płomienia w zakresie widzialnym stosuje się kamery, które z wykorzystaniem przetworników krzemowych przetwarzają obraz płomienia. Prowadzenie badań nad obrazem płomienia opiera się na wykorzystaniu metod opisu kształtu obiektu [10-11]. Analiza kształtu płomienia może być stosowana w zakresie jego obszaru oraz konturu [12].

[1] zaklin.gradz@pollub.edu.pl, Instytut Elektroniki i Technik Informacyjnych Wydział Elektrotechniki i Informatyki Politechnika Lubelska, www.pollub.pl

Metody opisu kształtu obiektu wykorzystujące kontur jaki i obszar można podzielić na [13]:

- Metody wykorzystujące konturu:
  - Strukturalne (np. kod łańcuchowy, wielobok, krzywa B-sklejana),
  - Globalne (np. kolistość, sygnatura kształtu, odległość Hausdoffa, deskryptory Fouriera, deskryptory Falkowe),
- Metody wykorzystujące obszar:
  - Strukturalne (np. szkielet, powłoka wypukła),
  - Globalne (np. Powierzchnia, liczba Eulera, niewspółśrodkowość, momenty Legendre'a, uogólnione deskryptory Fouriera) [13].

W niniejszy artykule zostaną opisane wybrane metody obrazowania procesu spalania [14-15]. Pierwszą z nich będzie analiza składowych głównych (ang. PCA – principal component analysis), która w zastosowaniu do analizy procesu spalania pozwala na przetworzenie obrazu płomienia w celu zredukowania ilości zmiennych z zachowaniem tych najbardziej reprezentowalnych. Kolejną metodą prezentowaną w niniejszej publikacji jest zastosowanie deskryptorów Fouriera (FD – ang. Fourier Descriptors) w diagnostyce procesu spalania. Metoda ta opiera się na analizie kształtu obiektu płomienia z wykorzystaniem transformacji Fourierowskich. Ostatnią metodą prezentowaną w tym rozdziale będzie wykorzystanie transformacja Falkowa (ang. Wavelet transformation) w analizie obrazu płomienia. Dzięki jej zastosowaniu możliwe jest przetworzenie poszczególnych cech obrazu płomienia w wielowymiarowy sposób.

2. Analiza składowych głównych (ang. PCA – principal component analysis)

Wykorzystanie analizy składowych głównych do monitorowania procesu spalania pozwala na zredukowanie ilości zmiennych, które są pozyskiwane z układów pomiarowych płomienia. PCA należy do grupy metod statystycznych, które umożliwią zmniejszenie liczby wymiarów danych z zachowaniem najbardziej reprezentowalnych zmiennych [16]. Na podstawie zmiennych pierwotnych, które są przekształcone linowo przy zachowaniu jak największej zmienności, wyznacza się zmienne wtórne (składowe główne) [12]. Wyznaczenie pierwszej składowej głównej odbywa się z wykorzystaniem następującej zależności [17]:

$$var\left(\gamma_{(1)}^{T}\boldsymbol{x}\right) = max\{var(\boldsymbol{a}^{T}\boldsymbol{x})\} \qquad (1)$$

gdzie:
$\gamma_{(1)}$ - zmienna losowa ($\gamma_{(1)} \in \mathbf{R}$),
$\boldsymbol{x}$ - wektor losowy należący do zbioru danych $\mathbf{R}$,
$\boldsymbol{a}$ - wektor jednostkowy ($\|\boldsymbol{a}\|^2 \equiv \boldsymbol{a}^T\boldsymbol{a} = \mathbf{1}$),
$\boldsymbol{a}^T\boldsymbol{x}$ - standardowa kombinacja liniowa.

W związku z powyższym poszukiwana jest standaryzowana kombinacja liniowa charakteryzująca się możliwe największą wariancją [12,17]. Wyznaczenie kolejnej – drugiej składowej głównej w technice PCA odbywa się za pomocą [17]:

$$var\left(\gamma_{(2)}^{T}\boldsymbol{x}\right) = max\{var(a^{T}x)\} \qquad (2)$$

przy zachowaniu warunku [17]:

$$E\left[\gamma_{(1)}^{T}(\boldsymbol{x}-\boldsymbol{m})\gamma_{(2)}^{T}(\boldsymbol{x}-\boldsymbol{m})\right] = \mathbf{0}. \qquad (3)$$

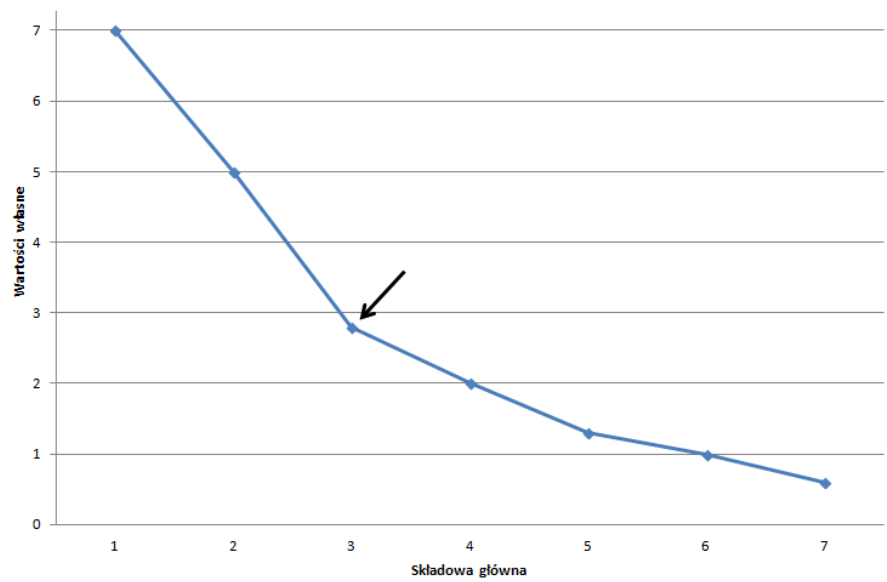
Druga składowa główna będzie miała możliwe największą wariancję, tylko wtedy gdy nie będzie skorelowana z pierwszą Zasada ta działa tak samo w przypadku każdej k + 1-szej składowej głównej [17]. Liczba składowych głównych jest zależna od konkretnego przypadku oraz może zostać wyznaczona za pomocą wykorzystania kryterium Kaisera lub wykresu osypiska Cattella [12].

Kryterium Kaisera

Kryterium Kaisera według definicji przedstawionej w [18] to: „Mając miarę ilości wariancji, którą wyodrębnia każdy kolejny czynnik, w postaci jego wartości własnej możemy zostawić tylko czynniki, które mają wartości własne większe niż 1, czyli gdy wyodrębniają przynajmniej tyle zmienności, ile jedna zmienna oryginalna". Kryterium to najlepiej sprawdza się gdy liczba zmiennych przekracza 20. W przypadku gdy liczba zmiennych jest mniejsza, może dochodzić do wyodrębniania za małej liczby czynników [18].

Test osypiska

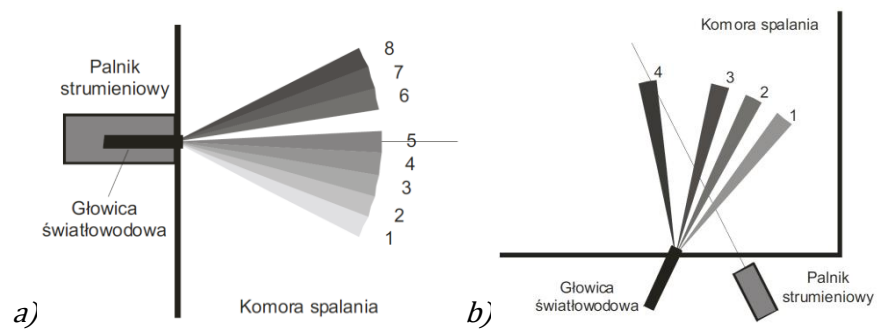
Wśród metod wyodrębniania składowych głównych należy wyróżnić wykres osypiska Cattella. W tej metodzie wartości własne są uporządkowane w sposób malejący oraz przedstawione na wykresie linowym. Na poniższym rysunku znajduje się przykładowy wykres osypiska.



Rys.1. Wykres osypiska z zaznaczonym miejscem załamania krzywej

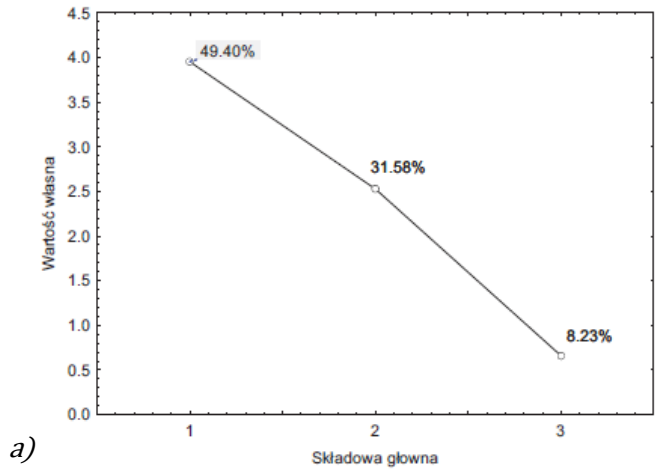
W teście osypiska poszukuje się miejsca załamania krzywej, od którego spadek wartości własnych staje się łagodny [18]. Na rysunku 1. miejsce załamania oznaczone jest czarną strzałką. Z powyższego wykresu wynika, że do analiz należy przyjąć trzy pierwsze wartości ze składowych głównych.

Obecnie w literaturze występuje wiele artykułów naukowych, które przedstawiają problematykę wykorzystania analizy składowych głównych do analizy danych procesowych. Przykładem tego jest publikacja [19], w której zastosowano PCA do redukcji ilości zmiennych wyjściowych pochodzących z układu monitorowania płomienia. W [19] autorzy przeprowadzili badania procesu spalania zachodzącego w kotle energetycznym w warunkach przemysłowych skupiając się na obserwacji wybranych obszarów płomienia. Na rysunku 2 przedstawione są dwa położenia pomiarowej głowicy światłowodowej wraz z oznaczonymi badanymi obszarami płomienia.

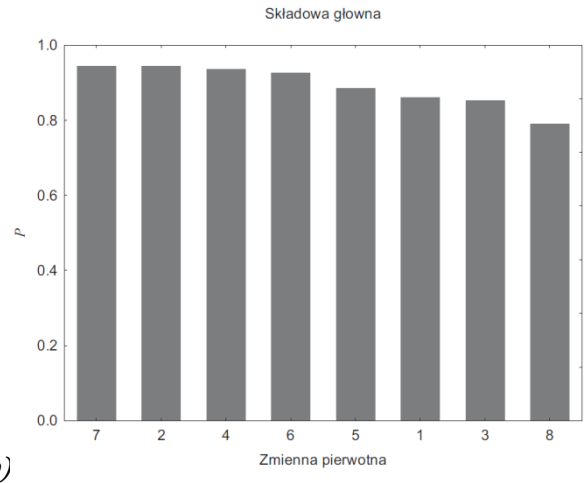


Rys.2. Położenie głowicy pomiarowej w komorze spalania a)poziome, b)pionowe wraz z oznaczonymi badanymi obszarami płomienia [19]

W publikacji [19] została zaprezentowana analiza głównych składowych w zastosowaniu do opracowania danych pomiarowych płomienia pozyskanych z dwóch położeń głowicy (poziomego i pionowego) w komorze spalania. Do analizy przyjęto 8 sygnałów z położenia poziomego i 4 sygnały z położenia pionowego. Po zastosowaniu kryterium Kaisera, dla danych pozyskanych z położenia poziomego głowicy, liczba wymiarów zmiennych może zostać zredukowana z 8 do 2 składowych głównych. Na poniższym rysunku znajduje się wykres osypiska dla trzech pierwszych składowych głównych oraz charakterystyka wartości współczynnika *P* dla kolejnych zmiennych w modelu PCA [19].



a)



b)

Rys. 3. a) Wykres osypiska dla trzech pierwszych składowych głównych układu z głowicą umiejscowioną w sposób poziomy b) charakterystyka wartości współczynnika *P* dla kolejnych zmiennych w modelu PCA [19]

Jak przedstawia rysunek 3, test osypiska został przeprowadzony dla wariancji pierwszej składowej głównej o wartości 49,40%, drugiej – 31,58% oraz trzeciej – 8,23% [19].

3. Deskryptory Fouriera (FD – ang. Fourier Descriptors)

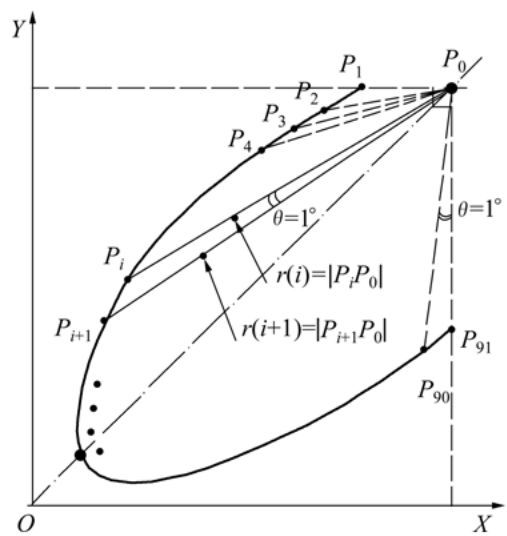
Wykorzystanie deskryptorów Fouriera w diagnostyce procesu spalania opiera się na analizie kształtu obiektu jakim jest płomień. W obrazie płomienia jego kontur jest opisywany przez piksele, które tworzą tzw. krawędź [12,20].

Opis krawędzi z wykorzystaniem deskryptorów Fouriera oparty jest na dobrze znanej teorii Fouriera. Metoda ta ma wiele zastosowań i jest jednym z najpowszechniej stosowanych sposobów opisu kształtu badanego obiektu. Głównym celem FD jest

wykorzystanie granicy przekształcenia Fouriera w opisie konturu [13,21]. Deskryptory Fouriera w porównaniu z innymi deskryptorami kształtu (np. Falkowymi) mają wiele zalet między innymi charakteryzują się [13]:

- prostymi obliczeniami,
- posiadaniem określonego znaczenia fizycznego,
- możliwością dopasowania kształtu do prostego zadania,
- posiadaniem cech globalnych, jaki i również lokalnych.

W większości publikacji opisywane jest zastosowanie FD do rozpoznawania zbiorów znaków i klasyfikacji obiektów. Przykładem artykułu, w którym przedstawiono deskryptor kształtu wykorzystywany do opisu płomienia jest [22].



Rys. 4. Kontur płomienia z zaznaczonymi pikselami [22]

Na rysunku 2 został przedstawiony przykładowy kontur obrazu płomienia, na który składa się zbiór pikseli *{(x(i), y(i)), i=0, 1, 2,..., N−1}*. Przy czym, *N* określa liczbę pikseli w analizowanym konturze. Powyższy kontur składa się z 91 oznaczonych pikseli *{P1, P2, ...,P91}*. W zależności od badanego przypadku kształtu płomienia, liczba pikseli N będzie zmienna. Jak przedstawiają autorzy w publikacji [22] deskryptor kształtu wyznaczany jest na podstawie następujących zależności [22]:

$$F(u) = \left(\frac{1}{N}\right) \sum_{t=0}^{N-1} r(t) \left(cos\left(\frac{2\pi ut}{N}\right) - i\, sin\left(\frac{2\pi ut}{N}\right)\right), \tag{4}$$

$$|F(u)| = \left(\frac{1}{N}\right) \sqrt{\left(\sum_{t=0}^{N-1} r(t)\, cos\left(\frac{2\pi ut}{N}\right)\right)^2 + \left(\sum_{t=0}^{N-1} r(t)\, sin\left(\frac{2\pi ut}{N}\right)\right)^2} \tag{5}$$

przy czym funkcję odległości *r(t)* wyraża się następująco [22]:

$$r(t) = \frac{1}{L_{P_o O}} \sqrt{\left(x_{P_t} - x_{P_0}\right)^2 + \left(y_{P_t} - y_{P_0}\right)^2}, \qquad (6)$$

gdzie:

$L_{P_o O}$ – odległość przekątnej P_0O,

x_{P_t}, y_{P_t} – współrzędne punktu P_t,

x_{P_0}, y_{P_0} – współrzędne punktu odniesienia.

W celu wyznaczenia deskryptora kształtu płomienia $\hat{F}(u)$ wybiera się pierwszych 40 współczynników $|F(u)|$. Ograniczenie ich liczby pozwala na zmniejszenie wymiarowości deskryptora konturu płomienia. Współczynnik $|F(0)|$ jest maksymalnym współczynnikiem wykorzystanym do przeprowadzenia analizy, a jednocześnie czynnikiem normalizującym w równaniu [22].

Deskryptor kształtu płomienia $\hat{F}(u)$ wyraża się następująco [22]:

$$\hat{F}(u) = \frac{|F(u)|}{|F(0)|}. \qquad (7)$$

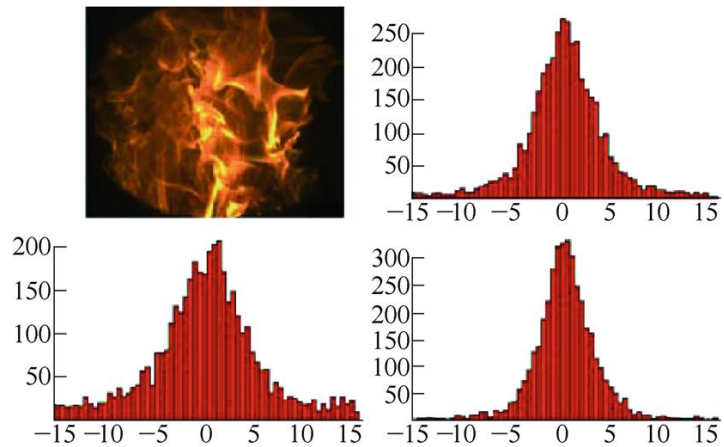
4. Transformacja Falkowa (ang. Wavelet transformation)

Metody transformacji Falkowej charakteryzują się dużą popularnością i dynamicznym rozwojem w zakresie narzędzi analizy czasowo-częstotliwościowej sygnałów niestacjonarnych. Transformacja Falkowa posiada liczne zastosowania w [23]:

- odszumianiu w zakresie współczynników transformacji [24],
- kompresji oraz kodowaniu sygnałów,
- czasowo-częstotliwościowej identyfikacji stanu obiektów, która jest oparta na analizie wygenerowanych sygnałów,
- detekcji punktów skokowej zmiany charakteru sygnału [23].

Metody analizy czasowo-częstotliwościowej sygnałów obejmują swoim zakresem pozyskiwanie informacji, które są zawarte w widmie częstotliwościowym analizowanego sygnału. Pozyskiwaną informacją może być ogólny kształt widma sygnału, jaki i również chwilowy rozkład częstotliwości, faz oraz amplitud sygnału. W przypadku gdy, analizie poddawane są sygnały przestrzenne otrzymywana jest informacja o zmianach w częstotliwości występowania konkretnych cech w przestrzeni. Falkowa dekompozycja obrazów może być stosowana do odszumiania, analizy oraz ich kompresji w dziedzinie współczynników transformaty. Dodatkowo, może zostać wykorzystana do rozpoznawania tekstur [23].

W publikacji [16] została poruszona problematyka zastosowania transformacji Falkowej w analizie obrazu płomienia. Jest to wielowymiarowy sposób reprezentacji poszczególnych cech tego obrazu. W diagnostyce płomienia Falki znajdują wiele zastosowań np. w kompresji obrazów oraz usuwaniu szumów. Dyskretna transformacja Falkowa 2D może być wykorzystywana do analizy obrazów 2D w skali szarości [16]. W przypadku gdy obraz płomienia jest traktowany jako sygnał wejściowy to z transformacji Falkowej można uzyskać 4 obrazy w zmniejszonej skali [16].



Rys.5. Obraz płomienia i wyniki jego analizy przeprowadzonej z wykorzystaniem transformacji Falkowej [16]

Na rysunku 5 przedstawione są wyniki analizy obrazu płomienia opracowane z wykorzystaniem transformacji Falkowej. Jak widać, każdy histogram jest podobny do rozkładu Gaussa o środku zero [16].

5. Wnioski

W niniejszym rozdziale zaprezentowane zostały wybrane metody analizy obrazów płomienia pochodzących z procesów spalania. Obecnie w literaturze występuje wiele artykułów naukowych, które przedstawiają problematykę publikacji. Diagnozowanie procesu spalania ma na celu zachowanie jego efektywność oraz ograniczenie ilość emitowanych do atmosfery gazów spalinowych i pyłów. Wpływ na poprawność tego procesu ma wiele czynników, a w szczególności ważny jest dobór odpowiedniej metody diagnostycznej. Badania na obrazie płomienia mogą być prowadzone w zakresie jego konturu bądź kształtu. Do analiz obraz płomienia z badanego układu jest pozyskiwany za pomocą kamer z przetwornikami krzemowymi, a następnie poddawany jest obróbce matematycznej w celu wyodrębnienia najbardziej charakterystycznych cech.

Przetwarzanie obrazów płomienia za pomocą analizy składowych głównych (ang. PCA – principal component analysis) daje możliwość redukcji liczby zmiennych, które są pozyskiwane z układu diagnostycznego płomienia oraz wyznaczenia charakterystycznych stref w jego obrazie.

Stosowanie deskryptorów Fouriera (FD – ang. Fourier Descriptors) w diagnostyce płomienia w głównej mierze opiera się na opisie jego krawędzi. W metodzie tej wykorzystane są transformacje Fourierowskie w zakresie analizy kształtu obiektu.

Wykorzystywanie transformacji Falkowej (ang. Wavelet transformation) do opisu obrazu płomienia sprowadza się do wielowymiarowego sposób reprezentacji poszczególnych, charakterystycznych cech tego obrazu. W analizie obrazu płomienia, transformacje Falkowe mogą również posłużyć do jego odszumiania oraz kompresji.

Dla poprawnej diagnostyki procesu spalania niezmiernie ważne jest stosowanie poprawnie dobranych metod badawczych oraz prowadzenie szczegółowych analiz jego przebiegu. Transformacje Falkowe, deskryptory Fouriera oraz PCA są bardzo popularnymi i dynamicznie rozwijającymi się metodami analizy sygnałów niestacjonarnych.

Literatura

1. Sawicki D., *Combustion process state classification based on flame image analysis*, Informatyka Automatyka Pomiary w Gospodarce i Ochronie Środowiska 4, s. 77-80, 2016.
2. Tanaś J., Kotyra A., *Application of optical flow algorithms and flame image sequences analysis in combustion process diagnostics*, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016 : conference [WOS], 2016.
3. Kordylewski, W. and Hardy T., *Niskoemisyjne techniki spalania: problem i peryspektywy*, Politechnika Wrocławska, Wrocław 2003.
4. Jones J. C., *A combustion scientist's view of thermocouple temperature measurement*, IEE Seminar Advanced Sensors and Instrumentation Systems for Combustion Processes, s. 11/1-11/4, 2000.
5. Guinee M. J., *Measurement of emissions from offshore combustion-user's requirements*, IEE Seminar Advanced Sensors and Instrumentation Systems for Combustion Processes, s. 7/1-7/3, 2000.
6. Ozanyan K. B. et al., *All-optoelectronic sensors for combustion-related processes*, IEE SeminarAdvanced Sensors and Instrumentation Systems for Combustion Processes, s. 8/1-8/4, 2000.
7. Wójcik W., *Application of fibre-optic flame monitoring systems to diagnostics of combustion process in power boilers*, Bulletin of the Polish Academy of Sciences – Technical Sciences, tom 56, nr 2, s. 177–195, 2008.
8. Kaczmarek Z., *Światłowodowe czujniki i przetworniki pomiarowe*, Warszawa, Agenda wydawnicza PAK ,2006
9. Wójcik W., Światłowodowy układ do monitorowania procesu spalania, PAK vol.53, nr 11/2007, s.24-28, 2007.
10. Docquier N., Candel S., *Combustion control and sensors, a review*, Progress in Energy and Combustion Science nr 28, s. 107-150, 2002.
11. Kurihara N. et al., A Combustion Diagnosis Method for Pulverized Coal Boilers using Flame-Image Recognition Technology, in IEEE Transactions on Energy Conversion, vol. EC-1, nr 2, s. 99-103, June 1986.
12. Kotyra A. „Diagnostyka procesu spalania pyłu węglowego z wykorzystaniem metod przetwarzania obrazu", Lublin : Politechnika Lubelska, 2010
13. Zhang D., Lu G., Review of shape representation and description techniques, Pattern Recogn, 37 (1) , s. 1-19, 2004.
14. Wójcik W., Kotyra A., Ławicki T., *Wskaźnik jakości spalania pyłu węglowego w oparciu o analizę obrazu transformatą curvelet*, Przegląd Elektrotechniczny, vol. 88 nr 10b/2012, s. 82-84, 2012.
15. Wójcik W., Kotyra A., Ławicki T., Pilek B., *Zastosowanie transformaty curvelet do analizy procesu spalania*, Przegląd Elektrotechniczny, vol. 86 nr 10/2010, s. 131-135, 2010.

16. Talu M. F., Onat C., Daskin M., *Prediction of Excess Air Factor in Automatic Feed Coal Burners by Processing of Flame Images*, Chinese Journal of Mechanical Engineering, vol.: 30, s. 722-731, 2017.
17. Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008.
18. Dobosz M., *Wspomagana komputerowo statystyczna analiza wyników*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2004.
19. Kotyra A., Analiza składowych głównych sygnałów wielokanałowego układu monitorowania płomienia, Przegląd Elektrotechniczny, vol. 86 nr 10/2010, s. 57-60, 2010.
20. Kotyra A., Wójcik W., Golec T., Ławicki T., *Assessment of pulverized coal combustion using Fourier Descriptors*, Przegląd Elektrotechniczny, vol. 86 nr. 7/2010, s. 241-243.
21. El-ghazal A., Basir O., Belkasim S., *Farthest point distance: a new shape signature for Fourier descriptors*,Signal Process Image Commun, 24 (7) , s. 572-586, 2009.
22. Zhang HL., Zou Z., Li J., Chen XT., *Flame image recognition of alumina rotary kiln by artificial neural network and support vector machine methods*, Journal of Central South University of Technology, vol. 15 (2018), s. 39-43.
23. Zieliński T. P., *„Cyfrowe przetwarzanie sygnałów : od teorii do zastosowań”* (Wyd. 2), Wydawnictwo Komunikacji i Łączności, Warszawa 2014.
24. Dzierżak R., Surtel W., Maciejewski M., Dzida G., *Improving the quality of the ECG signal by filtering in wavelet transform domain*, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016 : conference [WOS], 2016.

Wybrane metody obrazowania procesu spalania

Streszczenie

Prawidłowa diagnostyka procesu spalania polega na znalezieniu odpowiednio dobranych metod badawczych oraz prowadzeniu szczegółowej analiz przebiegu tego procesu. W publikacji zostały krótko scharakteryzowane wybrane metody analizy obrazów płomienia. Badanie obrazów z wykorzystaniem analizy składowych głównych daje możliwość redukcji zmiennych, a zastosowanie deskryptorów Fouriera pozwala na wyznaczenie obszaru obrazu. Stosowanie transformacji Falkowych daje możliwość odszumienia oraz kompresji obrazu płomienia.

Słowa kluczowe: proces spalania, analizy składowych głównych, deskryptory Fouriera, transformacji Falkowa, obraz płomienia

Selected methods of imaging the combustion process

Abstract

Correct diagnosis of the combustion process consists in finding the right research methods and carrying out a detailed analysis of the course of the process. The publication has briefly characterized selected methods of analysis of flame images. Study of images using analysis of the principal component analysis gives the opportunity to reduce variables. The use of Fourier descriptors allows you to define the image area. The use of Wavelet transformation gives the possibility to noise reduction and image compression flame.

Keywords: combustion process, principal component analysis, Fourier Descriptors, Wavelet transformation, flame image

Krzysztof Jarosz[1], Piotr Löschner[2]

Przegląd wybranych programów do komputerowej optymalizacji procesów skrawania

1. Wstęp

Współcześnie w projektowaniu procesów wytwarzania części maszyn szeroko wykorzystywane są techniki komputerowe. Taka sytuacja jest bezpośrednio związana z wiodącą rolą obrabiarek sterowanych numerycznie (CNC – ang. *Computerized Numerical Control*) w przemyśle wytwórczym. Większość stosowanych w obróbce skrawaniem maszyn technologicznych, takich jak przykładowo tokarki, frezarki, wiertarki czy szlifierki, posiada sterowanie numeryczne, w związku z czym obróbka realizowana jest w sposób zautomatyzowany według wcześniej zaprojektowanego i wczytanego do pamięci maszyny programu obróbki. Popularnym podejściem jest zaprojektowanie modelu 3D gotowego wyrobu w środowisku wybranego przez użytkownika programu CAD (ang. *Computer Aided Design*), i następnie bezpośrednie zaimportowanie wykonanego modelu do wybranego oprogramowania CAM (ang. *Computer Aided Machining*). Oprogramowanie CAM służy do generacji ścieżek narzędzia w oparciu o geometrię wyrobu finalnego oraz przygotówki oraz szeregu parametrów wejściowych. Wśród nich do najważniejszych zalicza się: geometrię narzędzi skrawających, rodzaj wykonywanych operacji technologicznych, typ ścieżki narzędzia oraz parametry technologiczne procesu jak posuwy, prędkości i głębokości skrawania. Po wygenerowaniu ścieżek narzędzia można dokonać wstępnej weryfikacji ich poprawności w module wizualizacji obróbki (przykładowo funkcje *Backplot* oraz *Visualization* w programach serii Mastercam) oraz wydać programowi polecenie generacji programu sterującego przeznaczonego na obrabiarkę w formie tzw. G-Code. W tym miejscu należy wspomnieć, iż zintegrowana w większości programów CAM opcja weryfikacji programu sterującego nie zawsze pozwala na zauważenie wszystkich potencjalnych błędów. Większość rozwiązań umożliwia jedynie sprawdzenie i ocenę kolizyjności narzędzi z przedmiotem obrabianym. Rozwiązanie takie często okazuje się niewystarczające.

Wraz ze wzrostem stopnia skomplikowania programów sterujących na obrabiarki CNC oraz coraz wyższymi wymaganiami stawianymi wyrobom końcowym, coraz większa popularność zdobywają programy przeznaczone do komputerowej symulacji, weryfikacji i optymalizacji procesów skrawania. Umożliwiają one nie tylko symulację projektowanego procesu technologicznego, ale także weryfikację jego poprawności,

[1]jarosz.krzysztof91@gmail.com, Katedra Technologii Maszyn i Automatyzacji Produkcji, Wydział Mechaniczny, Politechnika Opolska
[2]pleszner@op.pl, Katedra Technologii Maszyn i Automatyzacji Produkcji, Wydział Mechaniczny, Politechnika Opolska

porównanie kilku jego wariantów, a także jego optymalizację, przykładowo celem skrócenia czasu obróbki czy wydłużenia żywotności narzędzi skrawających. Wszystkie te czynności odbywają się w środowisku komputerowym i na etapie projektowania procesu technologicznego, bez konieczności czasochłonnej i kosztownej weryfikacji eksperymentalnej analizowanego procesu wytwarzania. Celem niniejszego opracowania jest prezentacja, opis oraz wzajemne porównanie najpopularniejszych wśród dostępnych na rynku programów do weryfikacji i optymalizacji programów sterujących na obrabiarki CNC. Ocena i porównanie zostaną przeprowadzone względem szeregu kryteriów, takich jak funkcjonalność oprogramowania, podejście do optymalizacji procesu technologicznego czy ilość dostępnych opcji i kryteriów optymalizacji/weryfikacji programów sterujących.

2. Podejścia do komputerowej weryfikacji i optymalizacji procesów technologicznych

Zgodnie z podejściem zaproponowanym w pracy [1], dostępne na rynku programy do optymalizacji programów sterujących na obrabiarki sterowane numerycznie charakteryzować mogą się dwojakim podejściem do problemu optymalizacji: podejściem procesowym bądź technologicznym. Podejścia te są od siebie wzajemnie odmienne i obejmują różne zakresy procesu technologicznego. Cechy charakterystyczne oraz różnice pomiędzy tymi dwoma odmiennymi podejściami zostały ukazane w tabeli 1.

Tabela 1. Cechy charakterystyczne wyróżniające technologiczne oraz procesowe podejście do komputerowej optymalizacji procesów technologicznych

Cechy charakterystyczne podejść do komputerowej ewaluacji i optymalizacji procesów technologicznych	
Podejście technologiczne	Podejście procesowe
Analiza ruchów poszczególnych członów roboczych obrabiarki	Kompleksowa analiza układu OUPN (Obrabiarka-Uchwyt-Przedmiot-Narzędzie)
Uwzględnienie stricte geometrycznych zależności pomiędzy obrabiarką, narzędziem skrawających i przedmiotem obrabianym	Uwzględnienie wzajemnych fizykalnych oddziaływań narzędzia i przedmiotu obrabianego;
Optymalizacja w oparciu o obliczanie wydajności skrawania	Implementacja modeli konstytutywnych oraz metod obliczeniowych opartych na MES (Metodzie Elementów Skończonych);
Czas obróbki jako jedyne dostępne kryterium optymalizacji	Wielokryterialna optymalizacja procesu (czas obróbki, warunki procesu, żywotność narzędzi skrawających)
Brak informacji ze strony oprogramowania o obciążeniu wrzeciona, składowych siły skrawania, temperaturze w strefie skrawania dla procesu bazowego i po optymalizacji	Informacja o parametrach i warunkach procesu w trakcie symulacji jego przebiegu

Źródło: Opracowanie własne

Jak można wywnioskować na podstawie zawartych w tabeli 1. informacji, podejścia procesowe oraz technologiczne są od siebie w znaczny sposób odmienne. Za podstawową wadę podejścia technologicznego w stosunku do procesowego należy uznać brak informacji o wpływie optymalizacji na warunki procesu, co w znaczący sposób może utrudniać właściwą ocenę zoptymalizowanego programu sterującego, a w szczególności jego wpływ na obciążenie obrabiarki, żywotność narzędzi skrawających czy spodziewaną jakość wyrobu końcowego. Kolejnym niedomaganiem w przypadku oprogramowania charakteryzującego się podejściem technologicznym jest możliwość optymalizacji programów sterujących na obrabiarki CNC jedynie w aspekcie minimalizacji czasu obróbki. Jako że programy oparte o to podejście nie uwzględniają oddziaływań fizykalnych pomiędzy przedmiotem i narzędziem, użytkownik końcowy nie uzyskuje informacji o warunkach procesu, co uniemożliwia zastosowania jako kryterium optymalizacji przykładowo składowych siły skrawania, czy obciążenia wrzeciona. Oznacza to, iż niemożliwa jest optymalizacja procesu pod kątem poprawy stabilności obróbki, bądź zmniejszenia obciążenia narzędzi skrawających czy też samej obrabiarki.

3. Wybrane oprogramowania do komputerowej weryfikacji i optymalizacji programów sterujących na obrabiarki CNC

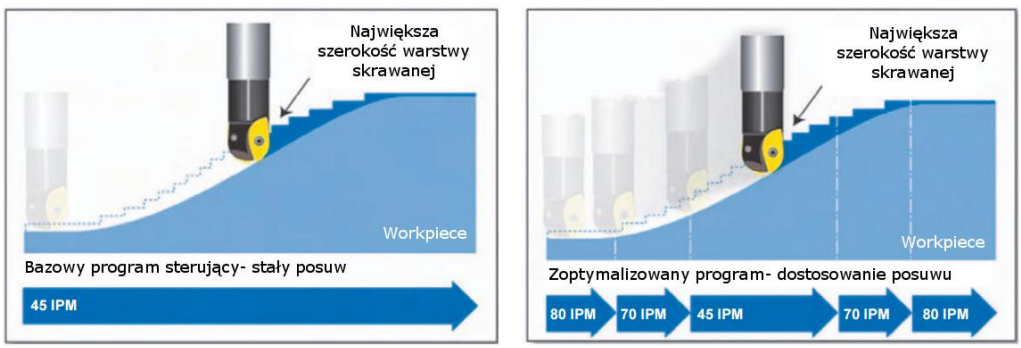
W niniejszej pracy opisane zostały trzy najpopularniejsze oraz najczęściej używane programy do komputerowej optymalizacji programów sterujących na obrabiarki CNC-Vericut firmy CGTech, NCSIMUL opracowany przez SPRING Technologies, oraz Production Module 2D/3D amerykańskiego developera Third Wave Systems. Pierwsze dwa z wymienionych programów opierają się o podejście technologiczne, ostatni zaś wykorzystuje bardziej kompleksowe podejście procesowe.

3.1. CGTechVericut

Szeroko wykorzystywanym przez technologów-programistów CNC jest oprogramowanie Vericut oferowane przez firmę CGTech. Podstawą programu jest symulacja procesu obróbki skrawaniem na wirtualnej obrabiarce CNC w oparciu o rozbudowane i dokładne modele geometryczne CAD 3D jej poszczególnych elementów. Możliwość optymalizacji procesu technologicznego jest zorientowana na możliwość skrócenia czasu obróbki. Dodatkowo, dzięki wizualizacji przebiegu procesu skrawania w czasie rzeczywistym w trójwymiarowym środowisku programu możliwe jest wychwycenie przez użytkownika błędów w programie sterującym obrabiarką, takich jak źle zaprojektowane ścieżki narzędzia, niewłaściwie dobrane naddatki obróbkowe, czy miejsca potencjalnych kolizji. Możliwość wykrywania kolizji nie jest ograniczona jedynie do kolizji między przedmiotem i narzędziem obrabianym, co umożliwia większość popularnych programów CAM. Dzięki implementacji dokładnych modeli 3D całej obrabiarki możliwe jest wykrycie o wiele poważniejszych i potencjalnie wielokrotnie bardziej kosztownych kolizji, przykładowo pomiędzy elementami głowicy narzędziowej i wrzeciona obrabiarki. Biorąc pod uwagę wzrastający stopień skomplikowania zarówno samych obrabiarek CNC (wyposażanych w między innymi chwytaki, manipulatory, stoły symultaniczne uchylno-obrotowe, systemy wymiany narzędzi),

a także samych przedmiotów obrabianych (a więc naturalnie również programów sterujących niezbędnych do ich wykonania), jest to rozwiązanie szczególnie cenne i pożądane. Dodatkową zaletą jest możliwość importowania samodzielnie wykonanych modeli CAD 3D obrabiarek do środowiska programu, co umożliwia praktycznie nieograniczone wykorzystanie programu Vericut, niezależnie od posiadanej przez użytkownika końcowego obrabiarki. Dodatkowo program wyposażony jest w obszerną bibliotekę gotowych modeli 3D obrabiarek z szerokiego wachlarza oferty renomowanych producentów [1, 2].

Sama optymalizacja ścieżek ruchu narzędzia odbywa się w programie Vericut poprzez ich analizę pod kątem miejsc występowania nagłych zmian kierunku czy wartości ruchu posuwowego, głębokości skrawania bądź grubości ścianek samego przedmiotu obrabianego. Obszary te są identyfikowane przez moduł oprogramowania nazwany *Optipath* jako „krytyczne". Po ich zidentyfikowaniu oprogramowanie proponuje zmianę wartości występujących w bazowym programie sterującym wartości posuwów narzędzia celem minimalizacji czasu obróbki [2]. Działanie modułu Optipath zostało zobrazowane na rys. 1



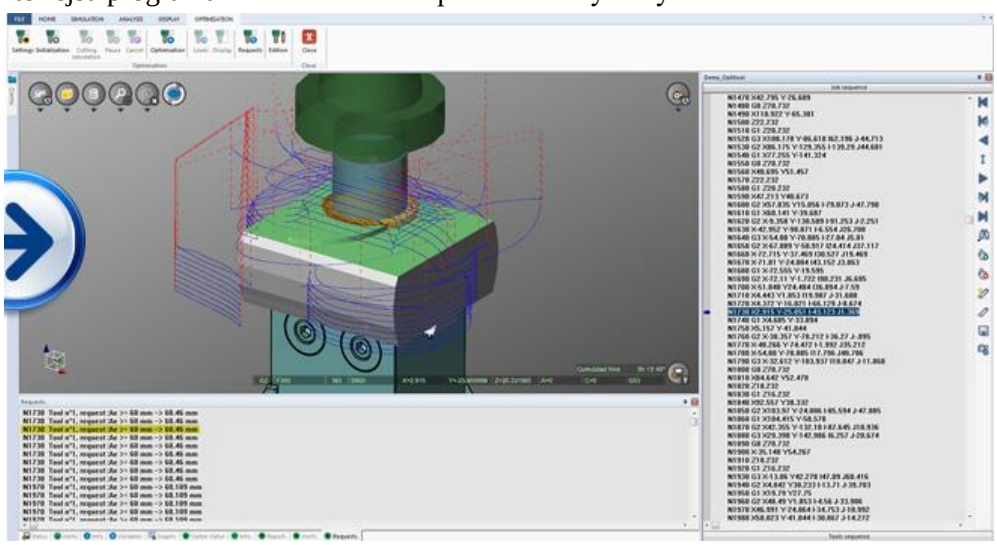
Rysunek1. Zasada optymalizacji w programie Vericut [2]

Zgodnie z zasadą działania przedstawioną na rys.1, Vericut zmniejsza wartość posuwu w obszarach, gdzie narzędzie skrawające usuwa większą ilość materiału, jednocześnie zwiększając wartości posuwu tam, gdzie ilość materiału usuwana przez narzędzie skrawające jest mniejsza. Zmiana wartości posuwów odbywa się automatycznie – obliczone przez program wartości są w czasie rzeczywistym wpisywane do zoptymalizowanej wersji programu sterującego. Należy w tym miejscu nadmienić, iż sama ścieżka ruchu narzędzia nie ulega zmianie [2].

Przykłady udanego wykorzystania oprogramowania Vericut odnaleźć można między innymi w literaturze [3-5]. Możliwe obszary zastosowania obejmują między innymi redukcję błędów kształtu w obróbce elementów cienkościennych, poprawę jakości wykończenia powierzchni po obróbce 5-osiowej czy wreszcie optymalizację procesów wytwarzania elementów o złożonej geometrii, przykładowo wirników sprężarek. Na koniec nadmienić należy, iż Vericut jest w pełni kompatybilny z wiodącymi systemami CAD/CAM, jak przykładowo Catia Pro/ENGINEER, Mastercam czy Edgecam [2].

3.2. SPRING Technologies NCSIMUL

Oprogramowaniem o funkcjonalności porównywalnej z pakietem Vericut jest NCSIMUL oferowany przez SPRING Technologies. Jego zasada działania jest zbliżona do opisanego w poprzednim podrozdziale oprogramowania. NCSIMUL określa geometryczne interakcje pomiędzy poszczególnymi podzespołami wirtualnej obrabiarki, narzędziami i przedmiotem obrabianym w oparciu o ich modele 3D. Umożliwia to wykrycie ewentualnych kolizji pomiędzy tymi elementami. Oprogramowanie zostało wyposażone w moduł wizualizacji procesu z uwzględnieniem stanu powierzchni obrobionej, dzięki czemu możliwa jest przybliżona ocena spodziewanej chropowatości i jakości powierzchni obrobionej. Po obserwacji przez użytkownika końcowego przebiegu symulacji procesu technologicznego, możliwa jest edycja programu sterującego obrabiarką CNC przy pomocy wbudowanego w oprogramowanie NCSIMUL edytora. Po ewentualnej korekcie programu istnieje oczywiście możliwość jego weryfikacji na drodze powtórnej symulacji procesu [1, 6]. Wygląd interfejsu programu NCSIMUL został przedstawiony na rys. 2.

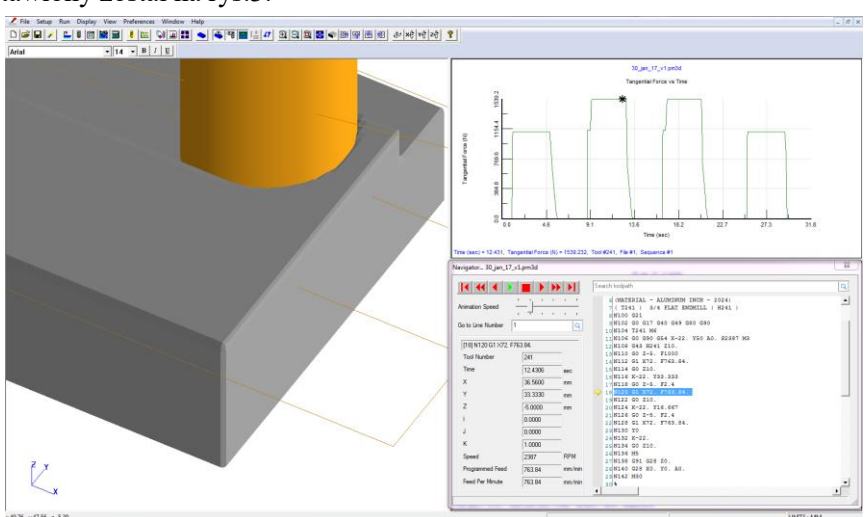


Rysunek2. Okno programu NCSIMUL [6]

Cechą charakterystyczną oprogramowania NCSIMUL są zintegrowane moduły mające ułatwiać zarządzanie procesem produkcyjnym, takie jak ToolSimul czy ToolExpert. Jedną z ich podstawowych funkcji jest szybkie wyszukiwanie odpowiednich dla danej operacji obróbkowej narzędzi skrawających z dostępnej w programie biblioteki. Dodatkowo możliwe jest obliczanie parametrów technologicznych dla konkretnego narzędzia i jego wybranego zastosowania, definiowanie układów współrzędnych narzędzi oraz przywołanie parametrów korekcyjnych dla danego narzędzia po uprzednim ich zapisaniu w bazie. Dodatkowo, oprogramowanie współpracuje z wiodącymi systemami CAD/CAM, jak chociażby Mastercam, PowerMill, HyperMill, Edgecam, Catia, Pro/ENGINEER [6].

3.3. Third Wave Systems Production Module 2D/3D

Pakiet oprogramowania Production Module 2D/3D autorstwa developera Third Wave Systems prezentuje diametralnie odmienne od dwóch uprzednio opisanych podejście do optymalizacji programów sterujących. Oprogramowanie to jest dostarczane w dwóch osobnych wersjach, przy czym zasada ich działania pozostaje taka sama.Pakiet Production Module 2D powstał z myślą o analizie i optymalizacji procesu skrawania na tokarkach sterowanych numerycznie, natomiast wersja 3D do przeznaczona jest do symulacji i optymalizacji procesów prowadzonych na frezarkach CNC. Podstawową cechą pakietu Production Module 2D/3D jest możliwość kompleksowej symulacji procesu skrawania z uwzględnieniem fizykalnych zależności pomiędzy narzędziem i przedmiotem obrabianym. Jest to możliwe dzięki implementacji modeli materiałowych oraz technik obliczeniowych opartych na Metodzie Elementów Skończonych. Dzięki takiemu rozwiązaniu użytkownik programu otrzymuje wyczerpujące informacje o występujących w procesie składowych siły skrawania, obciążeniu wrzeciona, parametrach technologicznych procesu jak posuwy, głębokości i prędkości skrawania, a nawet temperaturze w strefie skrawania. Materiał przedmiotu obrabianego może być wybrany z obszernej biblioteki programu Production Module 3D, bądź zdefiniowany przez użytkownika na podstawie wcześniej przeprowadzonych badań eksperymentalnych (opcja *CustomMaterial*). Wszelkie informacje są prezentowane graficznie w formie wykresów i mogą być śledzone w czasie rzeczywistym w trakcie trwania symulacji procesu dzięki opcji *Navigator*. Dodatkowo okno opcji *Navigator* umożliwia śledzenie w czasie rzeczywistym, która linia programu sterującego jest realizowana w danej chwili obróbki, umożliwiając natychmiastowe wychwycenie ewentualnych błędów i niedociągnięć oraz ich późniejszą edycją we wbudowanym edytorze programów sterujących *McdEdit* [7]. Interfejs programu Production Module 3D z widocznym modułem wizualizacji obróbki, graficzną reprezentacją wyników symulacji oraz oknem opcji Navigator przedstawiony został na rys.3.



Rysunek 3. Widok interfejsu programu Production Module 3D [opracowanie własne]

Optymalizacja procesu w pakiecie Production Module 2D/3D odbywa się na zasadzie zmiany posuwu narzędzia. Kryteria optymalizacji zostają odgórnie określone przez użytkownika programu. Może ona przebiegać według kryterium konkretnej, zadanej przez użytkownika wartości jednego z następujących parametrów: wybranej składowej siły skrawania, obciążenia na wrzecionie lub oporu właściwego skrawania. Procedura optymalizacji programu sterującego uwzględnia możliwość dopuszczalnych górnych i dolnych zakresów parametrów technologicznych, co pozwala na późniejsze wykorzystanie zoptymalizowanego programu sterującego w warunkach rzeczywistych bez obaw o uszkodzenie narzędzi skrawających oraz przeciążenie obrabiarki.

Łatwo wywnioskować, iż Production Module 2D/3D w znaczący sposób różni się od dwóch wcześniej opisanych w niniejszej pracy propozycji oprogramowani do optymalizacji procesów skrawania. Programy Vericut oraz NCSIMUL nie dostarczają informacji o warunkach samego procesu technologicznego, skupiając się na symulacji ruchów roboczych obrabiarki i symulacji geometrycznych zależności występujących w szerzej pojmowanym otoczeniu obróbki. PM 2D/3D skupia się na symulacji procesu obróbki w węższym ujęciu – jako interakcji narzędzia skrawającego i przedmiotu obrabianego- jednakże dostarcza znacznie bardziej obszernych informacji odnośnie warunków procesu, zarówno bazowego, jak i po optymalizacji. Dzięki temu użytkownik może przyjąć różnorodne kryteria optymalizacji, która może polegać już nie tylko na skróceniu czasu obróbki, ale także na poprawie stabilności procesu czy zmniejszeniu kosztów wytwarzania dzięki ograniczeniu zużycia narzędzi skrawających.

4. Podsumowanie

Jak można wywnioskować na podstawie opisów przedstawionego oprogramowania, rozwiązania proponowane przez różnych producentów są od siebie odmienne. Dlatego też nie należy ich ze sobą porównywać, pomimo występowania pewnych podobieństw, takich jak przykładowo możliwość wizualizacji procesu obróbki czy optymalizacja programów sterujących poprzez zmianę wartości posuwu narzędzia w obszarach identyfikowanych przez program jako kluczowe. Najbardziej ewidentne różnice można zauważyć pomiędzy oprogramowaniem Production Module 2D/3D, a programami Vericut/NCSIMUL. PM 2D/3D skupia się na symulacji fizykalnych aspektów procesu obróbki, nie dostarczając użytkownikom informacji o ewentualnych wzajemnych kolizjach elementów obrabiarki. W zamian umożliwia monitorowanie występujących w procesie sił, obciążeń obrabiarki i ogólnie pojmowanych warunków obróbki. Przykładowo, program dostarcza informacji o objętościowej wydajności skrawania, oporze właściwym skrawania czy temperaturze w strefie skrawania.

W tabeli 2. Zawarto krótkie podsumowanie przeglądu wybranych programów do optymalizacji procesów technologicznych na obrabiarkach CNC, wyszczególniając ich cechy charakterystyczne oraz wady i zalety.

Wybór właściwego programu do optymalizacji procesów skrawania zależny jest od potrzeb i preferencji użytkownika końcowego- powinien on zdecydować, czy ważniejsze jest dla niego uniknięcie kolizji na obrabiarce oraz skrócenie czasu obróbki (jednakże bez informacji o wpływie owej optymalizacji na warunki obróbki) czy też

woli skupić się na symulacji i monitorowaniu warunków obróbki, co może być pomocne w przeprowadzeniu optymalizacji procesu technologicznego w sposób racjonalny i niemający na celu jedynie skrócenie czasu obróbki.

Tabela 2. Podsumowanie cech, zalet i wad programów Vericut, NCSIMUL oraz Production Module 2D/3D

Vericut		
Cechy charakterystyczne	Zalety	Wady
-Symulacja obróbki w oparciu o modele 3D podzespołów obrabiarki -Implementacja własnych modeli 3D obrabiarek	-Możliwość wykrycia kolizji pomiędzy elementami obrabiarki -Optymalizacja pozwala skrócić czas obróbki	-Brak informacji o warunkach obróbki przed i po optymalizacji
NCSIMUL		
Cechy charakterystyczne	Zalety	Wady
-Rozszerzenie o moduły zarządzania procesem produkcji -Obliczanie parametrów technologicznych dla narzędzi i operacji	-Wykrywanie kolizji elementów obrabiarki -Wizualizacja stanu powierzchni obrobionej	- Brak danych o wpływie optymalizacji na warunki obróbki
Production Module 2D/3D		
Cechy charakterystyczne	Zalety	Wady
-Symulacja obróbki w ujęciu procesowym -Optymalizacja w zakresie ustalonych przez użytkownika parametrów technologicznych	-Obszerne informacje o warunkach obróbki i parametrach procesu -Możliwość wykorzystania do symulacji danych z eksperymentu	-Brak możliwości wykrycia kolizji pomiędzy elementami obrabiarki

Źródło: Opracowanie własne

Literatura

1. Niesłony P., *Wielokryterialna optymalizacja i racjonalizacja technologii obróbki CNC. Przykłady komercyjnych pakietów programów*, STAL Metale & Nowe Technologie, 11-12 (2012), s. 28-32.
2. http://www.cgtech.com/products/about-vericut/
3. Ratchev S., Liu S., Huang W., Becker A. A., *An advanced FEA based force induced error compensation strategy in milling*, International Journal of Machine Tools and Manufacture, 46/5 (2006), s. 542-551.
4. Lin Z., Fu J., Yao X., Sun Y., *Improving machined surface textures in avoiding five-axis singularities considering tool orientation angle changes*, International Journal of Machine Tools and Manufacture, 98 (2015), s. 41-49.
5. Chen S. L., Wang W. T., *Computer aided manufacturing technologies for centrifugal compressor impellers*, Journal of Materials Processing Technology, 115/3 (2001), s. 284-293.
6. https://www.ncsimul.com/
7. http://www.thirdwavesys.com/production-module/

Przegląd programów do komputerowej optymalizacji procesów skrawania

Streszczenie
W pracy przedstawione zostały przykładowe komercyjne programy CAE (ComputerAided Engineering) przeznaczone do optymalizacji programów sterujących przeznaczonych na obrabiarki CNC. Celem pracy było wzajemne porównanie dostępnych rozwiązań oraz próba zaproponowania najbardziej odpowiedniego z nich, w zależności od oczekiwań użytkowników końcowych. Wyróżniono dwa charakterystyczne i odmienne od siebie podejścia do optymalizacji procesów technologicznych: podejście procesowe oraz technologiczne. Omówiona została zasada działania każdego z programów oraz przyjęte w nich sposoby optymalizacji programów sterujących. W podsumowaniu zawarto wnioski końcowe wraz z wyszczególnieniem cech charakterystycznych, zalet i wad każdego z opisywanych programów.
Słowa kluczowe: optymalizacja, CNC, obrabiarki, skrawanie

Overwiev of software for CNC technological process optimization

Abstract
In this work, the examples of commercially available CAE programs for optimization of CNC toolpaths and technological processes were presented. The aim of this work was to compare the available solutions and to propose the most suitable solution for toolpath optimization in accordance to end user needs. Two distinct approaches to CNC toolpath optimization were distinguished: the process and technological approach. Working principle and optimization methods for each software were described. In the summary, final conclusions were formed, along with a comparison of described software solutions, taking into account their characteristic traits, advantages and shortcomings.
Keywords: optimization, CNC, machine tools, cutting

Agata Jaśkowiec[1]

Systemy i środki bezpieczeństwa w portach lotniczych

1. Wstęp

Bezpieczeństwo to istotny parametr przy wyborze środków transportu. Uważa się, że jednym z najbezpieczniejszych środków transportu jest transport lotniczy. Staje się on coraz powszechniej stosowanym środkiem przemieszczania się, co pokazują statystyki ruchu lotniczego. Podlega on surowym wymaganiom prawnym zarówno krajowym jak i międzynarodowym. O bezpieczeństwie w lotnictwie decyduje nie tylko to, co dzieje się w powietrzu, ale także w porcie lotniczym. Jest to miejsce, które skupia bardzo dużą ilość osób w tym samym czasie, na dużej powierzchni. Bezpieczeństwo w portach lotniczych to problem najwyższej wagi, dlatego też wymaga zaangażowania wielu służb. Wprowadzanie coraz to bardziej innowacyjnych systemów i środków ochrony sprawia, że poziom bezpieczeństwa w tej branży jest na wysokim poziomie. Ogromną rolę w tych systemach odgrywa także czynnik ludzki, jest to zarazem najsłabsze i najsilniejsze ogniwo. Pracownicy portów lotniczych wciąż muszą doskonalić swoje umiejętności, by dobrze wykonywać powierzoną im pracę. To od nich w dużej mierze zależy czy pasażerowie czują się na danym lotnisku bezpiecznie.

2. Akty prawne dotyczące ochrony lotnictwa

Transport lotniczy podlega wymaganiom prawnym międzynarodowym jak i krajowym. Jest to konieczne, aby zapewnić bezpieczeństwo, a także odpowiedni poziom oferowanych usług, co jest związane z koniecznością wdrażania systemów bezpieczeństwa.

Ustawy, rozporządzenia i zarządzenia organów władzy państwowej, odpowiadają za bezpieczeństwo lotnictwa i ochronę portów lotniczych oraz komunikację krajową i międzynarodową. Wśród krajowych aktów prawnych możemy wymienić między innymi:

- Ustawę z dnia 3 lipca 2002 r. Prawo lotnicze (Dz.U nr100 z 2006r., poz. 695 i 696) [1],
- Rozporządzenie Rady Ministrów z dnia 19 czerwca 2007r. w sprawie Krajowego Programu Ochrony Lotnictwa Cywilnego (Dz.U nr 116, poz. 803) [2],
- Rozporządzenie Ministra Infrastruktury z dnia 10 stycznia 2005r. w sprawie Krajowego Programu Kontroli Jakości w zakresie ochrony lotnictwa cywilnego (Dz.U nr 25, poz. 208) [3],
- Rozporządzenie Ministra Infrastruktury z dnia 25 października 2005r. w sprawie Krajowego Programu Szkolenia [4].

Krajowy Program Ochrony Lotnictwa Cywilnego zawiera ustalenia dotyczące przeciwdziałania zagrożeniom bezpieczeństwa w lotnictwie, a także portach lot-

[1]jaskowiec1994@gmail.com, Międzywydziałowe Studenckie Koło Naukowe Ergonomii i BHP, Zakład Ergonomii, Katedra Podstaw Techniki, Wydział Inżynierii Produkcji, Uniwersytet Przyrodniczy w Lublinie

niczych. Rozporządzenie to dostosowuje przepisy wykonawcze do przepisów unijnych związanych z ochroną lotnictwa. Podstawowym celem, dla którego program został opracowany, jest zwiększenie poziomu bezpieczeństwa w lotnictwie. Ustala on zasady organizacji ochrony, działań zapobiegawczych oraz naprawczych w przypadku bezprawnej ingerencji. Została w nim również opisana współpraca z zakresu administracji publicznej, służb odpowiedzialnych za bezpieczeństwo, zarządzających lotniskiem oraz innych służb wykonujących pracę związaną z kontrolą bezpieczeństwa osób, bagażu, ładunków oraz przesyłek pocztowych. Program ten podejmuję szczegółowy opis działań dotyczących profilaktyki antyterrorystycznej [5].

Weryfikacją i oceną zagrożeń oraz nadzorem nad przestrzeganiem obowiązujących procedur zajmuję się powołany przez program zespół ochrony lotniska. Tworzy go zarządzający lotniskiem lub osoba przez niego wyznaczona, komitet skupiający przewoźników lotniczych oraz przedstawiciele Służby Ochrony Lotniska, Policji, ABW, Służby Celnej oraz Straży Granicznej [5].

Krajowy Program Ochrony Lotnictwa Cywilnego składa się z 20 rozdziałów, które dotyczą między innymi: przewozu broni i amunicji, programów ochrony, współpracy międzynarodowej, ochrony statku powietrznego, metod oraz środków kontroli bezpieczeństwa bagażu i pasażerów oraz wykorzystania psów pomocnych w wyszukiwaniu materiałów wybuchowych [6]. Ma on charakter operacyjno-taktyczny. Program ten nie dokonuję analizy zagrożeń, ale ustanawia procedury postępowania, które powinny być stosowane w celu niedopuszczenia do jakichkolwiek zagrożeń bezpieczeństwa w lotnictwie [5].

Aby skutecznie realizować przepisy dotyczące organizacji portów lotniczych Programu Ochrony Lotnictwa Cywilnego został przyjęty Krajowy Program Kontroli Jakości. Ma on na celu wzmacnianie oraz identyfikację słabych stron systemu ochrony. Kontrolę jakości realizują audytorzy. Program obejmuje: audyt ochrony, przegląd ochrony, inspekcję ochrony, test ochrony, badania systemu kontroli [6].

Krajowy Program Szkolenia określa metody prowadzenia szkoleń dotyczących ochrony lotnictwa, chroniącymi od ataków bezprawnej ingerencji. Osoby wykonujące obowiązki w zakresie ochrony portów lotniczych, lotnictwa oraz prowadzących lotniczą działalność gospodarczą są szkolone by właściwie zdobywać kwalifikację w strefie ochrony. Szkolenia prowadzą instruktorzy umieszczani na liście instruktorów przez Prezesa Urzędu Lotnictwa Cywilnego, który nadzoruję krajowy system szkolenia lotniczego w Polsce [7].

Wyróżniamy szkolenia:

- Podstawowe,
- Okresowe,
- Doskonalące i specjalistyczne,
- Dotyczące świadomości ochrony lotnictwa [4].

3. System ochrony portów lotniczych w Polsce

Ochrona lotnictwa skupia się nad działaniami przed ataki bezprawnej ingerencji, w tym ochronie portów lotniczych. Głównym celem powstania systemów ochrony lotnisk jest niedopuszczenie do zaistnienia sytuacji kryzysowej. Gdy taka już się wydarzy, przeciwdziałanie kryzysowi na lotnisku również należy do zadań takiego

systemu. Powstał międzynarodowy system ochrony lotnictwa oraz portów lotniczych, zgodny z regulacjami prawnymi [7]. Obejmuje on:

- Procedury systemu kontroli dostępu personelu poruszającego się po strefach chronionych portu lotniczego w ramach wykonywania obowiązków służbowych,
- Procedury systemu kontroli bezpieczeństwa pasażerów i bagażu w ramach odprawy biletowo-bagażowej,
- Kontrolę jakości systemów bezpieczeństwa, szkolenie personelu oraz planowanie kryzysowe,
- Zbieżne z regulacjami międzynarodowymi normy wewnętrzne państw członków Organizacji Międzynarodowego Lotnictwa Cywilnego (ICAO) [8].

3.1. Techniczne systemy stosowane na lotniskach

- System sygnalizacji włamania i napadu,
- Zintegrowany system kontroli bezpieczeństwa (system telewizji przemysłowej wraz z kamerami, które monitorują wydarzenia zewnątrz i wewnątrz terminalu),
- System zabezpieczenia ogrodzenia w najbardziej niebezpiecznych miejscach.
- System kontroli pasażera i personelu,
- System rejestracji bagażu współpracujący z systemem kontroli odlotów w celu szybkiej identyfikacji bagażu z pasażerem,
- System ochrony wewnętrznych stref – obiektów o szczególnym znaczeniu (generatory prądu, anteny radiowe wysokiej częstotliwości, radiolatarnie, stacje meteorologiczne, magazyny i inne),
- Systemy przepustowe,
- Biometryczne systemy identyfikacji personelu w strefach najwyższej ochrony (identyfikacja kształtu twarzy, linii papilarnych, obrazu rogówki, kształtu dłoni, co pozwala na szybkie rozpoznanie osoby spośród tłumu),
- System monitoringu pojazdów na lotnisku,
- Techniczne środki ochrony w postaci elektronicznych systemów (wykorzystanie lokalnej sieci komputerowej, która synchronizuje współpracę poszczególnych podsystemów i powadzi bazę danych,
- Systemy odstraszania ptactwa: techniczne (akustyczne systemy dźwiękowe, armatki hukowe, metoda włączonych reflektorów) oraz metody biologiczne (pomiar wysokości trawy, odstraszanie przez wyszkolone psy, a także sokolnictwo) [6].

3.2. Systemy zarządzania bezpieczeństwem

Utrzymanie wysokiego poziomu bezpieczeństwa w lotnictwie to główne zadanie Systemu Zarządzania Bezpieczeństwem – (Safety Management System – SMS). Opracowany został przez Międzynarodową Organizację Lotnictwa Cywilnego. Wdrażany program obejmuje zasady organizacyjne oraz wytyczne funkcjonalne systemu zarządzania bezpieczeństwem [9].Jest to sformalizowany zbiór zasad postępowania i procedur bezpieczeństwa, stosowany przez struktury organizacyjne, połączone odpowiednimi relacjami. SMS wymaga posiadania w organizacji lotniczej struktury organizacyjnej do zarządzania bezpieczeństwem. Musi ona posiadać takie zdolności jak: generowanie celów w zakresie bezpieczeństwa, stosowanie procedur

audytu i kontroli, identyfikacji zagrożeń oraz wprowadzania działań korygujących i naprawczych, a także ciągle doskonalić się w zakresie bezpieczeństwa. Celem tego systemu jest wskazanie i wdrożenie zasad bezpieczeństwa, które gwarantują osiągnięcie wysokiego poziomu wydajności organizacji Rada ICAO wymaga od zarządzających organizacjami lotniczymi wdrożenia Systemu Zarządzania Bezpieczeństwem celem zapewnienia bezpieczeństwa rozumianego, jako stan, w którym ryzyko wystąpienia niebezpieczeństwa jest zredukowane do akceptowalnego poziomu, bądź poniżej tego poziomu i jest utrzymywane w takim stopniu poprzez zarządzanie i identyfikację zagrożeń i ryzyka [10].

4. Środki ochrony portów lotniczych

Lotnisko jest to obiekt o podwyższonym ryzyku zagrożenia. Dlatego ważne jest, aby zastosować określone działania, metody oraz środki ochrony przeciwko bezprawnej ingerencji. Najważniejsze obiekty infrastruktury krytycznej portów lotniczych, które podlegają ochronie to:

- Terminale pasażerskie oraz cargo,
- Wieża kontroli ruchu lotniczego,
- Generatory energetyczne,
- Magazyny paliwa,
- Systemy wentylacji,
- Ujęcia wody,
- Płyty postojowe statków powietrznych,
- Hangary,
- Inne uznane przez zarządzającego lotniskiem oraz Prezesa ULC urządzenia i obiekty [11].

Najważniejszymi elementami ochrony portów lotniczych jest: patrolowanie, kontrola dostępu oraz kontrola bezpieczeństwa.

Istotnym elementem ochrony lotniska jest kontrola dostępu. Poprzedzona jest procedurami związanymi z systemem przepustkowym pracowników, bądź osób ubiegających się o wejście na teren lotniska. Ma to na celu sprawdzenie, czy dana osoba nie stwarza zagrożenia oraz zapobiega przedostaniu się nieupoważnionych osób i pojazdów na teren zastrzeżony [12]. Zgodnie z obowiązującym prawem w Polsce za kontrolę dostępu, w tym nadzór nad systemem przepustkowym odpowiada w całości zarządzający lotniskiem [2].

W ramach tej kontroli realizowane są procedury związane z programem ochrony lotniska, który jest niejawny oraz instrukcją przepustkową [13].

System kontroli dostępu oparty jest na technologii RFID (Radio Frequency Identification). Wykorzystuje on karty RFID i składa się z centrali, połączonej do PC oraz rozmieszczonych na terenie obiektu modułów wykonawczych – czytników kart zbliżeniowych, klawiatur do wprowadzania kodu oraz czytników biometrycznych [14].

System przepustkowy - dla wszystkich osób i pojazdów wykonujących czynności służbowe na lotnisku musi posiadać ważną przepustkę osobową upoważniającą do przebywania w określonych strefach lotniska [14]. System ten kontroluje zarządzający lotniskiem, który wydaje przepustki po przeszkoleniu wnioskodawcy w zakresie świadomości ochrony lotnictwa. Zarządzający we współdziałaniu ze Strażą Graniczną

i Policją nadzoruje prawidłowość funkcjonowania systemu przepustowego i ustala system patrolowania stref zastrzeżonych lotniska, ogrodzenia zewnętrznego i miejsc do niego przyległych. Przepustki wydawane dla osób i pojazdów dzielą się na stałe, tymczasowe i jednorazowe.

Prawidłowo przeprowadzona kontrola bezpieczeństwa to podstawa bezpieczeństwa w portach lotniczych. Jest to zespół działań, środków oraz metod podejmowanych przez służbę ochrony w celu wyeliminowania niebezpieczeństwa w strefie zastrzeżonej lotniska oraz na pokładzie statku powietrznego. Przed wejściem pasażera do strefy zastrzeżonej musi odbyć się kontrola bezpieczeństwa [15].

Kontrola bezpieczeństwa obejmuję:

- Sprawdzanie zgodności i prawidłowości dokumentów, które uprawniają do wstępu na pokład samolotu,
- Sprawdzenie, czy pasażer nie posiada przedmiotów niedozwolonych do przewozu na pokładzie samolotu w szczególności materiałów i przedmiotów niebezpiecznych,
- Zapewnienie odpowiednich warunków, by pasażer i jego bagaż nie miał dostępu do osób niepoddanych kontroli bezpieczeństwa [7].

Jeżeli chodzi o metody kontroli bezpieczeństwa osób możemy wyróżnić:

- Detektory śladowych ilości materiałów wybuchowych (Rys. 1a),
- Ręczne detektory metalu (Rys. 1b),
- Bramki do wykrywania metali (WTMD) (Rys. 2),
- Kontrolę manualną (Rys. 3),
- Psy do wykrywania materiałów wybuchowych [16].

a)

b)

Rys. 1 a) detektor śladowych ilości materiałów wybuchowych [17] b) ręczny detektor metalu[18]

Metody kontroli bezpieczeństwa bagażu:

- Kontrola wzrokowa,
- Urządzanie rentgenowskie do prześwietlania bagażu (Rys, 4),
- Kontrola manualna,
- Detektory śladowych ilości materiałów wybuchowych,
- Ręczne wykrywacze metali (HHMD),
- Psy do wykrywania materiałów wybuchowych,
- Systemy do wykrywania materiałów wybuchowych (EDS),
- Systemy do wykrywania urządzeń wybuchowych (EDDS) [16].

W stosunku do kontroli przewożonych ciecz, żeli i aerozoli stosuję się:

- Kontrolę smakowa lub badanie na powierzchni skóry,

- Skanery cieczy w butelkach,
- Paski próbne do badania reakcji chemicznych,
- Kontrola wzrokowa,
- Detektory śladowych ilości materiałów wybuchowych,
- Psy do wykrywania materiałów wybuchowych,
- Systemy do wykrywania materiałów wybuchowych (EDS),
- Urządzenie rentgenowskie.

Rys. 3 Kontrola manualna [20]

Rys 4 Urządzenie rentgenowskie do prześwietlania bagażu [21]

W momencie, gdy jest umożliwione korzystanie z bramki do wykrywania metali, stosuję się wyrywkowe rewizję manualne kontrolowanych osób. Manualne kontrole znajdują zastosowanie u wszystkich pasażerów, którzy uruchomili alarm w urządzeniu. W takim wypadku pasażer ponownie przechodzi przez bramkę do wykrywania metali lub zostaje przeszukany przez ręczny detektor metali. W chwili, gdy pracownik nie jest w stanie określić czy podróżujący nie posiada przy sobie przedmiotów i materiałów niebezpiecznych, poddawany jest kontroli osobistej w wyznaczonym miejscu. Kontrola odbywa się z osobą tej samej płci [16].

5.

6. Służby odpowiedzialne za bezpieczeństwo portów lotniczych

W celu sprawnego funkcjonowania systemu ochrony portu lotniczego powołuję się tzw. Zespół Ochrony Lotniska na podstawie ustawy z dnia 3 lipca 2002r. Prawo Lotnicze (Dz.U nr 130, poz. 1112), w którego skład wchodzi zarządzający lotniskiem oraz po jednym przedstawicielu służb:

- Służby Ochrony Lotniska,
- Lotniskowej Straży Pożarnej,
- Policji,
- Straż Graniczna,
- Służba Celna,
- Jednostki Ratownictwa Medycznego,
- Przewoźników lotniczych działających na lotnisku,
- Innych podmiotów prowadzących lotniczą działalność gospodarczą [1].

Do zadań Zespołu Ochrony Lotniska należy: ocena niebezpieczeństwa ataków bezprawnej ingerencji oraz prowadzenie czynności im zapobiegającym, opracowanie poleceń i zarządzeń w dziedzinie ochrony lotniska, prezentowanie i zgłaszanie wniosków i opinii do programu ochrony lotniska [22].

Urząd Celny – jest odpowiedzialny za kontrolę bagażu przed załadunkiem w celu sprawdzenia czy nie znajdują się w nim materiały pirotechniczne oraz inne, których przewóz jest zabroniony lub podlega ocleniu. Urząd zajmuję się również ochroną strefy celnej lotniska, odprawą celną, a także sprawdzaniem zgodności obrotu towarowego z zagranicy z obowiązującymi aktami prawnymi [7].

Agencja Celna – zajmuje się obsługą lotniczych przesyłek towarowych przyjmowanych i nadawanych na lotnisku. Przestrzega procedury kontroli stosowane przez Straż Graniczną przy przesyłkach wychodzących z lotniska. Celem pracy Agencji jest: uzyskanie decyzji dopuszczającej towar do wywiezienia z kraju lub obrotu na terenie Polski, zgłaszanie towarów zwierzęcych lub spożywczych do odpowiedniej kontroli, obliczanie należności celnych i ustalanie taryfy celnej towarów, prowadzenie dokumentacji oraz magazynu celnego [7].

Spółka holdingowa – prowadzi działalność usługową w zakresie obsługi pasażerów i ich bagażu oraz statków powietrznych. W zakresie ochrony do obowiązków spółki należy znajomość obowiązujących procedur i planów postępowania w momencie zaistnienia zagrożenia a także treści programów ochrony i Programu Ochrony Lotniska, szkolenie personelu w zakresie znajomości zasad ochrony lotnictwa cywilnego, zapewnienie odprawy pasażerów i ich bagażu zgodnie z procedurami, współpraca z służbami ochrony na lotnisku, zaostrzenie obowiązujących zasad na trasach zwiększonego ryzyka [23].

Policja - funkcjonariusze policji w porcie lotniczym odpowiedzialni są za ochronę porządku publicznego. Prowadzą działania antyterrorystyczne. Systematycznie patrolują obszar całego lotniska oraz ogrodzenia. Policjanci mają również prawo legitymować, kontrolować i zatrzymywać osoby, pojazdy i bagaże oraz nakładać mandaty karne. Biorą oni również czynne uczestnictwo w pracach Zespołu Ochrony

Lotniska. Funkcjonariusze współdziałają ze służbami odpowiedzialnymi za bezpieczeństwo i ochronę w porcie lotniczym [22].

Straż Graniczna – odpowiada za ochronę granicy państwowej. Do głównych jej zadań należy: przeprowadzanie odpraw paszportowych pasażerów, kontrolowanie bezpieczeństwa pasażerów i ich bagażu oraz przesyłek towarowych i ładunków w połączeniach międzynarodowych, organizowanie i dokonywanie kontroli ruchu granicznego, współpraca z innymi służbami na lotnisku, rozpoznawanie materiałów niebezpiecznych [22].

Służba Ochrony Lotniska (SOL) – formacja mundurowa wchodząca w skład systemu ochrony fizycznej portu lotniczego. Jest to specjalistyczna służba zapewniająca bezpieczeństwo pasażerom, statkom powietrznym i pracownikom portu. Działają na terenie całego lotniska. Do głównych zadań tej służby należy:

- Ochrona strefy zastrzeżonej lotniska,
- Kontrola przepustek wydawanych przez zarządzającego lotniskiem,
- Prowadzenie kontroli bezpieczeństwa,
- Zatrzymanie i przekazanie policji lub Straży Granicznej osób, które usiłowały umieścić lub umieściły broń, urządzenia lub materiały wybuchowe oraz urządzenia i substancje zagrażające bezpieczeństwu na pokładzie statku powietrznego,
- Ochrona portu lotniczego i jego obiektów,.
- Kontrola bezpieczeństwa załogi, pasażerów, bagażu, przesyłek towarowych oraz ładunków ruchu krajowego.
- Ochrona obszaru lotniska, obiektów lotniskowych i część manewrowej [7].

Za ochronę lotnisk zgodną z Planem Ochrony Lotniska jest Komendant Służby Ochrony Lotniska. Do jego zadań należy:

- Sporządzenie planu i dokumentacji ochrony lotniska.
- Monitorowanie bezpieczeństwa pasażerów i bagażu w lotach krajowych.
- Kierowanie pracą oddziału, nadzorowanie oraz kontrola nad zadaniami wykonywanymi przez SOL.
- Przeprowadzanie okresowych analiz i ocen poziomu ochrony, nadzorowanie oraz przechowywanie broni i amunicji [22].

7. Podsumowanie

Podróżni coraz częściej wybierają transport lotniczy, jako szybki środek transportu. Wiąże się z tym częste przebywanie dużej liczby osób w porcie lotniczym, a co za tym idzie ogrom przedsięwzięć, jakie muszą wykonać pracownicy, aby zapewnić bezpieczeństwo podczas przebywania na lotnisku. Jest to osiągane poprzez ciągłe doszkalanie służb w instytucjach lotniczych, a także szczegółową selekcję kandydatów. Systemy bezpieczeństwa, akty prawne oraz metody i środki ochrony pozwalają utrzymywać poziom bezpieczeństwa w portach lotniczych na najwyższym poziomie.

Przedstawione systemy i środki ochrony bezpieczeństwa lotnisk to nowoczesne systemy łączące rozwiązania technologiczne z systemami zarządzania, dostosowujące się do ciągle to nowych zagrożeń występujących w lotnictwie. Największą rolę w tym

systemie odgrywa człowiek – jest jego najmocniejszym, jak i najsłabszym ogniwem. Ważne jest, aby wszyscy pracownicy znali zagrożenia występujące na lotnisku, wiedzieli jak im zapobiegać i chronić przed atakami bezprawnej ingerencji tworząc z wszystkimi technicznymi środkami ochrony jeden spójny system bezpieczeństwa portów lotniczych.

Literatura

1. Ustawa z dnia 3 lipca 2002 r. Prawo lotnicze (Dz.U nr100 z 2006r., poz. 695 i 696).
2. Rozporządzenie Rady Ministrów z dnia 19 czerwca 2007r. w sprawie Krajowego Programu Ochrony Lotnictwa Cywilnego (Dz.U nr 116, poz. 803).
3. Rozporządzenie Ministra Infrastruktury z dnia 10 stycznia 2005r. w sprawie Krajowego Programu Kontroli Jakości w zakresie ochrony lotnictwa cywilnego (Dz.U nr 25, poz. 208).
4. Rozporządzenie Ministra Infrastruktury z dnia 25 października 2005r. w sprawie Krajowego Programu Szkolenia.
5. Pomykała M. *Bezpieczeństwo operacji morskich i lotniczych*. [w] Bezpieczeństwo w lotnictwie w różnych aspektach działalności lotniczej, Wydawnictwo Wyższej Szkoły Sił Powietrznych. Dęblin. 2013.
6. Rajchel J., Grenda B., Nowak J. *Bezpieczeństwo i zarządzanie kryzysowe w lotnictwie.* Wydawnictwo Wyższej Szkoły Sił Powietrznych. Dęblin. 2014.
7. Rajchel J. *Bezpieczeństwo w porcie lotniczym*, Wydawnictwo Wyższej Szkoły Oficerskiej Sił Powietrznych. Dęblin.2010.
8. Dilling M. *Bezpieczeństwo w portach lotniczych* [w] Bezpieczne niebo. AON Warszawa. 2002.
9. http://www.amw.gdynia.pl/library/File/ZN%202010/16Zielinski%20M.pdf?PHPSESSID=bd4f5e989e2d5863870e23c6e1857209 (Dostęp 22.12.2016 r.).
10. Compa T., Rajchel J., Załęski K. *Bezpieczeństwo w lotnictwie w wybranych aspektach funkcjonowania portu lotniczego.*,Wydawnictwo Wyższej Szkoły Sił Powietrznych. Dęblin. 2013.
11. Jurgilewicz M. *Zarys prawnych uwarunkowań ochrony infrastruktury krytycznej na obszarze portów lotniczych*. [w] Bezpieczeństwo w lotnictwie w różnych aspektach działalności lotniczej,Wydawnictwo Wyższej Szkoły Sił Powietrznych. Dęblin. 2013.
12. Siadkowski A. *Bezpieczeństwo i ochrona w cywilnej komunikacji lotniczej na przykładzie Polski, Stanów Zjednoczonych i Izraela*, Wydawnictwo Wyższej Szkoły Policji w Szczytnie. Szczytno. 2013.
13. Rzadkiewicz H. *Bezpieczeństwo pasażerów i infrastruktury w portach lotniczych. [w] Bezpieczeństwo w lotnictwie w różnych aspektach działalności lotniczej*, Wydawnictwo Wyższej Szkoły Sił Powietrznych. Dęblin.2013.
14. Bagińska A. *Zapewnienie bezpieczeństwa w porcie lotniczym Łódź-Lublinek.* [w] Kwasiborska A. Bezpieczeństwo transportu lotniczego, Oficyna Wydawnicza ASPRA-JR. Warszawa.2016.
15. Szankin T. *Bezpieczeństwo w komunikacji lotniczej na tle stanu bezpieczeństwa państwa, ujęcie prawnokarne*. [w] A. Siadkowski, A. Tomasik. Bezpieczeństwo i ochrona Lotnictwa Cywilnego. Poznań.2012..
16. Ceglarski M. *Bezpieczeństwo w porcie lotniczym – wybrane procedury kontroli bezpieczeństwa pasażerów i ich bagażu*. [w] Compa T., Rajchel J., Załęski T.

Bezpieczeństwo w lotnictwie w różnych aspektach działalności lotniczej. Wydawnictwo Wyższej Szkoły Sił Powietrznych. Dęblin.2013..
17. http: //www.transactor.pl/images/mat_wybuchowe_03.jpg (Dostęp 02.12.2016 r.).
18. http://astrophysics.pl/userfiles/images/RTG1.JPG (Dostęp 02.12.2016 r.).
19. http://www.spyshop.pl/5632-thickbox_default/profesjonalna-bramka-do-wykrywania-metali-garrett-pd-6500i.jpg (Dostęp 02.12.2016 r.).
20. http://dlapilota.pl/files/Kontrola%20na%20lotnisku.jpg (Dostęp 02.12.2016 r.).
21. http://astrophysics.pl/userfiles/images/RTG1.JPG (Dostęp 02.12.2016 r.).
22. Kowalski R. Systemy, metody oraz środki ochrony w porcie lotniczym Poznań-Ławica. [w] Kwasiborska A. Bezpieczeństwo transportu lotniczego. Oficyna Wydawnicza ASPRA-JR. Warszawa.2016.
23. Sztucki J.,Gąsior M.,Zając G.,Szczelina M. *Zarządzanie bezpieczeństwem lotnictwa cywilnego.* Skrypt dydaktyczny. (Dostęp 12.11.2016r.).

Systemy i środki bezpieczeństwa w portach lotniczych

Streszczenie
Transport lotniczy jest najszybciej rozwijającą się gałęzią transportu. Uważa się, że jest jednym z najbezpieczniejszych środków przemieszczania się. O bezpieczeństwie decyduje nie tylko to, co dzieje się w powietrzu, ale także w porcie lotniczym, dlatego poświęca się mu tak dużo uwagi. Bezpieczeństwo w portach lotniczych to problem najwyższej wagi, dlatego też wymaga zaangażowania wielu służb. Wprowadzanie coraz to bardziej innowacyjnych systemów i środków ochrony sprawia, że poziom bezpieczeństwa w tej branży jest na wysokim poziomie.
Celem pracy jest opis systemów bezpieczeństwa w portach lotniczych. Scharakteryzowane zostały systemy oraz środki bezpieczeństwa stosowane w portach lotniczych, a także służby pracujące na lotnisku.
Słowa kluczowe: systemy bezpieczeństwa, port lotniczy, środki bezpieczeństwa

Systems and security measures at airports

Abstract
Air transport is the fastest growing branch of transport. It is believed to be one. The safest means of movement. Safety is what determines not only what is happening in the air, but also at the airport, so attention is paid so much to it.Airport security is a matter of utmost importance and therefore requires the involvement of many services.Introducing ever more innovative systems and protection means that the level of security in this industry is high.The purpose of the study is to describe safety management systems at airports. The systems and security measures at airports, and also the servants working at the airport are described.
Keywords: security systems, airport, security measures

Katarzyna Ignatiuk[1]

Bezpieczeństwo w medycznych systemach informacyjnych

1. Wstęp

Dokumentacja medyczna pacjenta stanowi kluczowy element podczas jego leczenia, gdyż zawiera wszystkie informacje o stanie zdrowia, wykonanych badaniach, pobytach w szpitalu i przeprowadzonych zabiegach na przestrzeni wielu lat. Jeszcze kilka lat temu dokumentacja medyczna występowała głównie w formie papierowej. Obecnie powoli zastępują ją formy elektroniczne. Na przestrzeni tych lat nie zmienił się jeden fakt – to, że to pacjent najczęściej odpowiada za jej dostarczenie do innej placówki leczniczej. W związku z tym, w nagłych wypadkach, podczas leczenia w nowej placówce, stan wiedzy o danym pacjencie jest zerowy. Problem ten rozwiązuje wprowadzenie EHR, czyli Elektronicznego Dokumentu Zdrowotnego (*ang. Electronic Health Record*) [1]. EHR jest wirtualnym dokumentem, na który składa się wszelaka dokumentacja medyczna w formie cyfrowej należąca do pojedynczego pacjenta. Dzięki temu rozwiązaniu informacje o pacjencie mogą być tworzone, przechowywane i używane w wielu różnych organizacjach ochrony zdrowia oraz udostępniane pacjentowi w jednym dokumencie w aplikacji internetowej.

Wprowadzenie elektronicznego systemu zarządzania dokumentami zdrowotnymi przynosi korzyści, ale również generuje nowe problemy. Wśród pozytywów można wymienić zwiększenie jakości opieki zdrowotnej pacjenta, sprawniejsze i dużo bardziej wydajne zarządzanie (system e-recept umożliwia kontrolę niepożądanych interakcji między lekami, przypisanymi i przyjmowanymi w jednym momencie), wspomaganie decyzji lekarzy i redukcja błędów medycznych nawet o 55% [1], leczenie na odległość, co jest dobre w dużych miastach, międzymiastowo, międzykontynentalnie. Najpoważniejszym skutkiem przeniesienia zasobów ze szpitalnych baz danych do sieci są problemy z kontrolą i ochroną informacji zawartych w dokumentach medycznych. Problem stanowi również integralność obiektów cyfrowych w przypadku wielomodułowych systemów EHR.

Niniejsza praca ma na celu przedstawienie podstawowych zagrożeń w stosunku do bezpieczeństwa medycznych systemów informacyjnych oraz metod ochrony ze zwróceniem uwagi na sposoby ukrywania wiedzy w medycznych bazach danych.

[1] ignatiukkatarzyna@gmail.com, Zakład Biocybernetyki i Inżynierii Biomedycznej, Wydział Mechaniczny, Politechnika Białostocka

2. Bezpieczeństwo systemów informacyjnych

Najpoważniejszym wyzwaniem stawianym nowoczesnym systemom informacyjnym jest zapewnienie bezpieczeństwa danych wrażliwych. Wynika to nie tylko z przepisów prawa, m. in. ustawy o ochronie danych osobowych z 29 sierpnia 1997 roku, ale również z lojalności wobec pacjentów, którzy powierzyli swoje dane publicznej placówce. Medyczne bazy danych narażone są na różnego rodzaju ataki, bez względu na to czy są przechowywane stacjonarnie w komputerach placówek medycznych, czy też udostępniane innym placówkom leczniczym poprzez aplikacje internetowe. W obu przypadkach dane z baz danych muszą być odpowiednio chronione przed dostępem osób nieupoważnionych. W podrozdziałach tego rozdziału zostaną opisane najpopularniejsze kategorie zagrożeń systemów medycznych ze zwróceniem szczególnej uwagi na bezpieczeństwo medycznych aplikacji internetowych oraz sposób analizy tych zagrożeń.

2.1. Bezpieczeństwo medycznych aplikacji internetowych

Dane medyczne mogą być udostępnione pacjentowi w aplikacjach internetowych za pośrednictwem przeglądarek internetowych. Na coraz większą popularność tej formy przekazywania informacji pacjentowi i lekarzom wpływa szybki i łatwy dostęp do Internetu [2, 3].

Najnowsze medyczne aplikacje internetowe są typu klient-serwer [2]. Charakteryzują się tym, że nie instaluje się ich na lokalnym komputerze użytkownika, a uruchamiane są za pośrednictwem przeglądarki internetowej. Użytkownik steruje aplikacją poprzez listy wyboru oraz edycję pól. Aplikacje posiadają warstwową architekturę i dzięki podziałowi na niezależne moduły odrębnie odbywa się zarządzanie bazą danych, implementacja logiki biznesowej czy obsługa interfejsu użytkownika. Poszczególne moduły mogą działać na różnych maszynach, niższe warstwy nie muszą posiadać wiedzy na temat wyższych warstw. Każda warstwa może być implementowana w różnych językach, przez różne zespoły [2, 3].

Tak wiele udogodnień niesie ze sobą jeszcze więcej możliwości ataku na dane wrażliwe, dlatego kluczową rolę przy tworzeniu kolejnych medycznych systemów informatycznych odgrywa właściwe zabezpieczenie każdej warstwy aplikacji oraz ciągłe monitorowanie skuteczności zabezpieczeń.

Celem mechanizmów zabezpieczających jest minimalizacja ryzyka przechwycenia poufnych, danych wrażliwych przez osoby do tego nieupoważnione. Aby minimalizować ryzyko związane z naruszeniem bezpieczeństwa aplikacji, projektuje się konkretne strategie zarządzania ryzykiem. Jednym z popularnie stosowanych modeli ryzyka zagrożeń jest model firmy Microsoft, który składa się z pięciu etapów, takich jak [2]:

1. Zdefiniowanie celów strategii bezpieczeństwa.
2. Wyszczególnienie charakterystyk aplikacji, które będą przydatne do identyfikacji zagrożeń w etapie czwartym.
3. Dekompozycja aplikacji w celu wydzielenia modułów, w których bezpieczeństwo jest kluczowe.

4. Identyfikacja poszczególnych zagrożeń na podstawie informacji z punktu drugiego i trzeciego.
5. Przegląd wszystkich warstw aplikacji, identyfikacja zagrożeń w każdej warstwie i ocena stopnia zagrożenia.

Do kategoryzacji zagrożeń wykorzystuje się metodykę STRIDE (ang. *Spoofing Identify, Tampering, Repudiability, Information Disclosure, Denial of Service, Elevation of Privilege*) [2]. W obrębie tej metodyki wyróżnia się sześć grup zagrożeń i proponuje się metody redukcji tych zagrożeń. Wyodrębniony podczas dekompozycji moduł powinien przejść weryfikację pod kątem wyszczególnionych zagrożeń.

2.1.1. Podstawowe kryteria zagrożeń

W medycznych systemach informatycznych spotyka się sześć głównych kategorii zagrożeń, które opisano poniżej. W każdej kategorii wspomniano również o sposobach walki z tymi zagrożeniami [2].

Podszywanie

Podszywanie się jest podstawowym zagrożeniem bezpieczeństwa aplikacji, polegającym na podszywaniu się pod tożsamość innego użytkownika, aby uzyskać dostęp do danych zastrzeżonych, gdy aplikacja pracuje z takimi samymi uprawnieniami dla wszystkich użytkowników. Najbardziej zagrożeni są użytkownicy administracyjni (z najszerszym zakresem uprawnień), którzy mogą modyfikować medyczne bazy danych. Do redukcji tego zagrożenia wykorzystuje się protokół SSL (ang. Secure Sockets Layer). Łączność przy użyciu SSL musi być stosowana w komunikacji ważnej dla bezpieczeństwa danych, przy pobieraniu lub zapisywaniu danych pacjentów, podczas przekazywania między komputerami w sieci ciasteczek (and. Cookies), które zawierają identyfikatory sesji, hasła, loginy. Aby uchronić się przed podszywaniem zaleca się również wykorzystywanie silnych mechanizmów uwierzytelniania, szyfrowania haseł, szyfrowania informacji uwierzytelniających przesyłanych przez sieć oraz kontrolowanie dostępu do profili administracyjnych [2].

Manipulacja na danych

Wśród zagrożeń tej kategorii można wymienić zagrożenie wynikające ze zbytniego zaufania dla walidacji danych przeprowadzanej po stronie klienta. Poprzez modyfikacje metod GET i POST protokołu HTTP można odpowiednio przygotować zapytania, które umożliwią pozyskanie danych wrażliwych. Zagrożeniem tej kategorii jest również nadpisywanie wartości zmiennych środowiskowych serwera WWW, aby uzyskać kontrolę nad aplikacją po stronie serwera. Zagrożenia tej kategorii można zredukować zabezpieczając warstwy danych protokołami zapewniającymi integralność (np. *IPSec*), stosując protokoły odporne na manipulacje czy stosując elektroniczny podpis [2].

Zaprzeczenia akcji

Zagrożenie to polega na utajaniu wprowadzenia modyfikacji w danych, gdy takie modyfikacje zostały jednak wprowadzone. W zastosowaniach medycznych konieczne jest stosowanie mechanizmów śledzenia i kontroli użytkowników oraz weryfikacja przebiegu procesu modyfikacji danych. Do redukcji tego zagrożenia wykorzystuje się podpis elektroniczny, mechanizmy generowania jednokrotnego hasła dla potwierdzenia dokonania modyfikacji, logowanie wszystkich ruchów użytkownika [2].

Ujawnienie informacji

Przyczyną tego zagrożenia może być błędne zachowanie przeglądarki internetowej. Tajne informacje mogą zostać ujawnione przez aplikacje, których klienci pracują na współdzielonych bazach danych. Aby zapobiec niepożądanemu ujawnianiu informacji, medyczna aplikacja internetowa powinna nie dopuszczać do zapamiętywania ważnych danych po stronie klienta, stosować silne mechanizmy autoryzacji oraz szyfrowania, dbać o separację warstwy danych od interfejsu użytkownika i zabezpieczać warstwy komunikacji protokołami poufności (SSL/TLS, IPSec) [2].

Zablokowanie dostępu do usługi

Zagrożenie to polega na zablokowaniu dostępu do aplikacji, które może być przeprowadzone w wielu warstwach aplikacji. Tego rodzaju atak nie przynosi korzyści atakującemu, ale wyrządzić może krzywdę pacjentowi, do którego dokumentacji medycznej dostęp zostaje utracony. Atak DOS (ang. *Denial of Service*) polega na zasypaniu serwera wieloma żądaniami pochodzącymi od różnych użytkowników, nad którymi została przejęta kontrola i zablokowaniu serwera. Najważniejsza jest umiejętność rozróżniania w serwerze wzrostu zainteresowania użytkowników aplikacją. Aby zabezpieczyć się przed tym zagrożeniem należy stosować mechanizmy na wielu poziomach, ale przede wszystkim kontrolować zasoby systemu, ruch sieciowy oraz stosować systemy IDS (ang. *Intrusion Detection System*) – system wykrywania włamań oraz IPS (ang. *Intrusion Prevention System*) – system zapobiegania włamaniom [2].

Nieuprawnione uzyskanie większych przywilejów

Zagrożenie to polega na pozyskaniu przez nieuprawnioną osobę większej liczby uprawnień i ról. Aby zredukować zagrożenie należy kontrolować poziom uprawnień, kierując się zasadą przyznawania najmniejszych wymaganych uprawnień do działania aplikacji, należy stosować mechanizmy separacji procesów i wirtualizację serwerów. Bardzo często aplikacje nie są odporne na ataki XSS (ang. *Cross-Site Scripting).* Atak polega na wykonywaniu kodu, który przyznaje prawa administracyjne użytkownikowi bez wiedzy administratora [2].

2.1.2. Analiza zagrożeń

Aby sprawować odpowiednią kontrolę nad bezpieczeństwem medycznych systemów informacyjnych, należy nieustannie kontrolować poziom zagrożeń wymienionych w podrozdziale 2.1.1. Należy oddzielnie rozpatrywać zagrożenia spowodowane wewnętrzną strukturą aplikacji oraz związane z wewnętrznym i zewnętrznym przepływem danych [1, 2]. Analiza zagrożeń, jakie mogą wystąpić na poziomie każdej wyodrębnionej warstwy aplikacji służy identyfikacji luk w systemie. Należy ocenić każde zagrożenie w skali 0-10, gdzie 0 oznacza najniższe zagrożenie, zaś 10 oznacza znaczny poziom ryzyka. W ocenie stopnia ryzyka posłużyć może metoda DREAD (ang. *Damage, Reproductibility, Exploitability, Affected Users, Discoverability)*, dzięki której określa się poziom pięciu zagrożeń w stosunku do aplikacji [2]. Rodzaje i ocena tych zagrożeń została zawarta w poniższej tab.1.

Tabela 1. Rodzaje i stopnie zagrożeń według metody DREAD

Rodzaj zagrożenia	Stopień zagrożenia
Poziom zniszczeń (w przypadku skutecznego ataku)	0 – brak zniszczeń 5 – ujawnienie poufnych danych użytkowników 10 – całkowite zniszczenie systemu i utrata danych
Trudność w odtworzeniu stanu systemu sprzed ataku	0 – niemożliwy lub trudny do odtworzenia stan 5 – możliwy do odtworzenia stan, pod pewnymi warunkami 10 – prosty do odtworzenia stan
Łatwość wykorzystania luki	0 – wymaga zaawansowanej wiedzy sieciowej i programistycznej oraz zaawansowanych narzędzi 5 – możliwy do wykorzystania z użyciem dostępnych narzędzi 10 – atak możliwy do przeprowadzenia przez osobę bez specjalistycznych kompetencji
Ilość zagrożonych użytkowników	0 – bliska zeru 5 – część użytkowników, nie wszyscy 10 – wszyscy użytkownicy
Poziom trudności w zlokalizowaniu luki	0 – bardzo trudna do zlokalizowania luka 5 – do zlokalizowania podczas monitorowania sieci 10 – łatwa do zlokalizowania nawet przez użytkownika bez specjalistycznej wiedzy

Źródło: [2]

3. Metody ochrony medycznych systemów informacyjnych

W związku z tym, że dane medyczne zawierają wiele danych wrażliwych, powinny podlegać ochronie jednocześnie na wielu płaszczyznach. Ochrona systemów medycznych obejmować powinna zabezpieczenia fizyczne, techniczne, personalne oraz organizacyjne.

Zabezpieczenia fizyczne dotyczą ochrony komputerów i pomieszczeń, w których te komputery się znajdują, przed osobami nieupoważnionymi. Zabezpieczenia techniczne obejmują takie zagadnienia jak: tworzenie kopii zapasowych danych, stosowanie programów antywirusowych i wirtualnych sieci prywatnych VPN (ang. *Virtual Private Network*), stosowanie zabezpieczeń poczty elektronicznej oraz systemów uwierzytelniania użytkowników [2, 3]. Zabezpieczenia personalne i organizacyjne dotyczą zabezpieczenia danych medycznych przed nieostrożnością osób korzystających z systemów legalnie. Na nic się zdadzą systemy identyfikacji użytkowników, jeśli osoby korzystające z systemu będą pozostawiały identyfikatory w miejscach dostępnych dla osób trzecich. Niewielka korzyść będzie płynęła z systemu archiwizacji danych, jeżeli personel zaniedba wykonywania kopii bezpieczeństwa. Dlatego ważnym jest, aby przeszkolić osoby upoważnione do korzystania z danych medycznych. Zabezpieczenia organizacyjne obejmują opracowanie i udostępnienie scenariuszy postępowań w sytuacjach wyjątkowych (atak hakera, dostanie się do systemu złośliwego wirusa komputerowego) osobom upoważnionym do korzystania z systemu informacyjnego oraz wyszkolenie tych osób. Każde z opisanych zabezpieczeń nic nie

znaczy, gdy funkcjonuje samodzielnie. Aby medyczny system informacyjny był bezpieczny potrzebna jest integracja zabezpieczeń technicznych, właściwego poziomu wyszkolenia pracowników oraz właściwych rozwiązań organizacyjnych [2, 4].

Na temat każdej z metod ochrony danych medycznych można byłoby napisać kolejny rozdział. W tej pracy przeglądowej większa uwaga zostanie zwrócona na podstawowe metody ochrony prywatności w medycznych bazach danych.

3.1. Metody anonimizacji danych medycznych

W literaturze wymienia się cztery podstawowe metody ataków prowadzonych w stosunku do tabel baz danych medycznych [5]:

- łączenie rekordów danych,
- łączenie do atrybutów,
- łączenie do tabel,
- metody probabilistyczne.

W metodzie łączenia do rekordów i do atrybutów zakłada się, że znana jest wartość pseudoidentyfikatora (skrót PID) ofiary i poszukuje się jej danych wrażliwych. W przypadku łączenia do tabel, atakujący ustala czy w opublikowanej tabeli znajduje się rekord ofiary. Metody probabilistyczne służą zdobyciu szerszej wiedzy o ofierze na podstawie danych zanonimizowanych [5-8]. W odpowiedzi na tego rodzaju ataki wykorzystuje się metody zapewniające anonimowość osób, których dane są przechowywane w tabelach. Kilka z tych metod zostało opisanych w poniższych podrozdziałach.

3.1.1. Anonimowość metodą k-anonimizacji

Metoda k-anonimizacji jest metodą przeciwdziałającą naruszeniu prywatności pacjentów poprzez dołączenie zewnętrznych rekordów do tabeli danych. W tabeli z danymi pacjentów wyłonić można atrybuty, które stanowią pseudoidentyfikator pacjenta PID. Wszystkie rekordy bazy danych są grupowane względem poszczególnych wartości pseudo-identyfikatora [5- 8]. Zazwyczaj taki podział wyznacza grupy niezbyt liczne. Mała liczba rekordów w grupie daje małą liczbę wartości, które można przypisać do pacjenta, którego dane chce się pozyskać. Dlatego też oszust, posiadający dostęp do innej wiedzy o pacjencie, może wyizolować rekord pacjenta z grupy wskazanej przed PID. Aby zabezpieczyć dane medyczne przed ujawnieniem należy przyjąć minimalną liczność rekordów k w grupach identyfikowanych pseudo-identyfikatorem PID [5-7]. W praktyce metoda ta wymaga korekty danych dla atrybutów tworzących PID. Tab. 2. przedstawia przykładowe dane pacjentów z bazy chorób tarczycy. Identyfikator PID tworzony jest przez trzy atrybuty: płeć, miasto oraz wartość hormonu TSH pacjenta.

Tabela 2. Fragment bazy danych pacjentów, u których zdiagnozowano chorą tarczycę. W bazie umieszczono rekordy pacjentów zidentyfikowanych wg trzech PID

Pseudo-identyfikator	Płeć (atrybut a)	Miasto (atrybut b)	TSH (atrybut c)	Choroba tarczycy (atrybut d)
PID_1	a_3	b_1	c_2	d_1
PID_2	a_2	b_1	c_3	d_2
PID_3	a_1	b_2	c_1	d_1
PID_1	a_3	b_1	c_2	d_1
PID_1	a_3	b_1	c_2	d_2
PID_3	a_1	b_2	c_1	d_2
PID_3	a_1	b_2	c_1	d_2
PID_1	a_3	b_1	c_2	d_2

Źródło: Opracowanie własne

Na podstawie tab. 2. można wskazać trzy grupy wyznaczane wartością pseudo-identyfikatora. Wśród tych grup jest jedna określana identyfikatorem PID_2, którą tworzy jeden rekord. Oszust, posiadając wiedzę z innych źródeł, np. z tab. 3., może wydobyć dane wrażliwe pacjentów.

Tabela 3. Dodatkowy fragment bazy z danymi pacjentów z chorą tarczycą

Pacjent (obiekt)	Płeć (atrybut a)	Miasto (atrybut b)	TSH (atrybut c)
x_1	a_3	b_1	c_2
x_2	a_2	b_1	c_3
x_3	a_1	b_2	c_1

Źródło: Opracowanie własne

Na podstawie tab.2. oraz tab. 3. oszust jest w stanie wydobyć dane wrażliwe pacjenta x_2 w postaci zdiagnozowanej u niego choroby tarczycy d_2. Metoda k-anonimizacji nie zabezpiecza medycznej bazy danych przed ujawnieniem danych wrażliwych metodami statystycznymi np. informacji o tym, że pacjent x_3 z prawdopodobieństwem 0.(6) może mieć chorobę d_2.

3.1.2. Anonimowość metodą (X,Y)-anonimizacji

Uogólnieniem metody k-anonimizacji jest metoda (X,Y)-anonimizacji. Metoda ta polega na podziale atrybutów na dwa zbiory X oraz Y. X jest zbiorem atrybutów tworzących pseudoidentyfikator, zaś zbiór Y tworzą wartości pseudoidentyfikatora. Wymaganym jest, aby dla każdej wartości atrybutu X występowało *k* różnych wartości atrybutu Y [5, 7]. Tab. 4. przedstawia podział danych metodą (X,Y)-anonimizacji.

Tabela 4. Fragment bazy danych pacjentów, u których zdiagnozowano chorą tarczycę. Fragment bazy zawiera rekordy wybrane metodą (X,Y)-anonimizacji

Pseudo-identyfikator	Płeć (atrybut a)	Miasto (atrybut b)	TSH (atrybut c)
Y	X		
PID_1	a_2	b_1	c_3
PID_2	a_2	b_1	c_3
PID_3	a_2	b_1	c_3
PID_4	a_2	b_2	c_2
PID_5	a_2	b_2	c_2
PID_6	a_2	b_2	c_2
PID_7	a_2	b_2	c_2
PID_8	a_2	b_2	c_2

Źródło: Opracowanie własne

3.1.3. Anonimowość metodą (α,k)-anonimizacji

Metoda ta łączy w sobie metodę k-anonimizacji oraz α-deasocjacji. Według metody k-anonimizacji minimalna liczba rekordów w grupie dla wskazanej wartości PID nie może być mniejsza niż *k*. Dodatkowo, dla zadanej wartości wrażliwej *w*, prawdopodobieństwo wystąpienia jest nie większe niż α we wszystkich klasach równoważności [5, 9]. Tab.5. przedstawia warunki (0.(6),3)-anonimizacji.

Tabela 5. Fragment bazy danych pacjentów, u których zdiagnozowano chorą tarczycę. Fragment bazy zawiera rekordy wybrane metodą (0.(6),3)-anonimizacji

Pseudo-identyfikator	Płeć (atrybut a)	Miasto (atrybut b)	TSH (atrybut c)	Choroba tarczycy (atrybut d)
PID_1	a_3	b_1	c_2	d_1
PID_2	a_2	b_1	c_3	d_2
PID_3	a_1	b_2	c_1	d_1
PID_1	a_3	b_1	c_2	d_2
PID_2	a_2	b_1	c_3	d_3
PID_3	a_1	b_2	c_1	d_2
PID_3	a_1	b_2	c_1	d_3
PID_1	a_3	b_1	c_2	d_3
PID_2	a_2	b_1	c_3	d_1

Źródło: Opracowanie własne

3.1.4. Anonimowość metodą (k,e)-anonimizacji

Metodę tę wykorzystuje się do ochrony danych medycznych, które mają postać numeryczną. W metodzie (k,e)-anonimizacji rekordy muszą być podzielone na grupy zawierające przynajmniej *k* różnych wartości wrażliwych, a maksymalna różnica między tymi wartościami wrażliwymi powinna wynosić *e* [5]. Tab. 6. przedstawia podział danych metodą (5, 2.3)-anonimizacji.

Tabela 6. Fragment bazy danych pacjentów, u których zdiagnozowano chorą tarczycę. Fragment bazy zawiera rekordy wybrane metodą (5, 2.3)-anonimizacji

Pseudo-identyfikator	Płeć (atrybut a)	Miasto (atrybut b)	TSH (atrybut c)
PID_1	a_2	b_1	1,26
PID_2	a_2	b_1	3,12
PID_3	a_2	b_1	0,59
PID_4	a_2	b_1	2,01
PID_5	a_2	b_1	2,88

Źródło: Opracowanie własne

Opisane w pracy metody są podstawowymi metodami anonimizacji danych medycznych, na podstawie których rozwinęły się też inne metody anonimizacji, takie jak: l-dywersyfikacji, (X,Y)-dołączalności, (X,Y)-prywatności, LKC-prywatności, t-bliskości, ograniczonego zaufania oraz personalizowanej prywatności [5]. Metodom tym warto jest poświęcić szerszą uwagę, dlatego będą one przedmiotem dalszej pracy naukowej.

4. Podsumowanie

Rozwój elektronicznych form przechowywania, przetwarzania i przekazywania danych medycznych wpłynął nie tylko na podwyższenie jakości opieki zdrowotnej pacjentów, ale również na rozwój nowych metod pozyskiwania wiedzy z medycznych baz danych przez osoby do tego nieupoważnione. Aby zapobiec ujawnieniu danych wrażliwych, medyczne systemy informacyjne cały czas powinny być poddawane kontroli bezpieczeństwa we wszystkich strukturach systemu. Zabezpieczenia techniczne powinny być uzupełniane zabezpieczeniami fizycznymi, personalnymi oraz organizacyjnymi. Jedynie integracja właściwego zabezpieczenia technicznego, właściwego poziomu wyszkolenia pracowników i rozwiązań organizacyjnych może zminimalizować prawdopodobieństwo odkrycia wiedzy strzeżonej przez system medyczny. Spośród wymienionych kategorii zabezpieczeń najdynamiczniej rozwijają się zabezpieczenia techniczne, w obrębie których jest tworzenie kopii zapasowych danych, programy antywirusowe czy też programowe i sprzętowe systemy uwierzytelniania użytkowników. Aby spotęgować ochronę baz danych przed ujawnieniem danych pacjentów stosuje się różne algorytmy podziału baz danych na mniejsze, specyficzne tabele. Najczęściej wykorzystuje się metody k-anonimizacji, (k,e)-anonimizacji, (X,Y)-ananonimizacji oraz (α,k)-anonimizacji, które dzielą rekordy baz danych na tabele biorąc pod uwagę liczbę *k* rekordów z określonym pseudo-identyfikatorem oraz najczęściej też inny czynnik np. statystyczny (α) czy różnicowy (e). Dzięki metodom ukrywania wiedzy w medycznych bazach danych można zminimalizować skuteczność różnych form ataków na dane pacjentów.

Wymienione metody anonimizacji stanowią bazę dla bardziej zaawansowanych metod ochrony danych pacjentów, których analiza będzie przedmiotem dalszej pracy naukowej.

Literatura:

1. http://it.rsi.org.pl/dane/Nowoczesne_systemy_informatyczne_w_ochronie_zdrowia.pdf.
2. Tadeusiewicz R.*Informatyka medyczna*, Uniwersytet Marii Curie-Skłodowskiej w Lublinie, Instytut Informatyki (2011), s. 183-212.
3. Zajdel R., Kęcki E., Szczepaniak P., Kurzyński M. *Kompendium informatyki medycznej*, Alfa-medica Press (2003).
4. Borucki B. *Ochrona poufności i bezpieczeństwa medycznych danych osobowych*, Kardionet e-book, Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytet Warszawski (2009), s. 9-20.
5. Liber A. *Problemy anonimizacji dokumentow medycznych*. *Część 1. Wprowadzenie do anonimizacji danych medycznych. Zapewnienie ochrony danych wrażliwych metodami f(a)- i f(a,b)- anonimizacji*., Puls Uczelni, 1(2014), s. 13-21.
6. Samarati P.,*Protecting respondents identities in microdata release*., IEEE Transactions on Knowledge and Data Engineering, Vol. 13(6) (2001), pp. 1010-1027.
7. Samarati P.,Sweeney L.,*Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*., Technical report. SRI International (1998), pp.1-19.
8. Sweeney L. *k-Anonymity: A model for protecting privacy*, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), (2002), pp. 557-570.
9. Chi-Wing Wong R., Li J., Fu, Wai-Chee Fu A., Wang K., *(α,k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing*. SIGKDD International Conference on Knowledge Discovery and Data Mining, (2006), pp. 754-759.

Bezpieczeństwo w medycznych systemach informacyjnych

Streszczenie

Największym wyzwaniem współczesnych medycznych systemów informacyjnych jest ochrona danych wrażliwych, które gromadzą, przetwarzają i przekazują między placówkami leczniczymi a pacjentami. Rozwój tych systemów wiąże się z pojawianiem nowych zagrożeń, dlatego też poszukuje się coraz doskonalszych metod zapewniających anonimowość danych medycznych.

W pracy przedstawione zostały podstawowe formy zagrożeń dla bezpieczeństwa systemów informacyjnych, przykładowy sposób oceny tych zagrożeń oraz metody walki z nimi. Szczególną uwagę zwrócono na podstawowe metody anonimizacji danych pacjentów, na których bazują wszystkie bardziej zaawansowane. Opisane i poparte przykładem zostały metody k-anonimizacji, (X,Y)-anonimizacji, (k,e)-anonimizacji oraz (k, α)-anonimizacji. Stanowić one będą punkt wyjścia dla dalszych prac naukowych nad algorytmami ukrywania i odkrywania wiedzy z medycznych bazach danych.

Słowa kluczowe: bezpieczeństwo medycznych systemów informacyjnych, anonimizacja

Safety of Healthcare Information Systems

Abstract

The greatest challenge of modern medical information systems is the protection of sensitive patient data which they are collected, processed and transferred between treatment centers and patients. The development of these systems involves the emergence of new threats. Therefore, we are looking for better and better methods to ensure the anonymity of medical data.

The paper presents the basic forms of threats to the security of information systems, example of how these threats are rated and fighting methods with them. Particular attention has been paid to the basic methods of patient anonymisation. They are a base for more advanced ones. The k-anonimization, (X,Y)-anonimization (k,e)-anonimization and (k,α)- anonimization methods have been described and supported by the example. They will be the starting point for further research into the algorithms of hiding and discovering knowledge from medical databases.

Keywords: safety of healthcare information systems, anonymisation

Anna Kasperczuk[1], Agnieszka Dardzińska-Głębocka[2]

Drzewa decyzyjne jako narzędzie wspomagające eksplorację wiedzy z medycznych systemów informacyjnych

1. Wstęp

Obecnie ogromnym wyzwaniem staje się efektywna analiza przechowywanych danych, także medycznych. Generuje to zapotrzebowanie na nowe metody i narzędzia informatyczne wspomagające odkrywanie wiedzy z gromadzonych danych. Szybki rozwój technologii generowania, gromadzenia oraz przetwarzania danych generuje konieczność analizowania zgromadzonych zbiorów. Odpowiedzią na potrzebę zaawansowanej oraz automatycznej analizy danych, przechowywania w bazach i hurtowniach danych jest technologia eksploracji wiedzy (ang. *data mining*). Zadaniem omawianych metod jest automatyczne odkrywanie nietrywialnych oraz dotychczas nieznanych zależności i wzorców. Jednym z istotnych zagadnień związanych z tematyką eksploracji danych jest indukcja drzew decyzyjnych [1, 2]. W niniejszej pracy opisane zostanie zastosowanie drzew decyzyjnych do wydobywania wiedzy z medycznej bazy danych.

2. Eksploracja wiedzy

Eksploracja danych (ang. *data mining*) jest stosunkowo nową dyscypliną nauki, której celem jest poznanie analizowanych procesów oraz generowanych przez nie danych. Mówiąc o eksploracji danych, należy odwołać się do analizy rzeczywistych, dużych zbiorów danych obserwacyjnych, badanych w celu generowania, interesujących z punktu widzenia oczekiwań użytkowników, rezultatów [2].

Istnieje wiele definicji eksploracji wiedzy z baz danych. Najbardziej znane są dwie, niżej przedstawione.

Pierwsza jest następująca: „Eksploracja danych jest analizą (często ogromnych) zbiorów danych obserwacyjnych, w celu znalezienia nieoczekiwanych związków i podsumowania danych w oryginalny sposób, tak aby były zarówno zrozumiałe, jak i przydatne dla ich właściciela" [3].

„Eksploracja danych jest międzydyscyplinarną dziedziną, łączącą techniki uczenia maszynowego, rozpoznawania wzorców, statystyki, baz danych i wizualizacji w celu uzyskiwania informacji z dużych baz danych" [3].

[1] a.kasperczuk@pb.edu.pl, Politechnika Białostocka, Wydział Mechaniczny, Zakład Biocybernetyki i Inżynierii Biomedycznej
[2] a.dardzinska@pb.edu.pl, Politechnika Białostocka, Wydział Mechaniczny, Zakład Biocybernetyki i Inżynierii Biomedycznej

2.1. Drzewa decyzyjne

Jedną z najważniejszych metod eksploracji wiedzy, mających ogromne znaczenie praktyczne, jest klasyfikacja. Celem klasyfikacji danych jest zbudowanie klasyfikatora, czyli modelu, który w oparciu o wartości pozostałych atrybutów (deskryptorów), przydziela każdej próbce testowej wartość atrybutu danej klasy [4].

Wśród metod klasyfikacji najbardziej popularną jest metoda wykorzystująca algorytm indukcji drzew decyzyjnych. Jest ona szczególnie atrakcyjna ze względu na intuicyjny, zrozumiały dla człowieka sposób reprezentacji wiedzy [5].

Początkowo drzewa decyzyjne (ang. *decision trees*) pojawiły się w latach 60. W obszarach badań dotyczących psychologii oraz socjologii. W informatyce po raz pierwszy znalazły swoje zastosowanie w pracach w latach 80. [6, 7].

W porównaniu z pozostałymi metodami klasyfikacji, drzewa decyzji mogą być konstruowane stosunkowo szybko. Podstawową ich zaletą jest czytelna reprezentacja wiedzy, możliwość zastosowania danych wielowymiarowych, skalowalność przy wykorzystaniu dużych zbiorów danych. Ponadto dokładność tej metody jest porównywalna z dokładnością innych metod klasyfikacji. Natomiast zasadniczą wadą omawianej metody jest duża wrażliwość na brakujące wartości atrybutów, bowiem u ich podstaw istnieje niewyrażone jawnie założenie o pełnej dostępności informacji zgromadzonych w bazie przypadkach. Do wad należy zaliczyć również brak możliwości wychwycenia korelacji pomiędzy atrybutami [8].

W metodzie indukcji drzew decyzyjnych wynikiem ich działania jest skierowany, spójny graf posiadający drzewiastą strukturę. Korzysta on z graficznej struktury danych oraz przedstawia ich możliwe konsekwencje, pomagając w podejmowaniu decyzji. Otrzymana w ten sposób struktura jest zbiorem węzłów decyzyjnych połączonych za pomocą „gałęzi", które rozchodzą się w dół od „korzenia" do kończących drzewo „liści".

Drzewa klasyfikacyjne wykorzystywane są do określania przynależności obiektów do klasy jakościowej zmiennej zależnej. Odbywa się to na podstawie pomiarów jednej lub więcej zmiennych predykcyjnych. Drzewo klasyfikacyjne przedstawia proces podziału zbioru obiektów na klasy jednorodne. Podział odbywa się w oparciu o wartości cech obiektów, liście odpowiadają klasom, do których należą obiekty, natomiast krawędzie drzewa reprezentują wartości cech, na podstawie których dokonano podziału [8].

Proces tworzenia drzewa decyzyjnego polega na rekurencyjnym podziale zbioru uczącego na podzbiory, co odbywa się do momentu uzyskania ich jednorodności ze względu na przynależność obiektów do klas. Celem jest utworzenie drzewa o jak najmniejszej liczbie węzłów, a w konsekwencji utworzenie jak najprostszych reguł klasyfikacyjnych [9].

Algorytm tworzenia drzewa decyzyjnego można zapisać następująco [4, 10]:

- Dla danego zbioru obiektów należy sprawdzić, czy należą one do tej samej klasy (jeżeli należą – zakończyć postępowanie, w przypadku, gdy nie należą –

rozważyć wszystkie możliwe podziały danego zbioru na możliwie najbardziej jednorodne podzbiory),
- Ocenić jakość każdego z tych podzbiorów zgodnie z uprzednio przyjętym kryterium i wybrać optymalny,
- Podzielić zbiór obiektów w wybrany sposób,
- Etapy wykonać dla każdego z podzbiorów.

2.1.1. Algorytm ID3

Jedną z najwcześniejszych propozycji realizacji systemów uczących oraz pozyskiwania wiedzy prezentowanej w postaci drzewa decyzyjnego jest algorytm ID3, który został opracowany przez Quinlana [7]. Generuje on drzewo decyzyjne na podstawie szeregu przypadków jednostkowych. Drzewo decyzyjne stanowi strukturalny zapis wydobytej wiedzy, pozwalającym, na podstawie wartości pewnych cech, przypisać konkretne wartości cechom decyzyjnym [10].

W celu wygenerowania za pomocą algorytmu ID3 drzewa decyzyjnego konieczny jest stosunkowo duży zbiór obiektów opisujących dany problem. Każdy obiekt ze zbioru przyjmuje określoną wartość dla każdego atrybutu z listy atrybutów warunkowych oraz atrybutu decyzyjnego. Każdy atrybut opisujący dany obiekt przyjmuje jedną wartość z istniejącej i opisanej dziedziny danego atrybutu. Tak zdefiniowany zbiór obiektów jest zbiorem uczącym. W momencie, gdy zbudowana zostaje lista atrybutów wraz z listami dostępnych wartości oraz zbiór uczący, możliwe jest rozpoczęcie budowy drzewa decyzyjnego [2].

Algorytm ID3:

Wejście: zbiór treningowy *D*, zbiór atrybutów warunkowych *A*, metoda wyboru punktu podziału *SS*.
Wyjście: drzewo decyzyjne ukorzenione w wierzchołku *N*.
procedure *BuildTree*(*D, A, SS*):
utwórz wierzchołek drzewa decyzyjnego *N*;
if wszystkie rekordy zbioru *D* należą do tej samej klasy *C* **then**
return wierzchołek *N* jako liść drzewa decyzyjnego i przypisz danemu wierzchołkowi etykietę klasy *C;*
end if
if lista_atrybutów *A* jest pusta **then**
return wierzchołek *N* jako ilość drzewa i przypisz temu wierzchołkowi etykietę klasy dominującej w zbiorze treningowym *D*;
end if
zastosuj metodę *SS* w celu wybrania *atrybutu-podziałowego* ze zbioru *A*;
przypisz wierzchołkowi *N* etykietę *atrybutu-podziałowego*;
for all wartości a_i atrybutu-podziałowego **do**
$S_i \leftarrow$ zbiór rekordów D o wartości *atrybutu-podziałowego* $= a_i$;
$N_i \leftarrow$ BuildTree(S_i, (lista_atrybutów A) – (atrybut-podziałowy), SS);
utwórz krawędź z *N* do N_i etykietowaną wartością a_i;
end for
return wierzchołek *N*.

Niewątpliwą wadą algorytmu ID3 jest jego działanie jedynie na atrybutach o wartościach numerycznych, przy niepełnych danych algorytm nie działa. Dodatkowo finalne drzewa są zbyt dopasowane do danych zaś miara przyrostu informacji faworyzuje cechy o dużej liczbie wartości. Opisywany algorytm cechuje brak odporności na zjawisko *overfittingu*, bowiem nie radzi on sobie z danymi zaburzającymi ogólną ich informację, co może prowadzić do wysokiego współczynnika błędów na danych testowych [2, 3]. Problemy te zostały wyeliminowane po wprowadzeniu kolejnych wersji algorytmu ID3 (m.in. C4.5).

2.1.2. Algorytm C4.5

Algorytm C4.5 jest jednym z dwóch najpopularniejszych algorytmów stosowanych w praktyce. Algorytm ten jest w istocie rozszerzeniem algorytmu ID3.

Algorytm C4.5 rekurencyjnie przechodzi przez wszystkie węzły, wybierając możliwy podział, dopóki dalsze podziały są możliwe. Dla zmiennych jakościowych algorytm ten z definicji tworzy osobne gałęzie dla każdej wartości atrybutu jakościowego. Może to być skutkiem większego rozgałęzienia drzewa niż jest pożądane, bowiem niektóre wartości mogą występować rzadko oraz być w naturalny sposób powiązane z innymi wartościami [3, 11].

Dla atrybutów ciągłych omawiany algorytm rozpatruje wszystkie możliwe podziały na dwa podzbiory, zdeterminowane punktem podziału *w*. W przeciwieństwie do atrybutów dyskretnych, atrybuty ciągłe mogą pojawiać się na wielu poziomach tej samej gałęzi drzewa decyzyjnego. Dla każdego z możliwych podziałów, ocenia się jego jakość poprzez miarę wartości względnego zysku informacji. Wybierana jest opcja, dająca maksymalny zysk informacji [2, 3].

Algorytm C4.5, oprócz metody indukcji drzew decyzyjnych, umożliwia przekształcanie drzewa w zbiór reguł. Reguły są tu traktowane jako odmienny od drzewa model klasyfikacji, ponieważ nie są one wierną reprezentacją drzewa.

3. Analiza wyników

Dane medyczne mogą być drążone w celu wydobycia z nich reguł łączących diagnozę z symptomami chorobowymi. Reguły takie mogą być wykorzystane do automatycznej klasyfikacji (odnalezienia schorzenia) nowych, dotychczas niezdiagnozowanych pacjentów jedynie w oparciu o występujące u nich symptomy. Mogą być również użyte do odnalezienia ukrytej zależności pomiędzy stanem medycznym a predykatami wpływającymi na opisywany stan medyczny. Medyczne bazy danych zawierają informacje o pacjencie, m.in. zrejestrowane podczas wizyty u lekarza lub pobytów w szpitalu, wyniki testów diagnostycznych [11].

W pracy zaprezentowano wykorzystanie metod drzew decyzyjnych w eksploracji wiedzy z medycznego systemu informacyjnego. Analizie poddano rzeczywiste dane dotyczące 286 pacjentek z diagnozą nowotworu piersi. Rak piersi jest najczęściej diagnozowanym nowotworem złośliwym wśród kobiet. Wiedza o przyczynach raka piersi może w znacznym stopniu zwiększyć prawdopodobieństwo uniknięcia choroby [12]. Dane pozyskano z repozytorium UCI [13]. Metody klasyfikacji mogą w sposób znaczący przyspieszyć postawienie diagnozy oraz zwiększyć szanse na wyzdrowienie, a w dalszej perspektywie uniknięcie choroby.

Do przeprowadzenia procesu klasyfikacji użyto oprogramowania WEKA, będącym oprogramowaniem z zakresu uczenia maszynowego oraz pozyskiwania wiedzy stworzonym w środowisku języka programowania JAVA. Program WEKA, stworzony na Uniwersytecie w Waikato w Nowej Zelandii jest zestawem algorytmów wykorzystywanych do realizacji zadań *data mining*. Założeniem pracy był wybór technik eksploracji danych (drzew decyzyjnych) pozwalających przewidzieć ryzyko pojawienia się nawrotu lub przerzutu w czasie 5 lat od momentu rozpoznania choroby. Uzyskane modele drzew decyzyjnych oceniono pod względem jakości za pomocą analizy statystyk modelu. Atrybuty opisywanego zbioru danych zestawiono w tabeli 1.

Tabela 1. Zestawienie atrybutów bazy danych raka piersi

Atrybut	Wartość
Wiek	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
Menopauza	lt40, ge40, premeno
Wielkość guza	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
Inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39
Node-caps	tak/nie
Stopień złośliwości	1, 2, 3
Pierś	lewa/prawa
Część piersi	lewa górna, lewa dolna, prawa górna, prawa dolna, centralna
Radioterapia	tak/nie
Klasyfikujący	wystąpienie nawrotu choroby/brak nawrotu choroby

Opis atrybutów:

- Wiek: wiek pacjentki w momencie diagnozy,
- Menopauza: stan pacjentki w czasie diagnozy (lt40 – wiek przed 40. rokiem życia, ge40 – wiek po 40. roku życia, premeno – stan przedmenopauzalny, około 35. roku życia,
- Wielkość guza: rozmiar guza w milimetrach,
- Inv-nodes: zakres (0-39) pachowych węzłów chłonnych potwierdzających nowotwór piersi w badaniu histopatologicznym,
- Node-caps: określa przerzuty nowotworu do pachowych węzłów chłonnych.
- Stopień złośliwości: histopatologiczny stopień złośliwości nowotworu w skali 1-3 (1 – prawdopodobnie zawiera złośliwe komórki nowotworowe; stopień 2 – zawiera charakterystyczne komórki złośliwe; stopień 3 – zaawansowany nowotwór złośliwy),
- Pierś: pierś (lewa/prawa) zajęta przez nowotwór,
- Część piersi: określa segment piersi objęty nowotworem,
- Radioterapia: czy pacjentka była poddana radioterapii,

- Atrybut klasyfikujący: dotyczy stanu po zakończeniu leczenia. Wartość „tak" oznacza, że nastąpił nawrót choroby, wartość „nie" stwierdza brak nawrotu choroby.

Oprogramowanie WEKA wykorzystuje w eksploracji danych następujące mierniki oceny jakości modelu:

- *TP Rate* pokazuje jaki odsetek obserwacji z danej klasy jest poprawnie klasyfikowany przez model, liczy przypadki *true positiv,*.
- *FP Rate* opisuje, jaka część obserwacji nienależących do danej klasy została błędnie do niej zaklasyfikowana – *false positive,*
- *Precisio*n jest miernikiem precyzji przyporządkowania danej obserwacji do odpowiadającej klasy,
- Kategoria *Recall* wskazuje poprawne pokrycie danej klasy,
- Miara *F-Measure* to ogólny wskaźnik jakości modelu,
- Statystyka *Kappa* jest miernikiem zgodności między proponowanym przydziałem instancji do klasy a stanem faktycznym, co stanowi o ogólnej trafności modelu [35].

W pracy użyto dwóch klasyfikatorów: J48 oraz Random Tree. W przypadku obu testowanych algorytmów zastosowano metodę walidacji krzyżowej (ang. *cross validation*). Jest to metoda statystyczna, polegająca na podziale badanej próby statystycznej na dwa podzbiory: zbiór uczący i testowy. Analizy przeprowadzane są na zbiorze uczącym, podczas gdy zbiór testowy służy do potwierdzenia wiarygodności uzyskanych wyników.

3.1. Klasyfikator J48

Metoda J48 używa algorytmu C4.5 do wygenerowania drzewa decyzyjnego. Algorytm C4.5 dzieli pierwotny zestaw danych względem każdego atrybutu.

Pierwszym etapem procesu modelowania była ocena jego najważniejszych statystyk. Pozwoliło to na wstępne określenie trafności wygenerowanych reguł decyzyjnych oraz dostarczyło informacji o rodzaju ewentualnych błędów. Tabela 2. pokazuje zestawienie podstawowych statystyk określających ocenę jakości modelu J48.

Tabela 2. Statystyki modelu J48

Czynnik	Wartość
Poprawnie sklasyfikowane przypadki	76%
Błędnie sklasyfikowane przypadki	24%
Statystyka Kappa	0,2899

Odsetek poprawnie sklasyfikowanych atrybutów przez drzewo decyzyjne wynosi 76%, Jest on wysokim, prawidłowym wynikiem oraz wskazuje na dobrą jakość generowanego drzewa decyzyjnego. Wskaźnik Statystyki Kappa jest względnie niski, co oznacza, że jest blisko 30% obserwacji, z którymi klasyfikator losowy sobie nie poradził.

W tabeli 3. przedstawiono zestaw mierników jakości modelu J48. Dla atrybutu klasyfikacyjnego charakteryzującego brak nawrotu choroby wartość miernika *true positive* okazała się być bardzo wysoka (96%), co jest zadowalającym wynikiem.

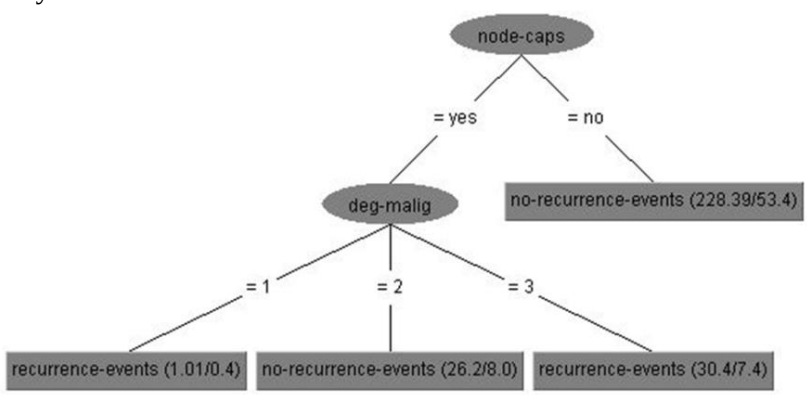
Stopa *false positive* jest niższa (73%), co może świadczyć o dostatecznej jakości wygenerowanego modelu.

Miara *F-Measure*, szacująca ogólną jakość modelu, posiada wysoką wartość (85%) i wskazuje na prawidłowo zbudowany proces medyczny.

Tabela 3. Mierniki jakości modelu J48

Klasa	TP Rate	FP Rate	Precision	Recall	F-Measure
Brak nawrotu choroby	0,965	0,729	0,758	0,965	0,849
Nawrót choroby	0,271	0,035	0,767	0,271	0,400

Wygenerowane drzewo decyzyjne J48 posiada 4 liście. W korzeniu drzewa znajduje się atrybut dotyczący przerzutów nowotworu do węzłów chłonnych. Test przeprowadzony na tym atrybucie podzielił zbiór na dwa podzbiory: według stopnia złośliwości nowotworu oraz braku nawrotu choroby. Otrzymane poddrzewa są niesymetryczne. Poddrzewo dotyczące stopnia złośliwości nowotworu jest głębsze ze względu na jego trzystopniową klasyfikację. Pełną postać drzewa decyzyjnego pokazano na rysunku 1.



Rysunek 1. Drzewo J48

3.2. Klasyfikator *Random Tree*

Metoda klasyfikacji drzew decyzyjnych *Random Tree* jest algorytmem łączącym drzewa decyzyjne i metody lasów losowych (*Random Forest*). Model samego drzewa decyzyjnego jest wystarczający wówczas, gdy badany jest mniejszy zakres zmiennych [14].

Tabela 4. Statystyki modelu Random Tree

Czynnik	Wartość
Poprawnie sklasyfikowane przypadki	98%
Błędnie sklasyfikowane przypadki	2,1%
Statystyka Kappa	0,9491

Tabela 4. przedstawia statystyki oceny jakości modelowanego procesu. Uzyskana wartość poprawnie sklasyfikowanych atrybutów jest bardzo wysoka (98%), co

sugeruje bardzo dobrą jakość drzewa. Ocena dobrej trafności modelu jest potwierdzona wysokim wskaźnikiem statystki Kappa. Fakt ten informuje o wysokiej, 95% liczbie klasyfikowanych obserwacji.

Tabela 5. Mierniki jakości modelu Random Tree

Klasa	TP Rate	FP Rate	Precision	Recall	F-Measure
Brak nawrotu choroby	0,995	0,059	0,976	0,995	0,985
Nawrót choroby	0,941	0,005	0,988	0,941	0,964

Tabela 5. zawiera zestaw mierników jakości modelowanego procesu. Atrybut stanowiący brak nawrotu choroby w kategorii miernika *true positive* posiada wartość 99%. Miernik *false positive* stanowi wartość absolutnie odmienną bliską 1%. Na podstawie tych dwóch wartości możemy wnioskować bardzo dobre dopasowanie modelowanego procesu. Pozostałe wskaźniki (*Precision, Recall* oraz *F-Measure*) również dowodzą trafności powyższego wniosku.

Otrzymane w wyniku klasyfikacji drzewo decyzyjne charakteryzuje się dużą liczbą poziomów. Korzeń drzewa zawiera atrybut odpowiadający pachowym węzłom chłonnym, który w badaniu histopatologicznym potwierdza nowotwór piersi. Test przeprowadzony na tym atrybucie podzielił zbiór na 4 podzbiory: wiek, pierś, stopień złośliwości, atrybut klasyfikujący. Ostatecznymi liśćmi drzewa są atrybuty klasyfikujące dotyczące braku nawrotu choroby po 5 latach od jej rozpoznania.

Głównym celem generowania drzewa decyzyjnego było sprawdzenie ilości przypadków z nawrotem nowotworu piersi w ciągu 5 lat od momentu rozpoznania choroby. Analizując uzyskane drzewa można stwierdzić, iż pojawienie się przerzutów nowotworu do pachowych węzłów chłonnych jest jednoznaczne z określeniem stopnia złośliwości nowotworu. Najwyższy stopień złośliwości nowotworu skorelowany jest z pojawieniem się nawrotu raka piersi. Pierwszy stopień złośliwości nowotworu charakteryzuje najniższą liczbą przypadków z nawrotem choroby. Brak przerzutów do pachowych węzłów chłonnych oznacza w większości przypadków remisję choroby.

4. Podsumowanie

Odkrywanie wiedzy w bazach danych jest prężnie rozwijającą się dziedziną, której szybki rozwój związany jest z rosnącą liczbą baz danych oraz wielkością gromadzonych w nich informacji. Zaistniała potrzeba opracowania nowych metod analizy danych. Odpowiedzią na to zapotrzebowanie okazały się metody rozwijane w ramach odkrywania wiedzy w bazach danych. Jedną z bardziej popularnych metod jest indukcja drzew decyzyjnych. Opisane metody można z powodzeniem stosować w tworzeniu medycznych modeli diagnostycznych, które mogą w przyszłości wspomagać pracę lekarzy. Warto przy tym nadmienić, że proponowane metody są jedynie narzędziem wspomagającym trafną diagnostykę schorzeń. Każdy zaś eksperyment powinno się konsultować z ekspertami danej dziedziny, w tym przypadku z lekarzami.

5. Uwagi ogólne

Badania przeprowadzono w ramach projektu MB/MW/8/2016 i sfinansowano ze środków MNiSW.

Literatura

1. Dardzinska A. *Action Rules Mining*. Springer, (2013).
2. Larose D.T. *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa (2013).
3. Hand D., Mannila H., Smyth P. *Eksploracja danych. Wydawnictwa Naukowo – Techniczne*, Warszawa (2005), s. 35-61, s. 91-127,s. 181-201.
4. Krętowski M. *Obliczenia ewolucyjne w eksploracji danych. Globalna indukcja drzew decyzyjnych*, Wydawnictwo Politechniki Białostockiej, Białystok (2008), s. 7-23, s. 27-51.
5. Morzy T. *Eksploracja danych*, Wydawnictwo Naukowe PWN, Warszawa (2013), s. 10-125, s. 196-325.
6. Breiman L., Friedman J.H., Olshen R.A., Stone C.J. *Classification and Regression Trees*, Wadsworth International Group, Belmont, (1984).
7. Quinlan J.R. *Introduction of decision trees. Machine Learning*, Kluwer Academic Publishers, (1986), s. 81-106.
8. Cios K., Pedrycz W., Swiniarski R., Kudrycz L. *Data mining. A Knowledge Discovery Approach*, Springer, (2007).
9. Misztal M. *Wykorzystanie drzew klasyfikacyjnych do wspomagania procesów podejmowania decyzji.* [w:] Zastosowania statystyki i data mining w badaniach naukowych. StatSoft, Kraków (2012).
10. Cios K., Pedrycz W., Swiniarski R., Kudrycz L. *Data mining. A Knowledge Discovery Approach*, Springer (2007).
11. Cabena P., Hadjinian P., Stadler R., Zanassi A. *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, Upper Saddle River, NJ (1998).
12. Jenzach M., Jordan P. Rak piersi - rozwój wczesnej diagnostyki nieinwazyjnej w celu ograniczenia umieralności i złagodzenia konsekwencji psychologicznych choroby, [w:] Psychoonkologia 2 (2013), s. 35-49.
13. https://archive.ics.uci.edu/
14. Pfahringer B. *Random model trees: an effective and scalable regression method* [w:] Working Paper 03/2010, June (2010).

Drzewa decyzyjne jako narzędzie wspomagające eksplorację wiedzy z medycznych systemów informacyjnych

Streszczenie

Intensywny rozwój technologii generowania, gromadzenia, przetwarzania danych oraz upowszechnienie systemów informatycznych wiąże się z powstaniem konieczności analizowania zgromadzonych zbiorów. Odpowiedzią na potrzebę analizy danych, przechowywania w bazach danych jest technologia eksploracji wiedzy (ang. data mining). Jednym z istotnych zagadnień związanych z tematyką eksploracja danych jest klasyfikacja, czyli metoda analizy danych, której celem jest przypisanie danego obiektu do jednej z przedefiniowanych klas w oparciu o zbiór wartości atrybutów opisujących badany obiekt. W pracy przedstawiono zastosowanie metod klasyfikacji do analizy medycznej bazy danych.

Słowa kluczowe: eksploracja wiedzy, drzewa decyzyjne, J48, ID3

Decision trees as a tool to support knowledge discovery from medical information systems

Abstract

Intensive development of technologies for generating, collecting, data processing and disseminating information systems involves the need to analyse the collected data. Data mining techniques are a response to the need for an advanced and automatic analysis of data that are stored in databases and data warehouses. The task of these methods is non-trivial and automatic discovery of previously unknown relationships and patterns. One of the important issues related to the topic of data mining is a classification, which is the method of data analysis. The purpose of classification, based on a set of attributes describing the test object, is to assign an object to one of the predefined classes. In this work, we show how we can use this methods in medical data base.

Keywords: knowledge discovery, decision trees, J48, ID3

Marlena Robakowska[1],Anna Tyrańska-Fobke[2], Piotr Robakowski[3], Daniel Ślęzak[4], Piotr Holajn[4]

Lean Six Sigma metodami poprawy efektywności w opiece medycznej

Lean Six-Sigma to improve efficiency in medical care

Jeżeli nie mierzymy, to nic nie wiemy,
Jeżeli nie wiemy, to nie możemy działać,
Jeżeli nie działamy to narażamy się na straty
dr Mikel J. Harry

1. Wprowadzenie

W celu wzrostu poprawy wydajności, a tym samym możliwości generowania zysków, w większości firm, z dowolnej branży, idąc w ślad za strategiami koncernów, jedną z podstawowych myśli przewodnich, na których w najbliższych latach należy skupić wysiłek organizacyjny, jest jakość [1]. Standard jakości winien być głównym założeniem rozwoju. O ile tzw. trójkąt ekonomiczny składający się z czasu, kosztów i jakości do tej pory przechylał się w stronę „twardych" kosztów, o tyle obecnie kadra kierownicza stara się koncentrować wysiłki na celach jakościowych. Istnieje kilka podstawowych celów działalności do jakich firma może dążyć, czas życia (utrzymanie na rynku), udział w rynku czy też zwiększenie opłacalności produkcji/usług może być jednym z nich, jednak założeniem każdej z działalności jest generowanie zysku. O ile czas prostej redukcji kosztów powoli dobiega końca, o tyle stawia się na poprawę jakości. Wszelkie działania staramy się koncentrować na możliwościach zarządzania bieżącym zasobem. Znaną od wielu lat na świecie jest koncepcja zarządzania Six Sigma i Lean. Obie koncepcje to zwrot w stronę nabywcy wraz z nadzorem nad „błędami" w podejmowanych decyzjach w celu dążenia do ich zredukowania [2].

W literaturze światowej spotykamy się z opinią dotyczącą zastosowania usprawnień procesów poprzez zastosowanie metod globalnego systemu zarządzania jakim jest metoda Lean lub Six Sigma bądź też metoda hybrydowa łącząca cechy obu systemów. W związku z celem pracy, jakim jest pokazanie możliwości obu metod, o ile metoda Six Sigma sprawdza się w produkcji a metoda Lean w miękkim podejściu do

[1]mrobakowska@gumed.edu.pl, Zakład Zdrowia Publicznego i Medycyny Społecznej, Wydział Nauk o Zdrowiu z Oddziałem Pielęgniarstwa i Instytutem Medycyny Morskiej i Tropikalnej, Gdański Uniwersytet Medyczny
[2]II Zakład Radiologii, Wydział Nauk o Zdrowiu z Oddziałem Pielęgniarstwa i Instytutem Medycyny Morskiej i Tropikalnej, Gdański Uniwersytet Medyczny
[3]Wydział Nauk Społecznych, Uniwersytet Gdański
[4]Katedra i Klinika Medycyny Ratunkowej, Wydział Nauk o Zdrowiu z Oddziałem Pielęgniarstwa i Instytutem Medycyny Morskiej i Tropikalnej, Gdański Uniwersytet Medyczny

zarządzania, o tyle połączenie obu w opiece medycznej może znacznie poprawić wydajność pracy poprzez zmniejszenie długości pobytu pacjenta na oddziale i/lub obniżkę kosztów funkcjonowania jednostek świadczących opiekę, bez wpływu na wielkość zasobu organizacji.

2. Założenia metod

2.1. 2.1. Six Sigma

Metoda Six Sigma opiera się na działaniu polegającym na podejmowaniu decyzji w oparciu o dane z analiz wielokierunkowych w celu dążenia do poprawy jakości, dążąc w kierunku perfekcjonizmu oraz zapobieganiu występowaniu błędów. Dobra analiza wielowątkowa jest w stanie zwrócić uwagę na możliwe błędy jeszcze przed ich wystąpieniem.

Sama metoda została zastosowana po raz pierwszy w latach osiemdziesiątych ubiegłego wieku przez B. Smitha i B. Galvina w firmie Motorola. Rosnąca konkurentcyjność tanich wyrobów japońskich, spowodowała, że firmy zza oceanu zaczęły szukać nowych, lepszych i bardziej efektywnych sposobów wzrostu jakości produktów, zgodnie z załażeniem ekonomicznego wpływu trzech czynników: jakości, kosztu i czasu. Firmy motoryzacyjne, takie jak General Motors, Ford i Chrysler podjęły wspólne wyzwanie by zastosować monitorowanie procesów poprzez Statystyczną Kontrolę Procesu (SPC – Statistical Process Control). Po tym czasie, sama już Motorola, dzięki pracy dużego profesjonalnego zespołu ekspertów, stworzyła system ciągłej poprawy jakości zwany właśnie "Six Sigma Initiative" [3, 4]. Firma, sama w sobie, uznawała iż warto stawiać na jakość, jednak wcale nie ma potrzeby by podniesienie jakości generowało koszt jednocześnie generując zyski. Monitorowanie kontroli jakości oraz permanentna jej poprawa była założeniem ciągłego doskonalenia się samej firmy. Sam proces zaś, dzięki monitorowaniu i zapobieganiu niezgodnościom mógł zapobiegać generowaniu strat wynikających z błędów produkcyjnych [5].

Strategia Six Sigma definiuje jakość usługi czy produktu jako wartość. Czyli dwubiegunowe jej widzenia będące zyskiem dla przedsiębiorstwa i spełnieniem oczekiwań dla nabywcy usługi/produktu. Zgodnie z takim założeniem jakość i wartość jest korzyścią ekonomiczną dla konsumenta (stosunek wartości do ceny i jakości) i dla producenta/usługodawcy – optymalny zysk [4]. Jakość i wartość może również być wyznaczana tzw. dostępnością czyli momentem w którym sprzedajemy i momentem zakupu – jak najbardziej zoptymalizowanym dla obu stron.

Jako sama metoda działa w kierunku stałej redukcji kosztów błędów, czyli poprawie jakości bądź też niedopuszczeniu do jej obniżenia [4]. Głębsze przyjrzenie się redukcji wskazuje na dwie strategie – strategię krótkoterminową związaną z redukcją/ograniczeniem błędów oraz strategią długoterminową opartą na ciągłym doskonaleniu całego systemu.

Ograniczenia kosztów dotyczące nadzoru nad produkcją i produktem może generować redukcję kosztów wytwórczych. Przedsiębiorstwa produkcyjne lwią część kosztów przeznaczać musza na poprawę błędów procesów produkcji, a skrócenie czasu wykrycia możliwego błędu obniża koszty całkowite, zapobieżenie zaś jemu jeszcze

bardziej wpływa na obniżenie kosztochłonności produkcji. Jak widać, zgodnie z zasadą Six Sigma zapobieganie błędom jest podstawą poprawy jakości i obniżania kosztochłonności produkcji. Motorola osiągnęła już w kilka lat po wdrożeniu metody redukcję kosztów niskiej jakości z ok. 40% wartości sprzedanej do około 1%, natomiast analiza jakości wyrobów wykazała jedynie 3,4 błędu na jeden milion możliwości popełnienia [4,5]. Wszystkie te działania i czynności opierają się na zasobach wewnętrznych przedsiębiorstwa, na kierownikach średniego i wyższego szczebla. Kierownik szczebla średniego zajmuje się projektem analiz (mierzalności efektywności), kierownik szczebla wyższego koordynuje projekt poprzez nadzór nad stałym jego połączeniem z celami i jego misją. A więc z założonymi celami jakości i jej poprawy, analizą procesu i czynników mających na niego wpływ wraz z ich pomiarem, kontrolą poprawności założeń projektu i wyciąganiem wniosków czyli propozycją zmian organizacyjnych [6].

W skrócie więc metodę tą możemy opisać jako metodę opartą na pozyskaniu ogromnej liczby danych dla wykorzystania ich do osiągania możliwej maksymalnej jakości. Dzięki temu pozwalamy na wykrycie i całkowite poznanie obecnych i ewentualnych, przyszłych błędów. Badamy wszystkie procesy istniejące w przedsiębiorstwie. Metoda ta ma też kilka zasad, które opisać możemy następująco: aktywne zarządzanie procesem ze szczególnym uwzględnieniem jego słabych punktów, koncentracja na kliencie, dane liczbowe, szeroka współpraca na każdym poziomie, dążenie do perfekcji. Jednocześnie z zasad metody wynikają jej możliwe cele tj. ograniczanie zmienności, satysfakcja klienta, zmiany obniżające budżet błędów/ napraw, poprawa wyników przedsiębiorstwa.

Jako sposób poprawy jakości, Six Sigma opiera się na metodzie wspomagającej zarządzanie jaką jest DMAIC (akronim od pierwszych liter angielskich słów Define – Measure – Analyze – Improve – Control) – cyklem doskonalenia procesu [4,8].

Tabela 1. Cykl doskonalenia procesu metodą DMAIC

DEFINIOWANIE	POMIAR	ANALIZA PROBLEMU	ULEPSZENIE	KONTROLA
Definiuj problem	Ważność danych	Przyczyna problemu	Proste rozwiązania?	Rekomendacje
Zakres problemu	Celność i dokładność danych	Weryfikacja przyczyny	Weryfikacja zaproponowanych rozwiązań	Wsparcie rozwiązań
Miara ważności	Stratyfikacja danych	Miejsce skupienia wysiłków	Pilotowanie wyborów rozwiązań	Plan wprowadzenia
Interesariusze	Sposób przedstawienia danych	Wnioski i wskazówki	Redukcja odchyleń	Oczekiwane rezultaty

Źródło: Opracowanie własne na podstawie [4,8]

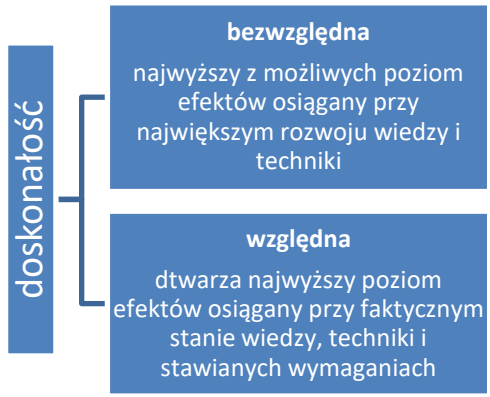
Metoda Six Sigma wyznacza określone fazy działania, do których należą właśnie definiowanie, pomiary, analiza, ulepszanie i kontrola. (tab. 1).

Sam proces tworzenia definicji wbudowuje się w etap opisu i poznawania procesów, skupiając się na ich brakach – słabej stronie. W etapie definicyjnym wyznaczamy również cele do osiągnięcia dla danego projektu wraz z kalkulatorem Six Sigma (odchylenia, liczba błędów) i tabelą opisu przyczynowoskutkowego. W momencie przystąpienia do pomiaru musimy pamiętać o dokładności i precyzyji tegoż, tak by uzyskane wyniki pomiarów dawały jasne przekazy dla analiz i wyciągania wniosków. Sama analiza skutkuje już identyfikacją przyczyn określonych skutków działań, czy tez określonego błędu. Jednocześnie zdobywamy dane dotyczące całego procesu na każdym etapie, poszukując przyczyn zmian oraz możliwych zależności skutków. Analiza przekrojowa procesu wraz z opisem i planami procesu może zostać zbadana najprostszymi testami statystycznymi. Po zakończeniu analizy statystycznej następuje etap ulepszania. Ulepszania, czyli ingerencji w obecny proces w celu jego poprawy. W etapie tym staramy się redukować odkryte i przeanalizowane wcześniej przyczyny błędów i odchyleń, wraz z realnym nadzorem nad skutkiem wdrożonych ulepszeń. Najlepszym rozwiązaniem jest badanie realnego skutku po wdrożeniu zmiany dzięki na przykład analizie Pareto czy też analizie kart kontrolnych. Tak dochodzimy do etapu kontroli, czyli obserwacji starającej się zapewnić określoną jakość.

Głównymi zadaniami Six Sigma jest miara wielkości satysfakcji odbiorcy definiowana na każdym etapie/poziomie projektu, określająca jakość na milion możliwości przed i po zmianie.

Należy też pamiętać, iż do wdrożenia tej metody nieodzowny jest określony, przeszkolony personel. Czynnik ludzki jest ważną częścią systemu a warunkiem skutecznego wdrożenia metody Six Sigma jest działanie we wszystkich komórkach organizacji jednocześnie i w taki sam sposób. Odpowiedni personel powstaje jedynie wraz z procesem budowania wzajemnych relacji, zaufania i wspólnej wizji. Wybór przywódcy determinuje prawdopodobieństwo sukcesu. Od kierowników wyższego szczebla oczekujemy stworzenia możliwości do działania, zasobów i określenia pełnomocnictw zakresu władzy – tak by pomóc we wprowadzaniu usprawnień i decydowaniu. Kierownicy średnia niższego i lub kierownicy zespołów odpowiadających za wdrożenie Six Sigma oraz eksperci z tego zakresu również muszą polegać na zaufaniu i możliwościach nadanych przez „zarząd". Zaś pracownicy realizując określone projekty wykorzystują do tego oprócz powyższych szereg doświadczeń i wiedzy zdobytych podczas spotkań przedsystemowych i szkoleń z tym związanych, na których uzyskuję stopnie/pasy wskazujące na doświadczenie.

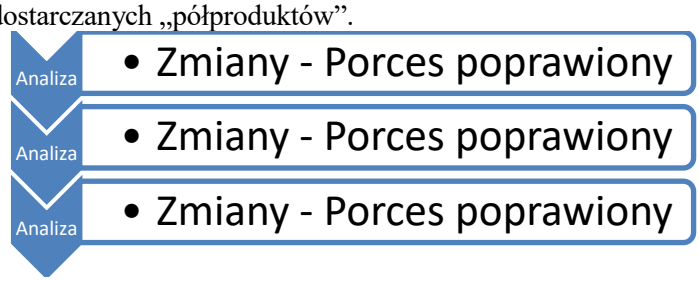
Potrzeba ciągłego doskonalenia jest podstawą systemów jakości w świecie i dąży do osiągnięcia doskonałości jako najkorzystniejszego z możliwych stanów wszystkich przymiotów rozpatrywanego przedmiotu analiz [7].



Rysunek 1. Podział doskonałości produktu/usługi
Źródło: Opracowanie własne na podstawie[7,8]

Six Sigma to proces dzięki któremu unikamy niskiej efektywności pracy zasobów organizacji. Proces, który eliminuje błędy każdej fazy i nie pozwala marnotrawić niczego – czasu, zasobów a nawet odpadów. Jednak samo wdrożenie jest procesem długotrwałym, wymagającym przede wszystkim dobrego przygotowania przedsiębiorstwa do pracy z tym narzędziem. Organizacja winna zrobić przegląd na każdym etapie swojej działalności tak by uzyskać niezbędną wiedze w zakresie przygotowania produkcji/usługi, samej produkcji lub też tworzenia usługi, techniki i technologii stosowanych, kultury organizacji oraz sprzedaży, a co za tym idzie, sposobu dbania o klienta-odbiorcę [8].

Dzięki podejściu strategicznemu Six Sigma można redukować koszty jakości w wielu obszarach. Podstawowym z nich jest obszar działalności operacyjnej i możliwości oraz umiejętności jego usprawniania. Innym może być przestarzała technika/ technologia czy też błędy w samych procesach budowy produktu/usługi. W obszarze planowania kosztów brak ciągłej kontroli z realnymi wydatkami czy też brak ciągłego badania jakości dostarczanych „półproduktów".



Rysunek 2. Cykl procesu analiz ciągłych. Źródło: Opracowanie własne na podstawie [8, 9]

Mikel Harry i Richard Schroeder, autorzy książki o metodzie Six Sigma, a także wdrożeniowcy w Motoroli powiedzieli kiedyś że „Przeważnie okazuje się, że po wprowadzeniu Six Sigma rentowność przedsiębiorstwa jest średnio o 6 – 8% wyższa od planowanej. Za tak imponujące rezultaty pracowników wdrażających Six Sigma należy odpowiednio wynagradzać – w przeciwnym razie mogą odchodzić do innych firm, oferujących lepsze warunki płacowe. Nie powinny mieć miejsca sytuacje, kiedy na sukcesie finansowym Six Sigma korzystają jedynie udziałowcy i liderzy firmy" [8].

Poziom jakości Six Sigma (6σ) w danym miejscu procesu oznacza, że odchylenie standardowe pomiarów mieści się w założonym przedziale specyfikacji 12 razy równe jest popełnianiu 3,4 błędu na 1 mln możliwych do popełnienia błędów[9].

2.2. Lean

Lean management, czyli tzw. „szczupłe" zarządzanie, stanowi rozszerzenie koncepcji produkcyjnej *lean, któ*ra wykorzystuje zasady i narzędzia Systemu Toyoty [10]. Jako jedna z technik zarządzania jednostką jest często wykorzystywana przy restrukturyzacji. Jej zadaniem jest pozbycie się zbędnych procesów – czyli jej odchudzenie.

Można wyróżnić pięć zasad, zgodnie z którymi powinna działać organizacja działająca w zgodzie z zasadami „lean"[11]:

- wartość dla klienta – cecha ważna dla odbiorcy w danym produkcie/usłudze,
- identyfikowanie strumienia wartości i czynności w tym strumieniu – analiza czynności od momentu złożenia zamówienia do dostarczenia produktu/usługi odbiorcy,
- ciągły ruch – płynność wytwarzania, brak zastojów, zakłóceń czy niezasadnych przerw,
- ciągnienie przez popyt – tempo produkcji zależne od rzeczywistego popytu,
- doskonałość – ciągłe doskonalenie wszystkich procesów.

Pierwszą jednostką stosującą ww. podejście do produkcji, opierające się na wysokiej jakości, ciągłym doskonaleniu, elastyczności i eliminowaniu marnotrawstwa była właśnie firma Toyota Motor Company, która produkowała bardzo wysokiej jakości auta zużywając mniej zasobów niż jej konkurenci [12].

Zarządzanie poprzez Lean dąży do ograniczenia zasobów produkcji, zarówno po stronie zasobów ludzkich jak i powierzchni, czasu, nakładów inwestycyjnych itd., przy jednoczesnym zwróceniu bacznej uwagi na optymalne ich wykorzystanie. Dążymy do wygenerowania produktu odpowiadającego w większym zakresie oczekiwaniom naszych klientów przy jednoczesnej produkcji o niższych kosztach w porównaniu do systemu tradycyjnego. Dążymy więc do stworzenia prostych struktur organizacyjnych wraz ze spłaszczeniem hierarchii i postawieniem na znaczenie zasobów ludzkich w dążeniu do osiągania perfekcji [13]. Wdrożenie powyższej metody może pozytywnie wpływać na konkurencyjność przedsiębiorstwa, ograniczenie kosztów działalności, poprawę elastyczności do zmian w popycie, poprawę rotacji zapasów, wzrost produktywności, poprawę jakości, poprawę przepływów finansowych, spadek liczby wypadków przy pracy [14, 15].

Błędy występujące można nazwać min. marnotrawstwami zasobów. Problemów marnotrawienia jest kilka, i związane są one z możliwością ich ograniczenia. Od nadmiernych zapasów i niskiej wydajności pracy – czyli chęci usunięcia czynności, których wykonywanie zużywało zasoby nie dodając wartości, po nadprodukcję, nieergonomiczne zachowania pracownika i surowców, okresy przestoju czy wreszcie niewłaściwe wykonanie usługi/produktu. Należy też pamiętać, iż często dużym problem jest niewykorzystanie potencjału pracowniczego przedsiębiorstwa [16].

Na koncepcję zarządzania Lean składać się muszą:

- praca zespołowa i optymalne zarządzanie zasobami ludzkimi stawiające na kierownictwo średniego szczebla,
- bezpośredni kontakt z dostawcami,
- kontrola jakości, bieżący nadzór i usuwanie błędów wraz z stałym – symultanicznymi zmianami w systemie.

Głównym celem jest przygotowanie przedsiębiorstwa do zmieniających się warunków otoczenia firmy, by mogło ono reagować elastycznie i stało się nowoczesną organizacją nakierowaną na potrzeby zmieniającego się rynku.

Głównym narzędziem jest tzw. „Mapowanie Strumienia Wartości" dla gromadzenia danych na temat realnego przepływu zasobów i informacji. System ten wspierany jest poprzez narzędzia związane z pięcioma Japońskimi słowami (tzw. 5S) nadającymi standardy pracy[13]:

- Seiri – selekcja – w miejscu pracy trzymamy tylko to co niezbędne,
- Seiton – magazynowanie – uporządkowane miejsce pracy, wszystko musi mieć swoje oznaczone miejsce,
- Seiso – sprzątanie – sprzątanie codzienną czynnością,
- Seiketsu – standaryzacja – dla wykonywania czynności z większą łatwością,
- Shitsuke – samodyscyplina – utrzymywanie standardów, utrzymywanie wcześniej ustalonych zasad.

Oprócz wyżej wymienionych narzędzi do wdrażania Lean stosuje się też między innymi metody JiT (Just in Time), SMED (Single Minute Exchange of Die), TPM (Total ProductiveMaintenance), zasadę przepływu jednej sztuki, TQM Total Quality Management.

Z obserwacji wynika, że w Polsce coraz większa liczba firm próbuje wdrażać Lean Management, jednak nie jest w stanie do końca utrzymywać i/lub wdrożyć założeń Lean, ze względu na liczne bariery. Do barier zaliczyć możemy brak wsparcia i zaangażowania członków najwyższego i średniego szczebla kierownictwa, oraz opór pracowników wynikający z braku zaufania do kierownictwa, negatywne doświadczenia z zarządem czy też wreszcie brak wiedzy tematycznej. Jednocześnie pojawia się też brak chęci zmiany kultury organizacyjnej oraz konsekwencji wdrażania, częsty brak powiązania z celami strategicznymi firmy [17].

Zaletami systemu na pewno jest wysoka kultura organizacyjna, co za tym iść może – wysoka konkurencyjność rynkowa i wydajność procesów, zorientowanie na klienta i na koniec, ale nie najmniej ważne – satysfakcja i zaufanie pracowników.

3. Lean Six Sigma

Połączenie obu metod w niektórych przypadkach pozwala na osiągnięcie maksymalnych możliwych efektów. Lean i Six Sigma wzajemnie się uzupełniają. Lean Six Sigma powoduje przyspieszenie, dostarczając lepszych wyników niż to, co zazwyczaj osiąga się przez Lean i Six Sigma indywidualnie.

Połączenie tych dwóch metod daje zespołowi kompleksowy zestaw narzędzi do zwiększenia skuteczności każdego procesu wewnątrz organizacji w wyniku czego wzrosnąć mogą przychody, obniżyć się koszty i nastąpić znacznie lepsza współpraca [18]. Lean to metoda usprawnienia procesu, Six Sigma to metoda efektywnego rozwiązania problemu wad, jeśli zdecydujemy się na Lean Six Sigma to pozwalamy na zwalczanie problemów i rozwijanie przedsiębiorstwa określonymi łączonymi sposobami.

Lean Six Sigma zwiększa przychody poprzez usprawnienie procesów, a uproszczone procesy generują produkt/usługę wykonaną szybciej, wydajniej, bez żadnych dodatkowych kosztów związanych z jakością. Lean Six Sigma zwiększa przychody dzięki zwiększeniu produkcji przy obniżeniu kosztów błędów i organizacji procesu czyli zarówno sprzedaż, produkcja jak i dostarczenie większej ilości produktów lub usług przy pomocy mniejszych zasobów [18].

Połączenie metod Lean i Six Sigma zmniejsza koszty danej organizacji i poprawia jej efektywność poprzez:

- usuwanie „odpadów" z procesu – tj. każdej aktywności w procesie, która nie jest wymagana do wytwarzania produktu/świadczenia usługi,
- rozwiązywanie problemów wynikających z procesu – kosztownej wady produktu lub usługi która generowała obciążenia,
- maksymalizację wysiłków dla zapewnieniu zadowalającego odbiorcę produktu/usługi,
- umożliwienie reorganizacji i realokacji zasobów min. wyrażonych przychodami, czasem pracy, zasobem kadrowym wypracowanym dzięki nowemu ulepszonemu procesowi.

Jednocześnie Lean Six Sigma poprawia i wpływa na skuteczność działań zasobu kadrowego dzięki zaangażowaniu pracowników w proces doskonalenia, budowaniu zaufania, przejrzystości działań i ich skutków na wszystkich poziomach organizacji oraz promowaniu ważności każdej osoby dla organizacji [18].

Lean Six Sigma pozwala ustalić procesy, które obciążają przedsiębiorstwo niezasadnym wykorzystaniem zasobów oraz pozwala na tworzenie tak efektywnych procesów, że jednostka może dostarczyć więcej produktów lub usług, dla bardziej zadowolonych klientów niż kiedykolwiek wcześniej. Jednocześnie rozwija poczucie własności i odpowiedzialności za siebie i współpracowników, zwiększając skuteczność działań zasobu ludzkiego w osiąganie.

Metoda poprawy wydajności i efektywności wymaga stosowania kilku zasad:

- Skoncentruj się na kliencie,
- Zidentyfikuj i zrozum, jak to ma działać *(strumienia wartości)*,
- Zarządzaj – popraw przebieg procesu, ogranicz zmienność i ścieżki,

- Usuń nie potrzebne działania,
- Angażuj i wyposaż pracowników w procesie,
- Systematycznie podejmuj działania w celu poprawy.

Lean Six Sigma daje skuteczność przy niskich kosztach poprawiając kulturę organizacji i jakość produktów/usług. Przyspieszenie cyklu produkcji zarówno towaru jak i usługi przy mniejszej ilości błędów i poświęconego czasu na poprawki [19].

W oczach odbiorców wartością jest przede wszystkim odpowiedni produkt i/lub usługa podana we właściwym czasie i optymalnej cenie, w odpowiedniej jakości. Zainteresowanie metodą łączoną wzrasta na całym świecie, w różnych branżach.

4. Lean Six Sigma w medycynie i nie tylko

Na przykładzie firmy Xerox możemy zauważyć że Lean Six Sigma jako proces doskonalenia opierający się na kulturze organizacji jako bazie, poprzez usprawnienie procesów, daje trwałą zmianę całej firmy. Najważniejsze fakty jakie można przytoczyć w związku z projektem w ww. firmie to min. ilość wyszkolonych pracowników od najwyższego zakresu tj. czarnych pasów (1300 przeszkolonych osób) po zielone pasy (9.000osób) czyli ponad 70% kadry kierowniczej, aż po najniższy szczebel w ilości 45.000 osób w żółtych pasach wraz z ponad 7.500 zrealizowanych projektów usprawniających w każdym obszarze businessu [20]. Firma GE natomiast już w latach osiemdziesiątych rozpoczęła działania w kierunku dbałości o jakość. Kulturę organizacji postawiono na szali biurokracji i bezrefleksyjnych działań codziennych – osadzając ja na jakości myślenia procesowego – w poprzek i na każdym poziomie i w każdej funkcji [21].

W Gronningen University Medical Centre w Holandii wykazano, iż 30% długości pobytu w szpitalu było niepotrzebne, co więcej ograniczono o 50% zbędne pobyty szpitalne najczęściej związane pacjentami urazowymi. Jednocześnie zmniejszając długości w obszarze „oczekiwania" z 11,9 godziny do 3,4 godziny. W szpitalu Valley Baptist w Texasie w USA nastąpił spadek długości pobytu o 19% i spadek kosztów w wysokości prawie 3,1 miliona dolarów rocznie [22].

W Floyd Medical Centre gdzie celem zastosowania metody była poprawa wydajności operacyjnej, oprócz celu głównego udało się osiągnąć poprawę rotacji łóżka o ponad godzinę a długości pobytu w trybie pilnym oddziału ratunkowego zmniejszyła się o 4 godziny. Jednocześnie wykazano oszczędności w wysokości 6,3 miliona euro rocznie.

W Miami BaptistHospital z kolei oszczędności wyniosły 4.2 miliony dolarów, nastąpiło usprawnienie zarządzanie łóżkiem i procedury wypisu, w celu zmniejszenia długości pobytu pacjenta. Zmniejszyła się liczba pacjentów, którzy zostali wypisani bez leczenia oraz zwiększyły roczne przychody przy jednoczesnym spadku kosztów eksploatacji. Proces wypisu przed 14:00 wzrósł z 41% do 80%, długość pobytu na SOR została zmniejszona o 41%, odsetek nieleczonych pacjentów spadł z 8% do mniej niż 1%, a długość pobytu pacjenta na łóżku SOR zmniejszył się o 37%. Techniki Lean Six Sigma zostały wdrożone również w izbie przyjęć w jednym ze szpitali w USA. Co

spowodowało wzrost przyjęć pacjentów o ponad 9% przy zmniejszeniu długości pobytu pacjenta i bez podnoszenia kosztów ponoszonych „na pacjenta"[23].

Musimy wziąć pod uwagę, że centralną rolą kadry kierowniczej średniego szczebla jest usuwanie barier związanych z ograniczeniami chęci do działania pracowników i tworzeniem nowych rozwiązań. Rola ta winna być spełniana poprzez tworzenie osiągalnych celów, wsparcie i narzędzia ułatwiające ww. działania i dążenie do stworzenia optymalnej kultury organizacyjnej.

Ministerstwo Zdrowia w Quebec w Kandzie wdrożyło projekt inwestycyjny "Lean Healthcare Six Sigma", w którym trzy szpitale otrzymały 10 milionów dolarów na jego realizację. Minister Zdrowia, Yves Bolduc, stwierdził że „szpital należy traktować jak przedsiębiorstwo, wizytę pacjenta jak proces produkcji, a chorego człowieka jak produkt". Zaletami programu miała być poprawa dostępu do opieki medycznej, zmniejszenie czasu oczekiwania pacjentów na łóżko, czy zabieg, usprawnienie procesów wspierających główne strumienie wartości, a tym samym obniżenie kosztów działających szpitali. Celem poprawy wydajności od 25 do 30 % z tym samym personelem. Jednocześnie żaden z pracowników nie straci pracy, nie będzie redukcji czasu pracy a pracownicy nie będą przeciążeni pracą (jedna z zasad "Lean"). Natomiast będą mieli bardziej racjonalny nakład pracy. Projekt zwrócić się miał w ciągu 3 do 6 miesięcy, gwarantując po pewnym czasie również płynność finansową. Priorytetem było zyskanie na jakości życia ludzi. Mimo faktu, iż nie wszystkie założenia udało się spełnić, minister nadal planuje lobbować, aby wszystkie zakłady opieki zdrowotnej zastosowały lean'owskie podejścia do zarządzania. Ponieważ nastąpił spadek średniego całkowitego czasu rejestracji z 50 minut do 29 minut, czyli obniżył się aż o 42 %, wyeliminowano długie kolejki i zoptymalizowano poszczególne fazy obsługi pacjenta od rejestracji aż do analizy próbki w laboratorium. Kilka pomieszczeń szpitalnych zostało przekształconych, aby ułatwić przemieszczenie się pacjentów, a godziny pracy pracowników zostały zmodyfikowane dla osiągnięcia wyznaczonych celów. Czas pobytu w Izbie Przyjęć (tj. zanim pacjent zostanie zbadany i skierowany do sali szpitalnej) trwający 48 godzin i dłużej dotyczył 14% pacjentów obecnie odsetek ten zmniejszył się trzykrotnie a średni czas pobytu w Izbie Przyjęć wykazał się ponad 20% spadkiem w ciągu miesiąca, z 24,1 do 19,1 godzin. Zmniejszyła się ilość godzin pracy w godzinach nadliczbowych, z 6.000 do 3.600 godzin, przy jednoczesnym braku zwolnień. Personel medyczny spędzał średnio 10% czasu na wykonywaniu innych zadań niż opieka nad pacjentem, dlatego też optymalizując proces, placówki medyczne, przy takim samych środkach i zasobach są teraz w stanie obsługiwać więcej ludzi, a jakość usług nie uległa pogorszeniu [22].

Oczywiście nie zawsze działania strategii Lean Six Sigma przynoszą jedynie pozytywny i planowany skutek. Pewne badania w systemie angielskim (NHS) wykazały, że metoda Lean może nie dostarczyć poprawy efektywności w sektorze opieki zdrowotnej, jednak wymaga zmiany paradygmatu filozofii zarządzania, by w niej przyjąć metodę Lean szerzej i bardziej skutecznie [24,25].

Szpitale, które już wdrożyły tę metodologię zaczęły już częściowo czerpać z niej korzyści. Nastąpił np. spadek odsetka zakażeń szpitalnych MRSA o 51% [26,27]. Wyniki zastosowania ww. metody, statystycznie pokazują możliwy przyrost marży

o 20%, wzrost wydajności o 12 – 18%, możliwość redukcji koniecznego zatrudnienia o 12% czy też możliwość obniżenia niezbędnego kapitału obrotowego o 10-30%.

Więc pamiętać należy, że ze względu na wzrost popytu a tym samym zaporzebowania na łóżka szpitalne w jednostkach publicznych, np. w Australii, wymagany jest wzrost o 62% łóżek, przy oczywiście ogromnych kosztach [28] a w większość szpitali angielskich średni współczynnik obłożenia wrósł w ciągu półrocza o blisko 100% – co może prowadzić do wzrostu prawdopodobieństwa zgonu w szpitalu o 30% [29,30].

Zauważamy więc jak ważny jest zintegrowany i efektywny proces zarządzania łóżkiem, i jak jest potrzebny do poprawy obecnej i przyszłej sytuacji związanej ze wzrostem popytu na usługi szpitalne. Jednocześnie z uwagi na fakt, iż system opieki medycznej w Polsce, w większości należy do systemu publicznego, obecnie koncepcje zarządzania zorientowane na doskonalenie jakości są możliwe do zastosowania w również tychże instytucjach [31].

5. Podsumowanie

Lean i Six Sigma jako metody poprawy jakości są szeroko stosowane zarówno w branży produkcyjnej, usługowej jak i dość specyficznej, szeroko pojętej branży medycznej. Połączenie obu metod jest natomiast koncepcją dość nową w opiece zdrowotnej, zwłaszcza w procesie zarządzania kosztem, obiegiem pacjentów i obłożeniem łóżka. Zmiany jakie następują w systemie opieki zdrowotnej, również w Polsce, przymuszają do szukania nowych rozwiązań. Gdyż stare podejście optymalizacji kosztów poprzez ograniczenie zasobów ma wpływ na jakość usług – a na obniżanie tejże jakości pozwolić nie może sobie żaden system.

Zastosowanie nowego, bardziej efektywnego procesu zarządzania systemem łączonych strategii jest tematem wartym rozważania. Wiele szpitali na świecie wdrożyło metody Lean Six Sigma bądź też metody Six Sigma lub Lean. Działania te związane są z chęcią poprawy wydajności i efektywności procesu, tym samym czasu pobytu pacjenta, oraz generowaniu wzrostu przychodów bez wzrostu zużycia kosztów i/lub zasobów. Oczywiście okazuje się też, że każdy z procesów ma swoje wady i ograniczenia. Wiadomym jest, iż w warunkach szpitalnych, kwestie kultury organizacyjnej, zrównoważonego rozwoju, zaangażowania kadry niższego i wyższego szczebla, metody komunikacji, itp. czasami pozostawiają pole do rozwoju. Głównym problemem w systemie polskim może więc być niechęć do wprowadzenia zmian czy też problem wykształcenia lub zaangażowania kadry pracowniczej – jednakże oba te ograniczenia można dość łatwo wyeliminować stosując systemy motywacyjne i szkolenia tzw. miękkie. A dzięki realizacji wyznaczonych krytycznych czynników sukcesu i przezwyciężeniu ograniczeń oraz barier Lean Six Sigma może być bardzo skutecznym narzędziem w poprawie procesu zarządzania łóżkiem szpitalnym, w celu sprostania przyszłemu popytowi na usługi medyczne.

Literatura

1. Barney M.,McCarty T., *Nowa Six Sigma*, Wydawnictwo Helion, Gliwice 2005.
2. Nowosielski S. *Koncepcja Lean management w małym przedsiębiorstwie. Możliwości i ograniczenia zastosowania*, Przedsiębiorczość i zarządzanie 2015 .
3. Pande P., Neuman R.,Cavanagh R., Six Sigma Sposób poprawy wyników nie tylko dla firm takich jak GE czy Motorola, Wydawnictwo K.E. Liber s.c., Warszawa 2003.
4. Harry M., Shroeder R., Sześć sigma, wykorzystanie programu jakości do poprawy wyników finansowych, Oficyna Ekonomiczna, Kraków 2001.
5. http://www.polishsixsigmaacademy.pl/resources.php?category=1.
6. https://mfiles.pl/pl/index.php/Six_sigma.
7. Kolman R., *Kwalitologia. Wiedza o różnych dziedzinach jakości*, Wydawnictwo Placet, Warszawa 2009, s. 25.
8. Mikel H.,Schroeder R.,*Six Sigma* Oficyna Ekonomiczna, Kraków 2005.
9. http://simul8healthcare.com/simulation-software.
10. Czarnecki J., SikorskiCz.: *Lean management[w:] I. Sobańska (red. nauk) Lean accounting integralny element lean management. Szczupła rachunkowość w zarządzaniu*. Warszawa: WoltersKluwer Polska, 2013, s. 11–13, 25.
11. Grycuk A. *Bariery w stosowaniu koncepcji lean management*. „Kwartalnik Nauk o Przedsiębiorstwie". 3, s. 73–74, 2016.
12. Miller J., Wróblewski M., Villafuerte J.: *Kaizen. Budowanie i utrzymanie kultury ciągłego doskonalenia*. Warszawa: MT Biznes, 2014.
13. http://lean-management.pl/lean-management/koncepcja-lean-management/.
14. Hill A.: *The Encyclopedia of Operations Management*. FT Press, 2011.
15. Liker J. (ed.): *Becoming Lean. Inside Stories of U.S. Manufacturers*. New York: Productivity Press, 1998.
16. Dekier Ł., Grycuk A.: *Programy sugestii pracowniczych. Doświadczenia polskich przedsiębiorstw*. Wrocław: Stowarzyszenie Lean Management Polska, 2014.
17. Podobiński M. „Bariery i ograniczenia wdrażania koncepcji Lean Management– wyniki badań", Nauki o zarządzaniu Management Sciences 3(24) 2015.
18. http://www.goleansixsigma.com/wp-content/uploads/2012/02/The-Basics-of-Lean-Six-Sigma-www.GoLeanSixSigma.com_.pdf.
19. http://www.wz.uw.edu.pl/pracownicyFiles/id12340-Six%20Sigma.pdf.
20. https://www.xerox.com/assets/pdf/2012_GRC_LeanSixSigma_Brochure.pdf.
21. http://www.ge.com/sixsigma/SixSigma.pdf.
22. https://simul8healthcare.com/our-customers/case-studies.
23. Dickson E., Singh S, Cheung S., Christopher C., Nugent N., Application of Lean manufacturingtechniques in the department*The Journal of Emergency Medicine, 2009,*Vol: 37 (2), pp. 177-182. .
24. Randor M., Lean in health care: The unfilled promise?*Social Science and Medicine, 2012* Vol: 74 (3), pp.364-371.
25. McintoshB.,CooksonS., Lean management in the NHS: fad or panacea*British Journal of Healthcare Management, 2012* Vol: 18(3), pp. 130-135.
26. Carboneau C., Benge E., Jaco T., Robinson M. A Lean Six Sigma Team Increases Hand Hygiene Compliance and ReducesHospital-Acquired MRSA Infections by 51%,*Journal of Healthcare Quality, 2010,*Vol: 32 (4), pp. 61-70.
27. Faulkner T., Stuenkel K. A community hospital's journey into Lean Six *SigmaJournal of Health Service Management, 2009,Vol: 26 (1), pp. 5-13.*

28. Scott A., Public hospital bed crisis: Too few or too misused?*Australian Health Review, 2010,*Vol: 34(3), pp. 317-324.
29. Jones R., Hospital bed occupancy demystified, *British Journal of Health care Management, 2011,*Vol: 17(6), pp. 242-8.
30. https://www.linkedin.com/pulse/feasibility-lean-six-sigma-hospital-bed-management-process-ullah.
31. Kobylińska U., *Orientacja na jakość w administracji publicznej*, Administracja publiczna. Studia krajowe i międzynarodowe" 2013 nr 2.

Lean Six Sigma metodami poprawy efektywności w opiece medycznej

Streszczenie
Tematyka zarządzania poprzez jakość w polskim systemie opieki medycznej, jak do tej pory, jest postrzegana jedynie poprzez pryzmat istniejących lub narzuconych systemów norm. Najpowszechniej znane są normy ISO. W prezentowanym opracowaniu proponujemy podejście systemowego optymalizowania jakości dzięki zastosowaniu znanych systemów Lean oraz Six Sigma. Przegląd zysków i strat wynikających z zastosowania połączenia obu systemów (metod) pokazuje pożądane, pozytywne efekty. W celu poprawy wyników finansowych, zwrotu inwestycji czy optymalnego zarządzania „łóżkiem", bazując wyłącznie na posiadanych zasobach warto pochylić się nad takim rozwiązaniem. Główną metodą analizy był przegląd literatury obcej z wymienionego zakresu. Niestety w literaturze przedmiotu brak Polskiej perspektywy oraz brak opisanych polskich doświadczeń.
Słowa kluczowe: Lean, Six Sigma, jakość, zarządzanie, koszt

Lean Six-Sigma to improve efficiency in medical care

Abstract
Quality management in the Polish health care system, so far, is seen only through the prism of standards, for example ISO. To go further, we propose an approach to optimize system quality through the use, of already known in the West for many years, systems, Lean and Six Sigma. Overview of profit and loss of both methods showed positive results. To improve financial performance, return on investment and bed management, based on the owned resource, shows that it is worth to think about such a solution. The main method of analysis was to review the literature in the field, but in Polish literature, there is no experience described.
Keywords: Lean, Six Sigma, quality, management, cost

Anna Kasperczuk[1]

Inteligentne systemy rekomendacyjne i ich zastosowanie

1. Wstęp

Systemy rekomendacji są obecnie ważnym ogniwem w działalności wielu instytucji i organizacji. Badania wskazują, że systemy te przyczyniają się, w przypadku sklepów internetowych, do wzrostu przychodów o jedną trzecią. Badanie systemów rekomendacji (RS) jest stosunkowo nowe w porównaniu do badania klasycznych narzędzi i technologii informacyjnych [1, 2]. Systemy rekomendacji odgrywają znaczącą rolę na wielu witrynach internetowych, np. Google, Yahoo. Obecnie organizowane są konferencje dedykowane tej dziedzinie. Na wielu uczelniach technicznych pojawiają się kierunki studiów całkowicie poświęcone metodom rekomendacji.

2. System rekomendacyjny

System rekomendacyjny stanowi zbiór narzędzi i technik, które mają na celu sugerowanie użytecznych propozycji. Propozycja, rozumiana jest jako element zalecany użytkownikom, na przykład płyta CD, książka lub w przypadku nauk medycznych – metoda leczenia. W celu uproszczenia, spersonalizowane zalecenia przedstawiane są w postaci listy elementów. Ranking ten jest generowany przez stosowanie algorytmów predykcyjnych, w oparciu o preferencje i ograniczenia użytkownika. Preferencje użytkowników mogą być zbierane poprzez interakcję z użytkownikiem lub pośrednio, poprzez śledzenie zachowania użytkowników [2, 3].

Rozwój systemów rekomendacyjnych opiera się na wiedzy z różnych dziedzin, takich jak: sztuczna inteligencja, technologie informatyczne, metody *Data Mining*, statystyka, systemy wspomagania decyzji lub marketing [4]. W porównaniu do innych narzędzi i technik informatycznych, systemy rekomendacyjne są stosunkowo nową metodą. Ten nowy, niezależny obszar badawczy powstał w połowie lat 90, ale w ostatnich latach obserwujemy gwałtowny wzrost zainteresowania tą gałęzią technologii informacyjnych. Obecnie są one z powodzeniem stosowane jako część wielu witryn e-commerce, które oferują wiele istotnych korzyści biznesowych: wzrost liczby sprzedanych towarów i usług, zwiększając tym samym satysfakcję użytkownika i jego lojalność.

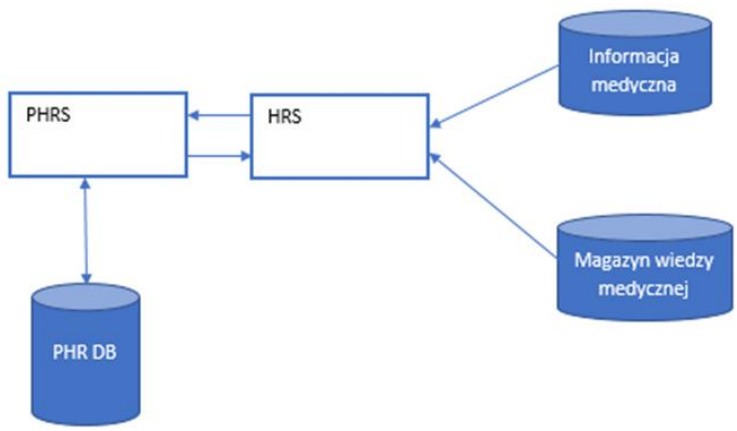
Dane wykorzystywane przez systemy rekomendacji dotyczą następujących typów obiektów:

- przedmiotów (np. produktów i usług),
- użytkowników (np. klientów, konsumentów),
- transakcji (relacji pomiędzy użytkownikami i przedmiotami).

[1] a.kasperczuk@pb.edu.pl, Politechnika Białostocka, Wydział Mechaniczny, Zakład Biocybernetyki i Inżynierii Biomedycznej

2.1. Medyczny system rekomendacyjny

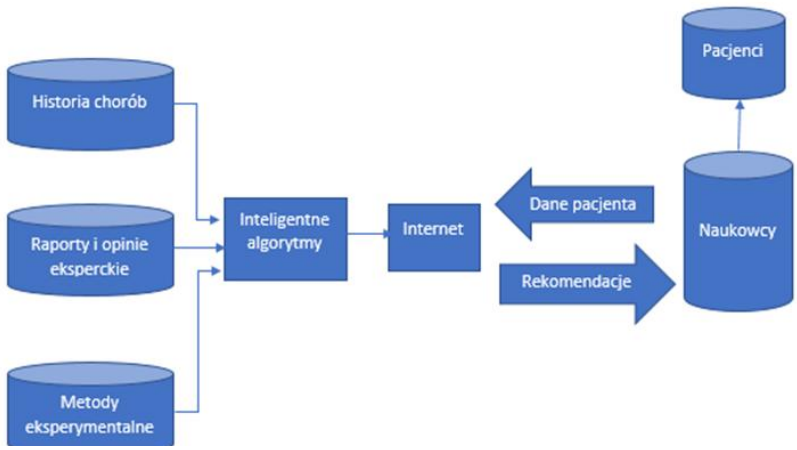
Mimo że systemy rekomendacyjne (ang. *recommender system* RS) są głównie opracowane na potrzeby e-commerce, podjęto próby przystosowania ich do zastosowań medycznych. Głównym celem tych prac jest potrzeba pomocy lekarzom w podejmowaniu decyzji bez konieczności bezpośrednich konsultacji specjalistów. Medyczny system rekomendacyjny (HRS) zaproponowany w [5], jest wyspecjalizowanym RS, w którym rekomendowany przedmiot zainteresowania stanowi część informacji medycznej, która nie jest związana z historią medyczną danej osoby.



Rysunek 1. Budowa systemu PHR powiększonego o HRS [5]

Wspomniane propozycje są tworzone na podstawie zindywidualizowanych danych zdrowotnych, takich jak np. indywidualny rekord pacjenta (ang. *Personal Health Record* PHR), który może być uznane jako "profil użytkownika" w systemie rekomendacyjnym. Medyczny system rekomendacyjny jest zaimplementowany jako rozszerzenie istniejącego systemu PHR, w którym istnieją wpisy w bazie danych (rys. 1), zaś HRS oblicza zestaw potencjalnie interesujących pozycji dla użytkownika docelowego.

W pracy [6] zaproponowano zastosowanie rekomendacyjnego systemu medycznego na potrzeby telemedycyny, a tym samym pomoc szpitalom położonym w odległych rejonach, w których jest niewystarczająca ilość lekarzy, zbyt małe doświadczenie w diagnozowaniu wszystkich rodzajów schorzeń oraz niewystarczająca infrastruktura. Architektura proponowanego systemu została przedstawiona na rysunku 2. System [6] wdraża inteligentne algorytmy, takie jak sieci neuronowe i techniki eksploracji danych. Ponad to wykorzystuje hybrydową technikę filtrowania: wspólną filtrację (ang. *collaborative filltering*), opartą na treści oraz filtrowanie oparte na wiedzy. Wspólna filtracja jest stosowana w historii medycznej pacjentów: filtrowanie oparte na treści wykorzystywane jest do badań eksperymentalnych, natomiast filtrowanie oparte na wiedzy stosowane jest do opisu przypadków oraz opinii ekspertów. Lekarze mogą uzyskać dostęp do systemu z dowolnego miejsca. Rekomendacje są dostarczane na podstawie wywiadu z pacjentem, po czym może nastąpić konsultacja poszczególnych przypadków w celu przeprowadzenia rozpoznania.



Rysunek 2. Koncepcyjny schemat medycznego systemu rekomendacyjnego [6]

3. Rekomendacja oparta na współpracy (ang. Collaborative recommendation)

Najbardziej popularnym podejściem, które używane jest m.in. w wielu księgarniach internetowych, badanie opinii oraz gustów dużej społeczności użytkowników (lub pacjentów w HRS) do generowania spersonalizowanych rekomendacji. Połączone techniki filtrowania, stosowane w tym podejściu, wykorzystują zalecenia sporządzone przez społeczność użytkowników (grupę pacjentów) do wydania rekomendacji dla aktywnego użytkownika (obecnego pacjenta). Metoda ta opiera się na fakcie, że ludzie często sugerują się opiniami i zaleceniami innych użytkowników w podejmowaniu codziennych decyzji. Osoby w podobnym wieku, ze zbliżonych klas społecznych, o takich samych zainteresowaniach mają podobny gust. Podobna zależność występuje w przypadku medycznych systemów: pacjenci o podobnej charakterystyce i objawach mogą być leczeni w zbliżony sposób [7-9].

Typowe wejście dla tego typu systemów stanowi macierz danych, natomiast typowe wyjście stanowi wskazanie, w jakim stopniu obecny użytkownik akceptuje daną pozycje. Jedna z pierwszych metod stosowanych w opisywanym podejściu polega na wywoływaniu przez użytkownika na podstawie zalecenia najbliższego sąsiada. Pierwszym krokiem jest identyfikacja użytkowników będących tzw. najbliższymi sąsiadami, czyli innych osób, które miały podobne preferencje w przeszłości. W następnym kroku, przewidywania dla pozycji i jest obliczana w oparciu o wskaźniki dla i przez użytkowników rówieśniczych. Metoda zakłada, że preferencje użytkownika pozostają stabilne i zgodne z upływem czasu.

Tabela 1. Przykładowa baza danych wejściowych [opracowanie własne]

	Produkt1	Produkt2	Produkt3	Produkt4	Produkt5
Aktywny użytkownik	5	3	4	4	?
Użytkownik1	3	1	2	3	3
Użytkownik2	4	3	4	3	5
Użytkownik3	3	3	1	5	4
Użytkownik4	1	5	5	2	1

Tabela 1 pokazuje bazę oceny aktywnego użytkownika oraz innych osób. Poszczególne pozycje były oceniane w skali od 1 do 5. W tym przypadku zadaniem systemu rekomendacyjnego jest przewidzieć, czy aktywny użytkownik lubi Produkt5, który nie został jeszcze użyty lub oceniony. W celu wykonania prognozy, system musi w pierwszej kolejności znaleźć użytkowników podobnych do aktywnego użytkownika. W drugim etapie, system ma za zadanie przewidzieć, czy badana osoba będzie lubić dany element w oparciu o oceny użytkowników, znajdujące się w pierwszym etapie [10].

W systemach rekomendacyjnych stosuje się różne miary, które mogą zostać użyte do określenia zbioru podobnych użytkowników. Najczęściej stosowaną miarą podobieństwa w systemach jest współczynnik korelacji Pearsona (równanie 1). Innymi miarami, które możemy wykorzystać w tym celu są: odległość euklidesowa, odległość Minkowskiego (będąca uogólnieniem odległości euklidesowej), odległość Mahalanobisa, cosinus podobieństwa lub norma L2.

$$sim(a,b) = \frac{\sum_i (r_{a,i} - \overline{r_a})(r_{b,i} - \overline{r_b})}{\sqrt{\sum_i (r_{a,i} - \overline{r_a})^2 (r_{b,i} - \overline{r_b})^2}} \quad (1)$$

Współczynnik korelacji Pearsona przyjmuje wartości w zakresie od +1 (silnie dodatnia korelacja) do -1 (silnie ujemna korelacja). Miara podobieństwa $sim\,(a,b)$ użytkowników a i b, dane w macierzy R, jest określona we wzorze (1), gdzie $r_{a,i}$ oznacza średnią ocenę użytkownika a pozycji i, $\overline{r_a}$ jest średnią ocenę użytkownika a. Po obliczeniu współczynnika między aktywnym użytkownikiem oraz każdym innym, znajdującym się w bazie, okazuje się, że Użytkownik1 i Użytkownik2 mieli zbliżone zachowania w przeszłości do zachowania osoby badanej (odpowiednio podobieństwa środków 0,85 i 0,7).

Po wybraniu użytkowników równorzędnych w stosunku do użytkownika aktywnego, możliwe jest obliczenie prognozy dla Produkt5. Jedną z możliwości jest przewidywanie wskaźnika dla użytkownika a na pozycji i, tak że czynniki względnej bliskości najbliższych sąsiadów N i średnia ocena $a's$ jest dana za pomocą równania 2.

$$pred(a,i) = \overline{r_a} \frac{\sum_b sim(a,b) * (r_{b,i} - \overline{r_b})}{\sum_b sim(a,b)} \quad (2)$$

W podanym przykładzie, przewidywania oceny dla aktywnego użytkownika w oparciu o najbliższych sąsiadów (Użytkownik1 i Użytkownik2) będą następujące:

$$(0{,}85 + 0{,}7) * \big(0{,}85 * (3 - 2{,}4) + 0{,}7 * (5 - 3{,}8)\big) = 4{,}87$$

Podejście probabilistyczne teorii prawdopodobieństwa może być również wykorzystane jako kolejny sposób dokonywania prognozy, jak dany użytkownik będzie oceniać pewną pozycję. W tym sposobie istnieje problem przewidywania, traktowany jako błąd klasyfikacji, który polega na przypisaniu przedmiotu do jednego z kilku predefiniowanych kategorii.

Jedna ze standardowych technik w klasyfikacji oparta jest na klasyfikatorze Bayesa, którego działanie zostało przedstawione na przykładzie przedstawionym powyżej. Zadania przewidywania zostały sformułowane jako problem obliczania najbardziej prawdopodobnej wartość znamionowej dla Produkt5, biorąc pod uwagę szereg innych ocen badanej osoby oraz ocen innych użytkowników. W tej metodzie, prawdopodobieństwo warunkowe jest obliczane dla każdej możliwej wartości znamionowej, a następnie wybierana jest wartość, dla której otrzymano wartość największą. Twierdzenie Bayesa jest stosowane do obliczania prawdopodobieństwa a posteriori $P\,(Y|X)$ przez warunkowe prawdopodobieństwo $P\,(X|Y)$, prawdopodobieństwo Y i prawdopodobieństwo X (wzór 3). W przedstawionym przykładzie, P (Y) może być prawdopodobieństwem wartości znamionowej 1 dla Produkt5 lub innej możliwej wartości znamionowej, zaś X stanowi zbiór innych ocen aktywnego użytkownika.

$$P(Y|X) = \frac{P(Y|X) * P(Y)}{P(X)} \quad (3)$$

3.1. Przykłady zastosowań

Jednym z aktualnie stosowanych przykładów zastosowań rekomendacji (ang. *collaborative recommendation*) jest *Google News* [4], który jest zbiorem kompletnych i aktualnych wiadomości zebranych ze źródeł na całym świecie. Proces wyświetlania odbywa się w sposób spersonalizowany dla zalogowanych użytkowników. Personalizacja opiera się na historii kliknięcia aktywnego użytkownika i historii większej społeczności, przy czym każde kliknięcie jest interpretowane jako pozytywna ocena.

W przeszłości istniało wiele innych systemów opartych o wspólną współpracę, opracowanych w środowisku akademickim oraz w przemyśle. *Grundy* [11] był pierwszym systemem rekomendacyjnym do modelowania użytkowników poprzez mechanizm, który był wykorzystywany do budowy poszczególnych modeli użytkowników. Ostatecznie system ten zalecał odpowiednie książki. *GroupLens*, *Video Recommender* i *Ringo* [11] były pierwszymi systemami do wykorzystania wspólnych algorytmów filtrowania w celu zautomatyzowania prognozy. Innym przykładem systemu rekomendacyjnego, opartego o wspólną współpracę, jest system *Amazon.com* [11].

Filtrowanie w medycynie realizowane w ramach współpracy wykorzystywane jest głównie przez organizacje komercyjne, w celu prognozowania preferencji użytkowników, co często jest istotne w kontekście oceny pacjentów. Ostatnio prowadzone są badania wykorzystujące systemy rekomendacyjne w zakresie opieki medycznej. Celem tych badań jest doradztwo lub przeprowadzanie konsultacji, co odbywa się na podstawie dokumentacji medycznej pacjentów z podobnymi schorzeniami. W związku

z tym koniecznym staje się rozwijanie systemów rekomendacyjnych do zastosowań medycznych.

Systemy rekomendacyjne w medycynie charakteryzują się następującymi właściwościami:

- pacjenci są identyfikowani z użytkownikami,
- wzory zawierające dane dotyczące historii choroby i wyników badań są identyfikowane za pomocą profili użytkowników,
- pojęcie podobieństwa jest stosowane w stosunku do pacjentów, jak i użytkowników,
- diagnozy pacjentów są identyfikowane z opiniami użytkowników.

W pracy [8] przedstawiono rekomendację w oparciu o wspólną filtrację (ang. *collaborative filtering*), jako metodę wyznaczania ryzyka klinicznej stratyfikacji. System wyszukuje podobieństwa, jednocześnie między poszczególnymi obiektami oraz przedmiotami. Prezentowane podejście polega na dopasowaniu pacjentów, jak również ich charakterystyk, do powikłań sercowo-naczyniowych. Odwołuje się ono do prognozowania ocen użytkowników dotyczących elementów oraz przewidywania ryzyka do powikłań. W [8] oceniane jest ryzyko poprzez porównanie nowych pacjentów do istniejących w bazie historii chorób, a także poprzez porównanie wyników zainteresowania innych efektów lub cech pacjenta w zbiorze danych (z wykorzystaniem informacji dotyczących sąsiednich pacjentów oraz cech klinicznych).

Przykłady przedstawione w pracy [8] są nadal stosowane w zawężonym zakresie domen: najbardziej popularne są filmy i książki, zatem większość algorytmów działa poprawnie tylko w tego typu zestawach danych.

4. Rekomendacja oparta na wiedzy

W podejściu rekomendacji opartej na wiedzy (ang. *content-based recommendation*) Systemy bazują na specyficznej wiedzy, o tym, jak cechy produktów mogą spełniać wymagania użytkowników. W tej metodzie wyznacza się całkowitą użyteczność danego produktu dla określonego użytkownika. Metoda polega głównie na wyszukiwaniu i filtrowaniu informacji. W rozróżnianiu elementów istotnych od nieistotnych, systemy wykorzystują informacje pochodzące z treści danej pozycji. Próbują one powiązać jej zawartość z profilem użytkownika. System uczy się rekomendować przedmioty podobne do tych, które użytkownik „lubił" w przeszłości. Można zatem powiedzieć, że przewidywanie nie jest oparte na podobieństwie między użytkownikami, lecz na funkcji podobieństwa pomiędzy elementami. Innymi słowy, systemy realizujące podejście rekomendacji opartej na wiedzy analizują zbiór elementów wcześniej ocenionych przez użytkownika, następnie budują model lub profil zainteresowań użytkownika na podstawie cech obiektów, które zostały ocenione przez daną osobę [11, 12]. Powyższe podejście sprawia, że opisywany system rekomendacyjny stanowi doskonały sposób personalizowania treści dla użytkowników lub jako narzędzie służące do modelowania użytkownika. Prognozowanie opiera się na dopasowywaniu nowych treści elementu do atrybutów profilu użytkownika.

4.1. Struktura danych i algorytmy

Zawartość reprezentacyjnych pozycji można opisać w najprostszy sposób poprzez listę funkcji, często nazywanych atrybutami, cechami lub profilami. Dla przykładu, książki można opisać używając cech charakterystycznych, takich jak tytuł, gatunek, autor, typ, cena (odpowiednio dla pacjentów mamy m.in. dane demograficzne, ocena opieki medycznej). Użytkownicy powinni zostać poproszeni o ocenę zestawu pozycji. Opierając się na tej istotnej informacji, systemy rekomendacyjne danych oparte są na ocenie, w jakim stopniu pozycja dotychczas niewybrana jest zbliżona do pozycji, którą lubił aktywny użytkownik w przeszłości. W tym celu używane są różne środki podobieństwa, w zależności od obszaru problemowego [3, 13].

Wektory Booleana (ang. *Boolean vectors*). Systemy oparte na treści zostały pierwotnie opracowane do filtrowania i rekomendowania elementów tekstowych, takich jak wiadomości e-mail. Jest to charakterystyczne dla reprezentacji tego typu elementów w postaci listy słów, które pojawiają się w dokumencie i mogą być zakodowane w różny sposób. Ponad to może ona stanowić zbiór wszystkich słów, które pojawiają się w dokumentach, gdzie "1" oznacza, że słowo pojawia się w dokumencie, natomiast "0" determinuje brak danego słowa (logiczna reprezentacja wektora). Odpowiednio, profil użytkownika może być opisany za pomocą listy, zaś dopasowania można dokonać poprzez pomiar nakładania zainteresowania i zawartości dokumentu [13]

Podobieństwo oparte na miarach podobieństwa dla przestrzeni wektora (ang. *vector-space*) opieraja się na następujących technikach [2]:

- metoda najbliższych sąsiadów (KNN),
- metoda sprzężenia zwrotnego, opracowana pod koniec 1960 roku przez Rocchio dla SMART System - technika, która pomaga użytkownikom stopniowo udoskonalać zbiór zapytań na podstawie wcześniejszych wyników wyszukiwania.

Metody klasyfikacji, tworzące kategorię wtórnego elementu filtrującego, obejmują:

- metody probabilistyczne (np. klasyfikator bayesowski),
- klasyfikatory liniowe (np. maszyny wektorów nośnych),
- drzewa decyzyjne (np. ID3),
- indukcja reguł (np. zbliżona do ekstrakcji reguł decyzyjnych, zbudowana na podstawie algorytmu RIPPER).

Ostatnia metoda, oparta na indukcji reguł, ma dwie zalety w stosunku do innych metod uczenia. Po pierwsze, reguły mogą stanowić podstawę generowania wyjaśnień dla rekomendacji systemu. Dodatkowo w tych modelach mogą być wprowadzane domeny.

4.2. Przykłady aplikacji

Systemy oparte na zawartości są rzadko spotykane w środowiskach komercyjnych. Większość z nich z powodzeniem została opracowana w środowiskach akademickich [3].

Internetowe systemy rekomendacyjne. *Letizia* stanowi rozszerzenie przeglądarki internetowej, która śledzi zachowanie użytkownika i na tej podstawie tworzy spersonalizowany model. Dodatkowo zbiera informacje użytkownika, które w przyszłości

mogą służyć do wnioskowania jego preferencji. *Watcher Personal Web* jest rozwiązaniem, które uczy się preferencji użytkownika poprzez śledzenie stron, które odwiedza. Odwiedzone strony internetowe są przetwarzane jako pozytywne przykłady, zaś te których nie odwiedził jako negatywne. *Amalthe*a stanowi rozwiązanie, które dodatkowo implementuje technologię agenta, wykonując filtrowanie informacji. *Syskill & Webert* to system, który ma wbudowanych asystentów przeglądania stron internetowych. Wykorzystywana jest tutaj przeszłość w celu przewidywania czy użytkownik może być zainteresowany odnośnikami znajdującymi się na danej stronie internetowej.

Rekomendacja książek, filmów, muzyki. System LIBRA poleca książki, wykorzystując opisy w sklepie Amazon.com wykorzystując za naiwny klasyfikator Bayesa do kategoryzacji tekstu. *CiteSeer* pomaga użytkownikom w poszukiwaniu literatury naukowej, wykorzystując cytaty w innych pozycjach. INTIMATE to system rekomendacji filmowej, która wykorzystuje techniki kategoryzacji tekstu w oparciu o informacje uzyskane z bazy Internet Movie Database. Użytkownicy tego systemu są proszeni o ocenę filmów, co umożliwia przewidywanie. W obszarze aplikacji rekomendacji muzycznych, najbardziej zauważalnym systemem jest Pandora oparta na treści. Wiele systemów rekomendacyjnych w opisywanym obszarze zastosowań opierają się na współpracy.

Opieka zdrowotna. W pracy [5] przedstawiono podejście oparte na zawartości odpowiedniej dla celów budowania medycznego systemu rekomendacji. Opisana metoda sprawdza poszczególne profile przez różnych użytkowników, co może pociągnąć za sobą zbyt duże zagrożenie dla zachowania bezpieczeństwa danych osobowych (dane osobowe pacjentów muszą być traktowane jako poufne).

Innym przykładem jest *Personal Health Explorer*, zaproponowany przez Morrell i Kerschberg (2012). Stanowi on semantyczny system rekomendacji zdrowotnej, którego działanie opiera się na pobieraniu zasobów internetowych związanych z osobistymi kartami zdrowia pacjenta.

Popularną formą rekomendacji są również portale internetowe dotyczące informacji medycznych, na przykład objawów i chorób. Jeśli użytkownik takiego portalu ma konto, RS mogą zapewnić indywidualne dopasowanie informacji zdrowotnych, istotnych z punktu widzenia użytkownika. Systemy tego typu kierowane są do ludzi niezwiązanych z branżą medyczną.

5. Podsumowanie

Gwałtowny rozwój technologii informacyjnych przyczynił się do powstania inteligentnych systemów rekomendacji. Najnowsze badania wskazują, że metody rekomendacji skutecznie przyczynia się do wzrostu sprzedaży w e-sklepach. Metoda ta odnajduje również zastosowanie w innych dziedzinach, m.in. w medycynie. W związku z tym istotnym staje się zagadnienie badania oraz rozwijania metod rekomendacyjnych.

6. Uwagi ogólne

Badania przeprowadzono w ramach projektu MB/MW/8/2016 i sfinansowano ze środków MNiSW

Literatura

1. Adomavicius G., Bockstedt J., Curley S., Zhang J. *Do Recommender Systems Manipulate Consumer Preferences?*, University of Minnesota (2011).
2. Dardzinska A. *Action Rules Mining* Springer (2013).
3. Paul B Kantor, Lior Rokach, Francesco Ricci, Bracha Shapira. *Recommender systems handbook*, Springer, (2011).
4. Jannac D., Zanker M., Alexander Felfernig, Gerhard Friedrich. *Recommender systems: an introduction*, Cambridge University Press, (2010).
5. Wiesner M, Pfeifer D., *Health recommender systems: Concepts, requirements, technical basics and challenges*. International journal of environmental research and public health, 11(3), (2014) , s. 2580–2607.
6. Sodsee S., Komkhao M. *Evidence-based medical recommender systems: A review*, International Journal of Information Processing & Management, 4(6), (2013).
7. Ekstrand M., J. Riedl, J. Konstan: Col*laborative Filtering Recommender Systems*, Now Publishers Inc, Hanover (2011).
8. Shahzaib Hassan, Zeeshan Syed. *From netflix to heart attacks: collaborative filtering in medical datasets*, Proceedings of the 1st ACM International Health Informatics Symposium, ACM, (2010) s. 128–134.
9. Schafer J. B., D. Frankowski, J. Herlocker, S. Sen: *Collaborative Filtering Recommender Systems*, (2007).
10. Jannach D., Zanker M., Felfernig A., *Recommender systems: an introduction*, Cambridge University Press, (2010).
11. Adomavicius G., Tuzhilin A. *Toward the next generation of recommender systems*: ‚A survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6), (2005), s.734–749.
12. Dell'Aglio D., I. Celino, D. Cerizza, *Anatomy of a Semantic Web-enabled Knowledge-based Recommender System*, Politecnico of Milano, (2010) .
13. Gudder S., Latrémolière F. Boolean inner-product spaces and Boolean matrices, Linear Algebra and its Applications, Volume 431, Issues 1–2, 1 July 2009, s 274-296.

Inteligentne systemy rekomendacyjne i ich zastosowanie

Streszczenie

Obecnie Internet oraz systemy informacyjne stają się nieodłączną częścią życia codziennego. Trend używania zaawansowanych systemów rekomendacji wciąż rośnie w różnych dziedzinach. Systemy rekomendacji można określić jako zbiór technik oraz narzędzi dostarczających użytkownikom sugestie na temat interesujących ich produktów lub usług. W pracy przedstawiono syntetyczny przegląd istniejących inteligentnych systemów rekomendacji. Szeroki zakres dziedzin, w którym wspomniane metody są stosowane świadczy o ogromnej potrzebie implementacji algorytmów rekomendacyjnych do innych zastosowań, także medycznych.

Słowa kluczowe: system rekomendacyjny, rekomendacja, systemy inteligentne

Intelligent recommendation systems and their application

Abstract

Today, the Internet and information systems become an integral part of everyday life. The trend of using advanced recommendation systems is still growing in various areas. Recommendation systems can be defined as a set of techniques and tools that provide users with suggestions about their products or services. The paper presents a synthetic overview of existing intelligent recommendation systems. A wide range of fields where these methods are used show the need to implement recommendation algorithms for other application, including medical systems.

Keywords: recommender system, recommendation, intelligent systems

Magdalena Fryc[1]

Wykorzystanie środowiska Matlab w diagnostyce onkologicznej

1. Wstęp

Rak płuc jest jedną z najbardziej rozwijających się chorób w Polsce i na świecie, a zwłaszcza w krajach rozwijających się. Choroba ta dotyka coraz więcej osób zarówno mężczyzn jak i kobiet, wśród których większość to nałogowi palacze tytoniu oraz osoby narażone na ekspozycję szkodliwych środków chemicznych tj. nikiel, azbest czy radon. Rak płuc należy do tych nielicznych, praktycznie nieuleczalnych chorób. Niewielki odsetek chorych na raka płuc zostaje uleczonych, jednakże dzięki postępowi medycyny oraz takich nauk jak fizyka medyczna czy też inżynieria biomedyczna możliwe jest znaczne przedłużenie życia osobom dotkniętym tą chorobą.

Istnieje wiele możliwości leczenia raka płuc w zależności od jego rodzaju, czyli podziału histopatologicznego raka płuc. Sposoby leczenia są wszelakie, od chemioterapii poprzez bronchoskopię, czy też radioterapię, aż do leczenia chirurgicznego, które w nielicznych przypadkach pozwala na całkowite usunięcie nowotworu. Dzięki rozwojowi takich dziedzin nauki jak fizyka medyczna i inżynieria biomedyczna diagnostyka raka staję się coraz to prostsza i szybsza. Głównymi narzędziami do wykrycia raka płuc jest badanie rentgenowskie (RTG), tomografia komputerowa (TK), pozytronowa tomografia emisyjna połączona z tomografią komputerową (PET CT), ale i w szczególnych przypadkach rezonans magnetyczny (MRI) i właśnie temu ostatniemu należałoby poświecić szczególną uwagę ze względu na praktycznie całkowitą bezinwazyjność dla człowieka. Jeżeli chodzi o przypadek raka płuc, MRI nie pokazuje znaczącej wyższości nad badaniem TK klatki piersiowej, jednakże posiada wiele zalet i udoskonaleń, których inne urządzenia obrazowania medycznego nie posiadają. MRI uwidacznia doskonale nawet drobne zmiany nowotworowe na etapie wczesnego stadium choroby, których nie widać np. podczas badania TK. Dodatkowo MRI pozwala lepiej uwidocznić naciek nowotworu na naczynia krwionośne i struktury nerwowe. Dlatego też znajduje zastosowanie w obrazowaniu guzów Pancoasta znajdujących się w szczycie płuca w pobliżu wymienionych struktur. MRI znajduje również szeroki wachlarz zastosowań w szczególności przy obrazowaniu nowotworów mózgu, ale i innych chorób klatki piersiowej. Ponadto badanie MRI pozwala na zobrazowanie całego ciała ludzkiego podczas jednej, dosyć krótkiej sesji badania, co podczas innych badań tj. RTG jest niemożliwe, dlatego też MRI jest bardzo cenionym urządzeniem obrazowania medycznego, jeśli u chorego nastąpiły przerzuty z jednego organu do innych. Oprócz szerokiego pola zastosowań w diagnostyce różnych chorób,

[1] Magdalena_fryc@o2.pl, Uniwersytet Śląski, wydział Informatyki i Nauki o Materiałach, Uniwersytet Śląski, www.wiinom.us.edu.pl

MRI nie wykorzystuje promieni podczerwonych, przez co nie naraża badanego ani obsługi na promieniowanie jonizujące, a stosowane środki kontrastowe podczas badania rzadziej wywołują reakcje alergiczne niż jod stosowany w konwencjonalnych metodach rentgenowskich TK. MRI obrazuje praktycznie cały ludki organizm: tkanki miękkie, układ nerwowy, nieprawidłowości które mogą być zasłonięte przez kości w innych metodach obrazowania, ale i drogi żółciowe, naczynia krwionośne i układ moczowy, co w przypadku innych metod jest bardzo inwazyjne oraz wymaga podawania dodatkowo kontrastów, natomiast w trakcie badania rezonansem magnetycznym kontrast stanowią naturalne płyny tj. żółć, krew oraz mocz, ale przede wszystkim jest nieocenionym narzędziem w zakresie wczesnej diagnostyki i oceny wielu zmian ogniskowych i nowotworów, ponieważ identyfikuje zmiany nieuchwytne w innych metodach obrazowania. Dodatkowym atutem MRI jest bezinwazyjna diagnostyka kobiet w ciąży oraz płodu.

2. Cel pracy oraz główne założenia

Celem poniższej publikacji jest propozycja metody komputerowego przetwarzania i analizy obrazów, która ma wspomóc diagnostykę zmian nowotworowych płuc przez onkologów w zakresie dostosowania odpowiedniego leczenia dla poszczególnych pacjentów.

Obrazami poddawanymi przetwarzaniu i analizie są sekwencje obrazów dla różnych płaszczyzn płuc, uzyskanych z rezonansu magnetycznego.

W szczególności, zostały rozważone następujące zagadnienia :

- obiektywna ocena możliwości zastosowania komputerowej analizy obrazów z rezonansu magnetycznego dla zmian nowotworowych płuc,
- segmentacja obszaru zmiany nowotworowej płuca z wykorzystaniem własnych propozycji wydobycia zmian poprzez binaryzację obrazu oraz wyodrębnienie interesujących fragmentów,
- obliczenie odległości pomiędzy dwoma najbardziej odległymi od siebie punktami zmiany nowotworowej,
- obliczenie pola powierzchni wyodrębnionych zmian nowotworowych płuca oraz objętości końcowej zmiany nowotworowej płuca dla każdej z wczytanych serii obrazów,
- zobrazowanie na wykresach wyznaczonych długości zmian nowotworowych w najszerszym miejscu,
- zobrazowanie na wykresach zmian wielkości pola powierzchni zmian nowotworowych na kolejnych obrazach serii wczytanych obrazów,
- automatyczne zapisanie wyników do arkusza kalkulacyjnego wraz z obrazami wyodrębnionych zmian nowotworowych,
- postawienie wstępnej diagnozy oraz dopasowanie odpowiedniego leczenia dla przykładowego pacjenta.

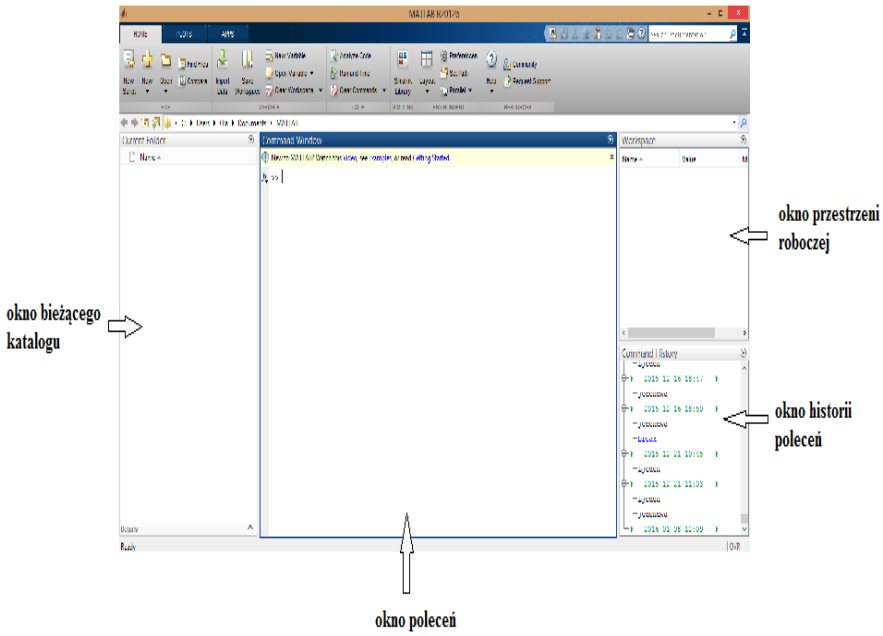
3. Matlab – co to jest i do czego służy ?

Nazwa Matlab pochodzi od MATrix i LABoratory. Program Matlab jest pakietem o uniwersalnym środowisku umożliwiającym przeprowadzenie złożonych obliczeń naukowo-technicznych, poprzez dostęp do różnorodnych i efektywnych algorytmów obliczeniowych. Pozwala na wizualizację uzyskanych wyników w postaci różnych wykresów oraz grafiki trójwymiarowej. Program Matlab to interaktywne środowisko oferujące użytkownikowi szeroki wachlarz wbudowanych technicznych funkcji obliczeniowych, graficznych oraz animacyjnych. Ponadto program pozwala na samodzielną rozbudowę zakresu zastosowań poprzez tworzenie własnych skryptów i programów w wybranym języku programowania wysokiego poziomu. Użytkownik ma dostęp do niezawodnych algorytmów matematyki stosowanej oraz do licznych modułów rozszerzeń (Simulink, biblioteki toolbox, blackset, Real Time Workshop, Stateflow). Podstawowymi typami danych w programie Matlab są tablice o elementach zespolonych lub rzeczywistych oraz struktury i obiekty. Wszystkie zmienne zdefiniowane przez użytkownika do programu są przechowywane w wygodny sposób w przestrzeni roboczej. Program jest przyjazny dla użytkownika, ponieważ automatycznie rozpoznaje typ zmiennej oraz informuje o każdorazowych błędach. Uniwersalność programu pozwala na jego szerokie zastosowanie, które można znaleźć między innymi w medycynie, energetyce, transporcie, przemyśle, fotografii i innych dziedzinach nauki. Tak szerokie zastosowanie programu Matlab wynika z jego wielu zalet i funkcji jakie posiada, a zwłaszcza umożliwia szybkie uzyskanie rezultatów złożonych obliczeń oraz wizualizację wyników[1,2]. Ponadto pakiet Matlab umożliwia analizę zdjęć nie tylko pozyskanych z aparatów fotograficznych, ale i z rezonansu magnetycznego oraz innych urządzeń medycznych tj. TK, RTG czy USG. Dlatego też, wykonanie celu pracy, którym było przedstawienie analizy zmian nowotworowych w stosunku do objętości płuc, było możliwe dzięki funkcjonalności pakietu Matlab.

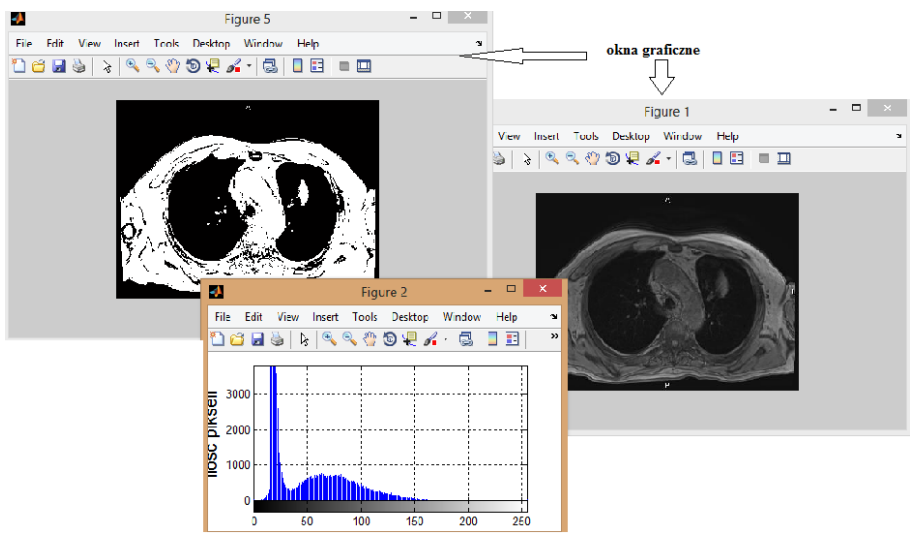
Program Matlab obsługiwać można przy pomocy trzech głównych okien[1]:

- Pulpit Matlab – pierwsze okno programu (Rysunek 1), otwierające się po uruchomieniu Matlab'a i składające się z czterech podokien:
  - Okno poleceń (Comand Window) –główne okno programu, gdzie wyświelny jest znak zachęty „>>", oznaczający możliwość wprowadzenia poleceń przez użytkownika,
  - Okno bieżącego katalogu (Current Folder) – okno, w którym wyświetlane są wszystkie pliki zawarte w bieżącym katalogu,
  - Okno przestrzeni roboczej (Workspace) – okno, w którym przechowywane są wszystkie utworzone w danym folderze zmienne oraz ich typ i rozmiar;
  - Okno historii poleceń (Command History) – okno magazynujące wszystkie wpisane przez użytkownika w oknie poleceń polecenia,
  - Okno graficzne – okno wyświetlające wszystkie rezultaty poleceń graficznych zdefiniowanych przez użytkownika w oknie poleceń (Rysunek 2). Użytkownik może otworzyć praktycznie nieskończoną ilość okien graficznych,

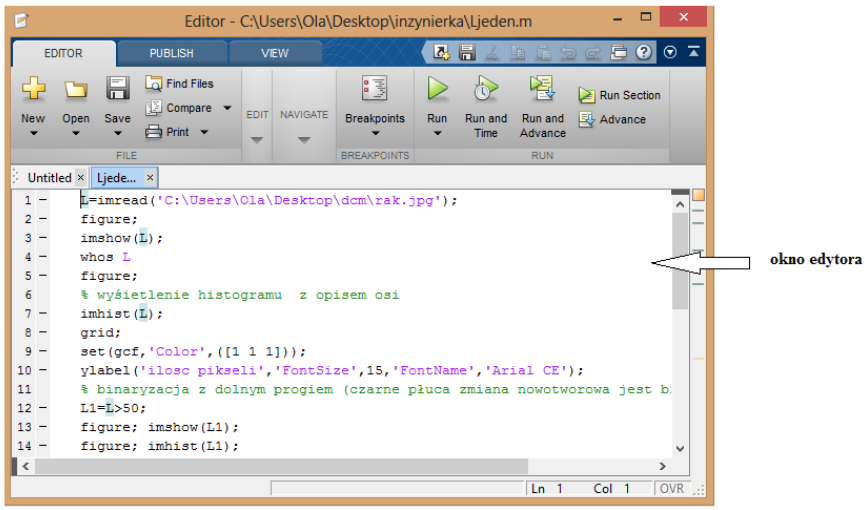
- Okno edytora – okno umożliwiające otwieranie, tworzenie, edytowanie oraz zapisywanie własnych programów przez użytkownika (Rysunek 3). Są to tzw. M-pliki [1].



Rysunek 1 Pulpit Matlab'a z podoknami [opracowanie własne]



Rysunek 2 okna graficzne programu Matlab [opracowanie własne]



Rysunek 3 Okno edytora programu Matlab [opracowanie własne]

4. Klasyfikacja tnm niedrobnokomórkowego raka płuca

Stopień zaawansowania raka niedrobnokoórkowego płuca ustala się poprzez międzynarodową klasyfikację TNM, w której bierze się pod uwagę trzy cechy :

- T–cecha guza pierwotnego w płucu,
- N–cecha oceniająca obecność przerzutów w węzłach chłonnych,
- M–cecha oceniająca obecność przerzutów odległych [9].

(Tabela 1) przedstawiona poniżej obrazuje analizę poszczególnych cech.

Tabela 1 Klasyfikacja raka niedrobnokomórkowego płuca TNM

T (tumor) - guz pierwotny nowotworu	
Tx	obecność guza udowodniona na podstawie obecnych komórek nowotworowych w wydzielinie oskrzelowej, ale bez cech guza w badaniach radiologicznych klatki piersiowej i bronchoskopii
T0	nie stwierdza się guza pierwotnego
T1	guz o średnicy poniżej 3 cm, otoczony tkanką płucną lub opłucną płucną bez naciekania oskrzela głównego
T2	guz mający przynajmniej jedną z następujących cech: średnica większa niż 3 cm, zajęcie oskrzela głównego w odległości nie mniejszej niż 2 cm od ostrogi głównej, naciekanie opłucnej, towarzysząca niedodma lub zapalenie płuc dochodzące do wnęki (ale niezajmujące całego płuca)
T3	guz każdej wielkości z naciekaniem następujących struktur anatomicznych: ściana klatki piersiowej, przepona, opłucna osierdziowa, opłucna śródpiersiowa, osierdzie, nerw przeponowy, guz z towarzyszącą niedodmą lub zapaleniem całego płuca
T4	guz każdej wielkości naciekający jedną ze struktur: śródpiersie, serce, wielkie naczynia, tchawica, przełyk, ostroga główna, trzony kręgów, nerw krtaniowy wsteczny, guz z wysiękiem opłucnowym lub osierdziowym, guz z oddzielnymi guzkami satelitarnymi w obrębie tego samego płata

N (lymph nodes) - stan regionalnych węzłów chłonnych	
Nx	nie można ocenić regionalnych węzłów chłonnych
N0	brak przerzutów do regionalnych węzłów chłonnych
N1	obecne przerzuty w węzłach okołooskrzelowych (międzypłatowych, płatowych, segmentarnych) lub/i wnękowych po stronie guza
N2	obecne przerzuty w węzłach chłonnych śródpiersia po stronie guza (górne śródpiersiowe, górne okołotchawicze, przedtchawicze, pozatchawicze, dolne okołotchawicze, aortalne, podaortalne, okołoaortalne, okołoprzełykowe, okolicy więzadła płucnego) lub/i w węzłach poniżej rozwidlenia tchawicy
N3	obecne przerzuty w węzłach chłonnych śródpiersia lub wnęki po stronie przeciwnej lub w węzłach chłonnych pod mięśniem pochyłym przednim szyi lub w węzłach chłonnych nadobojczykowych po stronie guza lub po stronie przeciwnej
M (metastases) - przerzuty odległe	
Mx	nie można ocenić przerzutów odległych
M0	brak przerzutów odległych
M1	stwierdza się przerzuty odległe (w tym inne ogniska nowotworu w innym płacie po tej samej stronie lub po stronie przeciwnej)

Źródło: [3]

Kolejnym krokiem do oceny stopnia zaawansowania klinicznego raka niedrobnokomórkowego płuca jest analiza zależności poszczególnych cech TNM, co ma znaczenie w wyborze metody leczenia choroby, a szczególnie do kwalifikacji pacjenta do leczenia operacyjnego [9]. Zależność stopnia poszczególnych cech w klasyfikacji TNM przedstawiono w (Tabela 2).

Tabela 2 Stopień zaawansowania klinicznego raka niedrobnokomórkowego płuca

Stopień zaawansowania raka	Cecha klasyfikacji TNM		
	T	N	M
Rak utajony	x	0	0
0	is	0	0
IA	1	0	0
IB	2	0	0
IIA	1	1	0
IIB	2 3	1 0	0 0
IIIA	1 2 3 3	2 2 1 2	0 0 0 0
IIIB	Każdy 4	3 Każdy	0 0
IV	Każdy	Każdy	1

Źródło: [3]

5. Leczenie

Dobór leczenia niedrobnokomórkowego raka płuc zależy od stopnia zaawansowania zależnego od klasyfikacji TNM co demonstruje (Tabela 3).

Tabela 3 Metody leczenia raka niedrobnokomórkowego płuca w zależności od stopnia jego zaawansowania

Stopień zaawansowania raka	Metoda leczenia
I II	Leczenie operacyjne i chemioterapia po operacji
IIIA	Leczenie operacyjne oraz ewentualnie chemioterapia i/lub radioterapia przed operacją
IIIB	Radioterapia ewentualnie połączona z chemioterapią
IV	Chemioterapia paliatywna lub leczenie objawowe wydolności narządów organizmu człowieka i ogólnego stanu pacjenta

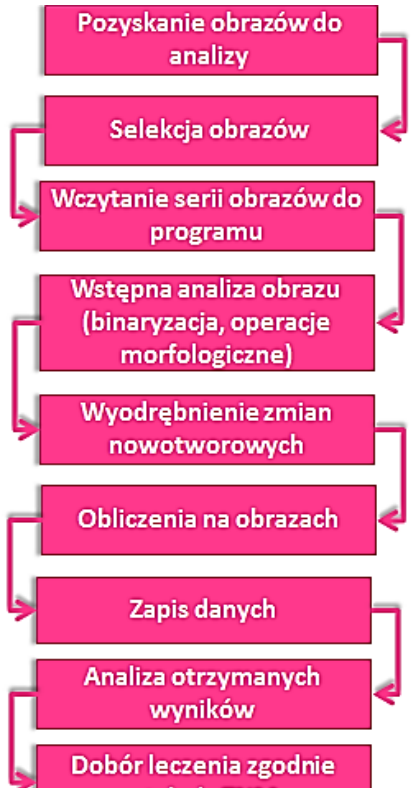
Źródło: [3]

6. Tworzenia programu za pomocą pakietu Matlab

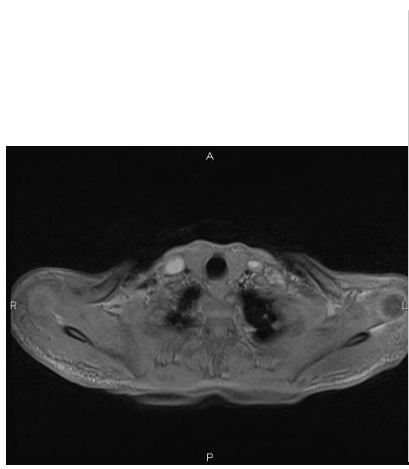
W ramach przedstawionej pracy za pomocą pakietu Matlab stworzono program służący do diagnostyki i analizy zmian nowotworowych płuc. Posługując się procedurą przedstawioną na Rysunku 2.

6.1. Selekcja obrazów do analizy

W każdej serii otrzymanych zdjęć do analizy została przeprowadzona selekcja, w której odrzucono obrazy na których zmiany nowotworowe nie były widoczne (Rysunek 3) oraz te obrazy, na których pomimo odpowiednio przeprowadzonej binaryzacji oraz operacji morfologicznych, zmiana nowotworowa była niemożliwa do wyodrębnienia. W efekcie otrzymywano wyodrębniony znacznie większy fragment (Rysunek 4). Drastyczna zmiana progu binaryzacji powodowała, iż pole powierzchni znacznie się zmniejszało, a na wielu obrazach zmiana nowotworowa całkowicie znikała, co powodowało błędną diagnozę, dlatego też zdecydowano, iż lepszym wyborem będzie odrzucenie kilku obrazów niż zmiana progu binaryzacji, przy którym ilość odrzuconych obrazów była by jeszcze większa, jednocześnie zmniejszając wynik końcowy otrzymanej objętości zmiany nowotworowej.



Rysunek 2 Proces postępowania podczas analizy i diagnostyki zmian nowotworowych płuc w środowisku Matlab [opracowanie własne]



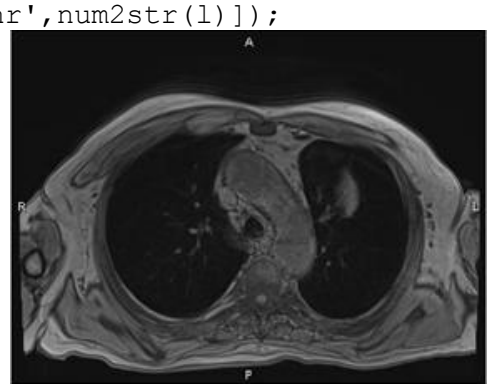
Rysunek 3 Przykładowy obrazy bez widocznych zmian nowotworowych [opracowanie własne]

Rysunek 4 Obraz z zaznaczoną zmianą nowotworową po nieudanej binaryzacji [opracowanie własne]

6.2. Wczytanie serii obrazów, binaryzacja oraz operacje morfologiczne na obrazach binarnych

Poniżej przedstawiony kod programu został stworzony w celu ułatwienia komputerowej diagnostyki i analizy zmian nowotworowych płuc u pacjentów, którzy zostali poddani badaniu MRI. Na początku seria zdjęć zostaje wczytana w pętli i wyświetlona w kolejnych oknach graficznych (Rysunek 5). Każdy wczytany z serii obraz został odpowiednio opisany oraz ponumerowany automatycznie poprzez wywołanie zdefiniowanej na początku pętli zmiennej 'i'. Co zobrazowane zostało na poniższym kodzie programu:

```
pam=[];
for i = 1:1:7
L=imread(['C:\ścieżka_dostępu\dcm\',mat2str(i),'.jpg']);
figure;
imshow(L);
title(['obraz nr',num2str(l)]);
```

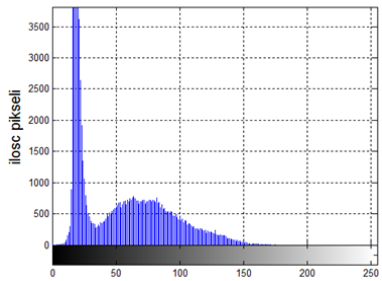


Rysunek 1 wczytanie obrazu ze zmianą nowotworową płuc przekrój w płaszczyźnie poprzecznej [opracowanie własne]

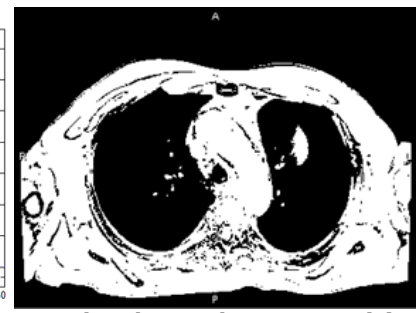
Następnie dla każdego zdjęcia wyświetlony zostaje histogram w celu dobrania neutralnego progu binaryzacji(Rysunek 6). Po otrzymanych histogramach dobrano jeden próg binaryzacji odpowiedni dla wszystkich wczytanych obrazów danej serii. Przykładowo dla pierwszej wczytanej serii obrazów próg binaryzacji wynosi 50 [X] w skali szarości lub po przejściu na double 0.25 [X], ponieważ obowiązuje wówczas skala 0-1 (Rysunek 7). Dodatkowo wyświetlono histogram po binaryzacji (Rysunek 8), w celu upewnienia się, czy binaryzacja została przeprowadzona prawidłowo. Wszystkie powyższe operacje zostały przedstawione na poniższym kodzie:

```
  L=double(L)/255;
  figure;
  imhist(L);
  grid;
  set(gcf,'Color',([1 1 1]));
  ylabel('ilosc  pikseli','FontSize',15,'FontName','Arial
CE');
  L1=L>0.25;
```

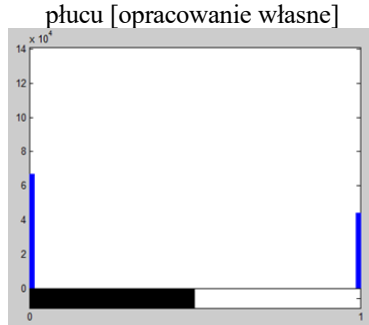
```
figure;
imshow(L1);
figure;
imhist(L1);
```



Rysunek 6 Histogram wczytanego obrazu [opracowanie własne]



Rysunek 7 obraz po binaryzacji z dolnym progiem z uwidocznioną zmianą nowo- tworową w lewym płucu [opracowanie własne]

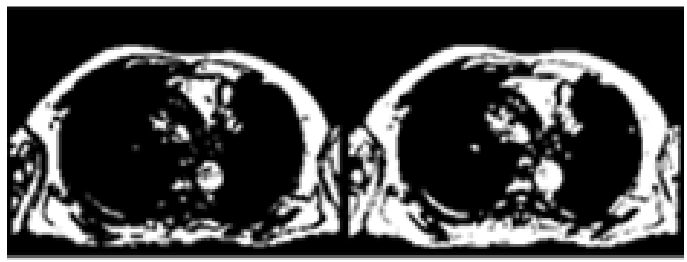


Rysunek 8 Histogram obrazu po binaryzacji [opracowanie własne]

Najlepszą binaryzacją okazała się binaryzacja z dolnym progiem, po której otrzymaliśmy oczekiwany wynik tzn. obrazy binarne z uwidocznioną odpowiednio zmianą nowotworową oraz z odpowiednim zarysem płuc. Ze względu na zamiar otrzymania obrazów binarnych, na których płuca mają być obszarem o wartości 0, czyli mają być czarne. Tym samym można stwierdzić, iż dolny próg binaryzacji powinien być wybierany każdorazowo, bez względu na to, jaką serię zdjęć wczytujemy w celu ich analizy.

Po poprawnej binaryzacji obrazów zostały przeprowadzone kolejne operacje morfologiczne (Rysunek 9), przedstawiona na poniższym kodzie programy:

```
BW2 = bwmorph(L1,'erode');
figure ;
imshow(BW2) ;
BW3 = bwmorph(BW2,'dilate');
    figure ;
    imshow(BW3) ;
bw=bwmorph(BW1,'open');
    figure ;
    imshow(bw) ;
 end
```

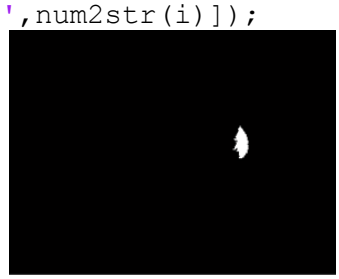


Rysunek 9 Operacje morfologiczne : erozja i dylatacja [opracowanie własne]

6.3. Wyodrębnianie zmian nowotworowych płuc.

Wyodrębnianie zmian nowotworowych płuc przeprowadzono w pętli. Zastosowana funkcja „ginput" pozwala na ręczne zaznaczenie na obrazie binarnym zmiany nowotworowej oraz dodatkowo na każdym z obrazów można zastosować ją wielokrotnie, jeśli analizowane zmiany w płucach są rozsiane, jak w przypadku raka drobnokomórkowego płuca, lub jeżeli wystąpiły by przerzuty do innych narządów lub węzłów chłonnych. Funkcja ta, pomimo iż jest czasochłonna dla użytkownika pozwala uniknąć błędu komputera, który mógłby ominąć którąś ze zmian lub zaznaczyć inną strukturę, która nie zaliczana jest do zmian nowotworowych. Poniższy kod odpowiada za wyodrębnianie zmian nowotworowych oraz wyświetlenie zaznaczonej zmiany w nowym oknie graficznym (Rysunek 10).

```
L5=bwlabel(BW3);
figure;
imshow(L5);
[x,y,k]=ginput(1);
nr=L5(round(y),round(x));
L6=L5==nr;
figure;
imshow(L6);
title(['zmiana nr',num2str(i)]);
```



Rysunek 10 Wyodrębniona zmiana nowotworowa w nowym oknie graficznym [opracowanie własne]

W celu zobrazowania poszczególnych etapów przetwarzania obrazów wyświetlamy je w odpowiedniej kolejności w nowym oknie graficznym (Rysunek 11), co pozwala na dodatkową analizę obrazu krok po kroku, aby upewnić się, iż każda operacja wykonana na obrazie została przeprowadzona prawidłowo oraz, czy nie wymaga dodatkowych poprawek. Poniższy kod pokazuje jak prawidłowo wyświetlić wszystkie

obrazy w jednym oknie graficznym wraz z opisem oraz numeracją kolejnych serii obrazów.

```
figure;
imshow([L,L1,BW2,BW3,L6]);
title(['1-obraz pierwotny 2-obraz po binaryzacji 3-obraz po erozji 4-obraz po dylatacji 3-zmiana nowotworowa',num2str(i)])
```

Rysunek 112 Okno graficzne z poszczególnymi etapami przetwarzania obrazu [opracowanie własne]

6.4. Obliczenia na obrazie

W celu dokonania analizy oraz postawienia jakiejkolwiek wstępnej diagnozy istotne jest dokonanie pewnych obliczeń na analizowanych obrazach. Ze względu na podział TNM nowotworów pierwszym parametrem, który zostanie obliczony jest długość zmiany nowotworowej w jej najszerszym miejscu. Następnie obliczone zostanie pole powierzchni każdej wyodrębnionej wcześniej zmiany nowotworowej płuca. Ostatnim krokiem będzie obliczenie objętości całkowitej zmiany nowotworowej płuca.

6.4.1. Obliczanie długości zmiany nowotworowej w najszerszym miejscu

Ze względu na nieregularny kształt zmian nowotworowych, długość każdej zmiany obliczamy w jej najszerszym miejscu, innymi słowy pomiędzy dwoma najbardziej oddalonymi od siebie punktami. W celu obliczenia tego parametru została stworzona dodatkowa pamięć, która przechowuje dane dotyczącego długości zmiany oraz pętla, w której została wprowadzona nowa zmienna ‘m’ wywołująca kąt obrotu dla każdej wyodrębnionej wcześniej zmiany nowotworowej. Dodatkowo cały wynik trzeba odpowiednio przemnożyć, aby uzyskać odpowiednie jednostki. W naszym przypadku oczekiwany wynik powinien być podany w milimetrach [mm], dlatego też, jednostka przez którą są reprezentowane nasze obrazy tzn. piksele [X] musi być zamieniona odpowiednio na milimetry poprzez pomnożenie danej sumy przez odpowiadającą 1 X jednostkę mm. Seria obrazów wykorzystywana w przedstawionej pracy była wykonana rezonansem magnetycznym, w którym 1x1x1 [X] odpowiada 1 wokselowi reprezentowanemu przez 3x3x3 [mm]. Poprawne obliczenie długości zmiany nowotworowej prezentuje poniższy kod:

```
pam1=[]
for m= 0:180
 L10=imrotate(L6,m);
 pam1=[pam1;[m sum(sum(L10,1)>1)]*3];
end
pam1=sortrows(pam1,-2);
pam1(1,2)
```

6.4.2. Pole powierzchni oraz objętość zmiany nowotworowej

Pole powierzchni jest parametrem prostym do wyznaczenia, który można wyznaczyć z dużą dokładnością. W celu wyznaczenia pola powierzchni trzeba zliczyć punkty fragmentu obrazu, którego pole nas interesuje. Następnie po ustaleniu rzeczywistej odległości jakiej odpowiadają wyznaczone punkty interesującego nas fragmentu, otrzymujemy wynik pomiaru w odpowiednich dla naszego pomiaru jednostkach. W celu uzyskania pola powierzchni w odpowiednich jednostkach [mm^2], konieczne jest przemnożenie otrzymanej wartości dwukrotnie przez 3. Wszystkie obliczone dane zapisane zostają do wspólnej pamięci utworzonej na początku programu. Operacje te prezentuje poniższy fragment programu:

```
bwarea(L6)
pam=[pam;[i,sum(sum(L6))*3*3,pam1(1,2)]];
```

W celu otrzymania końcowego wyniku tzn. objętości wyodrębnionej zmiany nowotworowej dla danej serii obrazów, trzeba pomnożyć odpowiednią wartość ponownie przez 3, w ten sposób otrzymujemy wynik w [mm^3]. Poniższy fragment kodu przedstawia poprawne obliczenie objętości zmiany nowotworowej płuca:

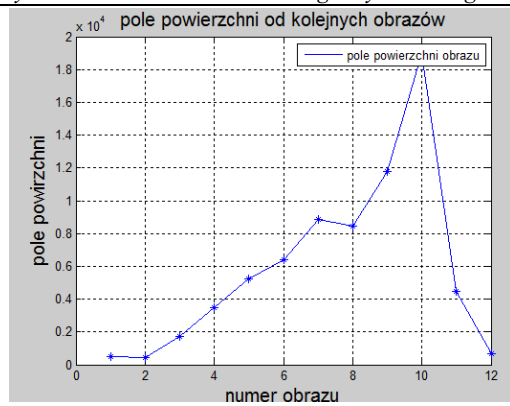
```
sum(pam(:,2)*3)
```

6.5. Tworzenie wykresów

6.5.1. Wykresy zależności pola powierzchni oraz długości zmian nowotworowych dla poszczególnych obrazów wczytanej serii.

Po obliczeniu pola powierzchni oraz długości poszczególnych zmian nowotworowych płuca wygenerowano wykresy, na których pokazano zmianę pola powierzchni dla kolejnych obrazów wczytanej serii jak i długości zmiany nowotworowej pomiędzy dwoma najbardziej oddalonymi od siebie punktami, w celu porównania wyników oraz przeprowadzenia dzięki nim dalszej analizy obrazów. Wybrano wykres liniowy otrzymywany dzięki wywołaniu funkcji ‘plot’ (Rysunek 12). Otrzymany wykres można modyfikować automatycznie poprzez wywołanie odpowiednich komend odpowiadających min. za wybór stylu linii, kolor, rozmiar, dodanie markerów oraz odpowiedni opis osi X i Y, nadanie tytułu oraz legendy, wybraną i zdefiniowaną w kodzie czcionką

```
Figure;
  plot(pam(:,1),pam(:,2));
  line (pam(:,1),pam(:,2),'linestyle','-','marker','*')
  xlabel ('numer obrazu','FontSize',15,'FontName','Arial
CE')
  ylabel                                              ('pole
powierzchni','FontSize',15,'FontName','Arial CE')
  title('pole   powierzchni   od   kolejnych   obrazów',
'FontSize', 15, 'FontName','Arial CE')
  grid on ;
  legend ('pole powierzchni obrazu')
```

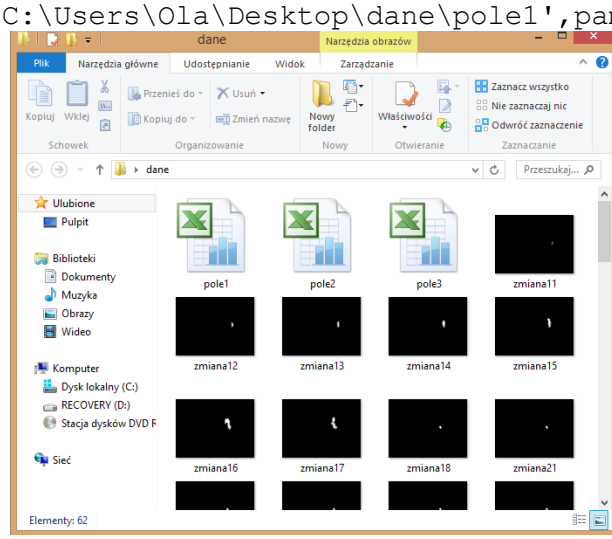


Rysunek 12 wykres pola powierzchni dla przykładowej serii obrazów [opracowanie własne]

6.6. Zapis sekwencji obrazów wyodrębnionej zmiany nowotworowej do folderu wraz z arkuszem kalkulacyjnym zawierającym dane obliczeniowe

Zapis sekwencji obrazów wyodrębnionych zmian nowotworowych płuca do nowego folderu (Rysunek 13) pozwala na późniejszą dokładną analizę interesujących zmian nowotworowych, bez konieczności ponownego wyodrębniania ich z pierwotnych obrazów. Jest to dużym ułatwieniem dla techników oraz lekarzy, ponieważ analizują już odpowiednio wyodrębnioną zmianę bez konieczności przeprowadzania ponownej obróbki obrazów. Dodatkowo zaimportowanie danych dotyczących długości zmiany nowotworowej oraz pola jej powierzchni do arkusza kalkulacyjnego jest kolejnym udogodnieniem, pozwalającym na analizowanie danych oraz tworzenie w arkuszu dowolnych wykresów i przeprowadzanie własnych analiz.

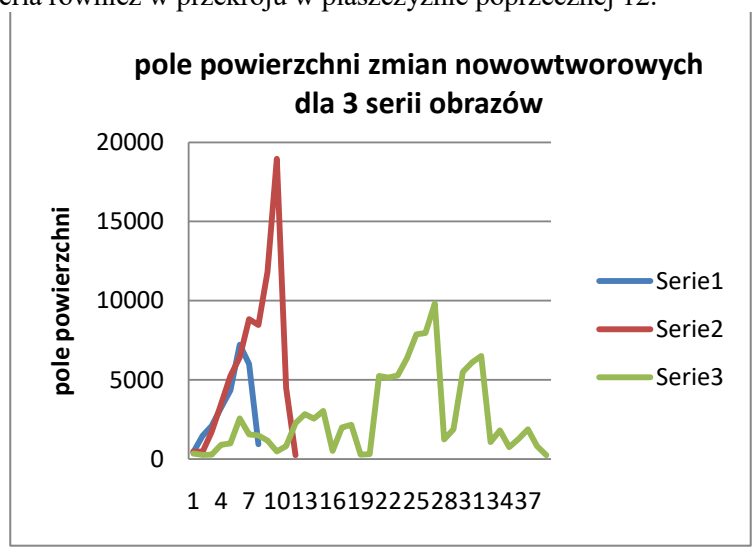
```
imwrite(L6,['C:\Users\Ola\Desktop\dane\zmiana1',mat2str
(i), '.jpg'])
xlswrite('C:\Users\Ola\Desktop\dane\pole1',pam);
```



Rysunek 3 Folder z arkuszami kalkulacyjnymi przechowującymi dane dotyczące pola powierzchni zmian nowotworowych oraz obrazy wyodrębnionych zmian nowotworowych [opracowanie własne]

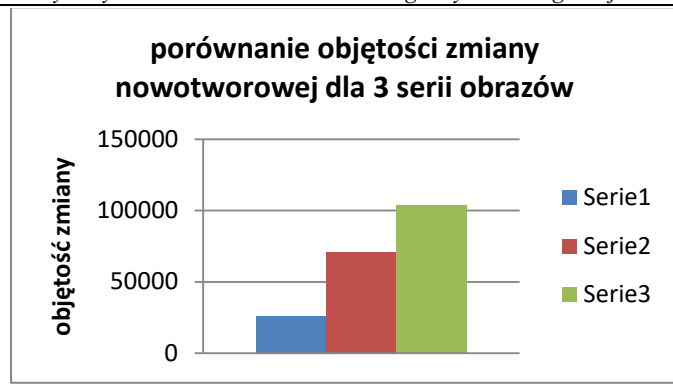
7. Analiza wyników

Po otrzymanych wynikach dla zmian nowotworowych u jednego pacjenta z trzech serii obrazów rezonansu magnetycznego dla różnych przekrojów ciała przeprowadzono analizę wyników pola powierzchni wyodrębnionych zmian dla wyselekcjonowanych obrazów, gdzie była widoczna zmiana nowotworowa oraz, dla których było możliwe wyodrębnienie jej w nowym oknie graficznym. Po zestawieniu wyników można stwierdzić, iż liczba obrazów poddanych analizie jest różna dla każdego z pacjentów. Wyniki długości pomiędzy dwoma najbardziej oddalonymi od siebie punktami, pola powierzchni jak i objętość końcowa zmiany nowotworowej różnią się znacznie pomiędzy analizowanymi seriami obrazów trzech przykładowych pacjentów, przy czym dla ostatniej serii obrazów wykonanych w przekroju płaszczyzny czołowej analizie poddana została największa liczba obrazów tzn. 39, natomiast pierwsza seria obrazów liczyła w przekroju w płaszczyźnie poprzecznej 8 zmian nowotworowych oraz druga seria również w przekroju w płaszczyźnie poprzecznej 12.



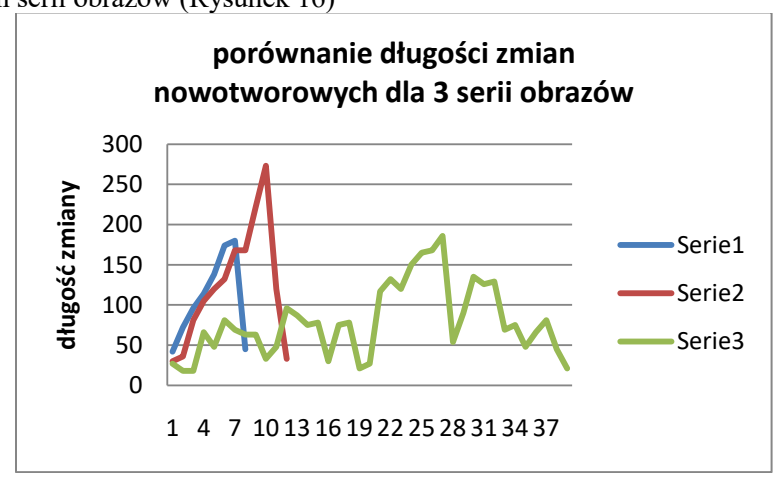
Rysunek 14 wykres porównawczy zmiany pola powierzchni dla poszczególnych serii obrazów [opracowanie własne]

Pomimo iż największe zanotowane pole powierzchni zmiany nowotworowej wystąpiło w drugiej serii wczytanych obrazów, to największa objętość końcowa wystąpiła w trzeciej wczytanej serii obrazów, gdzie obrazów poddanych analizie było najwięcej. Objętość końcową poszczególnych serii obrazów przedstawia (Rysunek 15).



Rysunek 4 wykres objętości zmiany nowotworowej dla poszczególnych serii obrazów [opracowanie własne]

W celu wydania wstępnej diagnozy niezbędne jest podanie parametru T z klasyfikacji TNM, który odpowiedzialny jest min. za średnicę zmiany nowotworowej, czyli odległość pomiędzy najbardziej oddalonymi między sobą dwoma punktami analizowanej zmiany nowotworowej. By wydać wstępną diagnozę przedstawiono wykres porównawczy zestawienia długości analizowanych zmian nowotworowych dla trzech wczytanych serii obrazów (Rysunek 16)



Rysunek 16 wykres porównania długości zmian nowotworowych poszczególnych serii obrazów [opracowanie własne]

Po przeanalizowaniu poszczególnych wykresów można stwierdzić, iż w serii drugiej gdzie zmiana nowotworowa osiągnęła największe pole powierzchni oraz długość. Po przeanalizowaniu danych dotyczących pola powierzchni objętości końcowej oraz długości zmian nowotworowych, można zauważyć iż jest znaczny związek między polem, powiechni, a długością zmian nowotworowych, natomiast objętość wynikająca z pola powierzchni wszystkich zmian nowotworowych danej serii nie wpływa na postawienie diagnozy.

Maksymalne długości zmian nowotworowych z każdej serii obrazów mogą być podstawą do postawienia wstępnej diagnozy bazując na parametrze T z klasyfikacji TNM.

Po otrzymanych wynikach oraz zestawieniu ich z tabelą TNM (Tabela 1) cecha T wynosi T4 , ze względu iż długość guza jest większa niż 3 cm oraz guz nacieka na struktury tj. Np. śródpiersie, serce, wielkie naczynia. Występowanie nacieków na inne struktury stwierdzono z braku możliwości rozdzielenia zmian nowotworowych od takich narządów jak serce, podczas procesu binaryzacji. Natomiast cecha N oraz cecha M otrzymują wartości X, tzn. że ocena regionalnych węzłów chłonnych jak i ocena występowania przerzutów odległych do innych narządów są niemożliwe. Obie cechy: N oraz M są niemożliwe do oceny ze względu na przeprowadzanie analizy obrazów jedynie obszaru klatki piersiowej ograniczających się do struktury płuc.

Po uzyskanych wynikach T4, Nx oraz Mx ,kolejnym krokiem do wystawienia wstępnej diagnozy jest określenie stopnia zaawansowania raka (Tabela 2). Po zestawieniu wyników z wartościami tabeli możemy zakwalifikować analizowaną zmianę nowotworową jako stopień IIIB zaawansowania raka.

Ostatnim krokiem jest wybór leczenia pacjenta (Tabela 3). Zgodnie ze stopniem zaawansowania raka IIIB odpowiednim leczeniem dla pacjenta według Tabeli 3 jest radioterapia ewentualnie połączona z chemioterapią.

8. Wnioski

Przeprowadzona binaryzacja na obrazach ma ogromny wpływ na wynik pola powierzchni. Nieodpowiednio dobrany próg binaryzacji może drastycznie zmienić pole powierzchni zmian nowotworowych. Co wiąże się z fałszowaniem wyników w konsekwencji doprowadzając do błędnej diagnozy.

Na objętość końcową zmiany nowotworowej ma wpływ ilość poddanych analizie obrazów, nawet gdy pole poszczególnych zmian nowotworowych w danej serii nie jest zbyt duże. Jednakże nie jest to wystarczające, by podać wstępną diagnozę.

Maksymalna długość zmiany nowotworowej może być podstawą do postawienia wstępnej diagnozy bazując na parametrze T z klasyfikacji TNM.

Stworzony program do analizy i wstępnej diagnostyki nowotoworów może być ułatwieniem dla lekarzy analizujących zdjęcia z rezonansu magnetycznego w celu dobrania odpowiedniego leczenia dla pacjentów cierpiących na nowotwory.

Dzięki funkcji archiwizacji danych obliczeniowych programu można szybko powrócić do wyników pacjentów bez konieczności przeprowadzania ponownej analizy obrazów zmian nowotworowych. Funkcja ta jest dodatkowo użyteczna przy porównywaniu nowych wyników badań po odpowiednim leczeniu z wynikami sprzed rozpoczętego leczenia w celu zweryfikowania, czy podjęte leczenie pacjenta przynosi oczekiwane rezultaty.

Sam program jest jedynie ułatweniem dla lekarzy i techników pracujących w szpitalach, jednakże bez odpowiedniej wiedzy lekarzy program nie jest w stanie sam zadecydować o dalszych losach pacjenta, dlatego też, czynnik ludzki jest nieodzownym elementem diagnostyki pacjentów.

Literatura

1. Pratap R., *Matlab 7 dla naukowców i inżynierów*, Warszawa: Wydawnictwo Naukowe PWN SA, 2013, str. 280.
2. Wróbel Z., Koprowski R., *Praktyka przetwarzania obrazów z zadaniami w programie Matlab*, Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2012, str. 278.
3. Rzyman W., *Rak płuca*, Gdańsk: Katedra i Klinika chirurgii klatki Piersiowej Akademii Medycznej, T. 2, nr 6/2008, str. 407-419.

Wykorzystanie środowiska Matlab w diagnostyce onkologicznej

Streszczenie
Przedstawiona praca ma na celu stworzenie programu ułatwiającego diagnostykę i analizę zmian nowotworowych płuc w środowisku Matlab. Program ma na celu ułatwienie pracy onkologów, co za tym idzie przy pomocy programu lekarz może wskazać interesujący go fragment, czyli zmianę w płucach w celu automatycznego obliczenia pola jej powierzchni oraz dobrania odpowiedniego leczenia pacjenta. Rak płuc to jedna z najbardziej rozwijających się chorób na świecie, dodatkowo zaliczająca się do tych praktycznie nieuleczalnych chorób, ale dzięki postępowi medycyny i inżynierii biomedycznej możliwe jest przedłużenie życia chorych. Głównymi narzędziami do wykrycia raka płuc jest RTG,TK,PET CT, a w szczególnych przypadkach MRI, a ze względu na praktycznie całkowitą bez inwazyjność dla człowieka MRI zasługuje na szczególną uwagę. Program został stworzony w środowisku Matlab, ponieważ jest to pakiet o uniwersalnym środowiskiem umożliwiającym przeprowadzenie złożonych obliczeń naukowo-technicznych, poprzez dostęp do różnorodnych i efektywnych algorytmów obliczeniowych .Pozwala na wizualizację uzyskanych wyników w postaci różnych wykresów i grafiki trójwymiarowej. Matlab to interaktywne środowisko oferujące użytkownikowi wachlarz wbudowanych technicznych funkcji obliczeniowych, graficznych i animacyjnych. Program pozwala na samodzielną rozbudowę zakresu zastosowań poprzez tworzenie własnych skryptów i programów w wybranym języku programowania wysokiego poziomu. Użytkownik ma dostęp do niezawodnych algorytmów matematyki stosowanej oraz do licznych modułów rozszerzeń. Uniwersalność programu pozwala na jego szerokie zastosowanie, które można znaleźć w medycynie, energetyce, fotografii i innych dziedzinach, co wynika z jego wielu zalet i funkcji jakie posiada, a zwłaszcza umożliwia szybkie uzyskanie rezultatów złożonych obliczeń i wizualizację wyników. Matlab umożliwia analizę zdjęć o różnych formatach. Dlatego też cel naszej pracy, którym było przedstawienie analizy zmian nowotworowych w stosunku do objętości płuc było możliwe dzięki funkcjonalności pakietu Matlab.
Słowa kluczowe: Matlab, nowotwory płuc, rezonans magnetyczny.

Using the Matlab environment for oncological diagnosis

Abstract
The presented work aims to create a program facilitating the diagnosis and analysis of lung cancer lesions in the Matlab environment. The program is designed to facilitate oncologists' work, and with the help of a program, a physician can indicate what part of the lung it is, a change in the lung to automatically calculate the area of the lung and to select the appropriate treatment for the patient. Lung cancer is one of the most developing diseases in the world, in addition to these virtually incurable diseases, but with the advancement of medicine and biomedical engineering it is possible to extend the lives of the sick. The main tools for detecting lung cancer are RTG, CT, PET CT, and in special cases MRI, and because of its virtually complete non-invasive nature, MRI deserves special attention. The program was created in Matlab because it is a package with an universal environment enabling Perform complex scientific and technical calculations by accessing a variety of efficient and efficient computing algorithms. It allows to visualize the results obtained in various graphs and 3D graphics. Matlab is an interactive environment that offers a range of built-in technical computing, graphics and animations. Range of applications by creating their own scripts and programs in the selected high-level programming language. User has access to reliable mathematical algorithms used and to numerous expansion modules. Universality of Matlab makes it possible to analyze images of different types, which can be found in medicine, power engineering, photography and other fields, as a result of its many advantages and functions. Therefore, the purpose of our work, which was to present the analysis of tumor changes with respect to lung volume was made possible by the functionality of the Matlab package.
Keywords: Matlab, poumon cancer, MRI

Wiktoria Sapota[1], Sebastian Stach[2], Zygmunt Wróbel[3]

Możliwości programu Matlab w zakresie analizy i wizualizacji powierzchni zespoleń kostnych stosowanych w leczeniu urazów twarzoczaszki

1. Wprowadzenie

Dynamiczny rozwój uprzemysłowienia, urbanizacja i wszechobecna mechanizacja to czynniki znacznie wpływające na rosnącą skalę wypadków, w wyniku których dochodzi do czasowych lub trwałych uszkodzeń ciała [1]. Ze względu na niewielką odporność na działanie sił fizycznych to urazy twarzoczaszki stanowią szczególnie ważny, acz wyjątkowo skomplikowany i interdyscyplinarny problem kliniczny. Na chwilę obecną w chirurgii twarzowo-szczękowej stosuje się dwie grupy metod leczenia urazów: metody chirurgiczne oraz zachowawczo-ortopedyczne, które polegają na zakładaniu wyciągów międzyszczękowych opartych na szynach standardowych oraz fundach elastycznych i gipsowych. Działania zachowawcze często niekorzystnie wpływają na komfort psychiczny pacjenta, dlatego też coraz częściej wdraża się leczenie operacyjne metodą osteosyntezy. W praktyce wymaga ona wykorzystania narzędzi takich jak: druty, pręty, śruby, wkręty i różnego rodzaju płytki zespalające. Umieszczenie odpowiednio zwymiarowanych i ukształtowanych implantów pozwala na stabilizację uszkodzonych fragmentów czaszki, regenerację tkanki kostnej i powrót pacjenta do zdrowia. Metoda ta, mimo wielu zalet nie jest pozbawiona niedoskonałości. Mimo usilnych starań doboru jak najbardziej biozgodnych materiałów wciąż pojawiają się wyniki badań wykazujących niepokojące zmiany zachodzące zarówno w organizmie pacjenta jak i na powierzchni wierzchniej implantów [1]. Świadomość występowania tego problemu przyczyniła się do zdefiniowania problemu badawczego i próby jego rozwiązania przy pomocy środowiska programistycznego Maltab.

2. Materiał i narzędzia badawcze

2.1. Przedmiot analiz

Do badań wykorzystano tytanową płytkę zespalającą typu MODUS Trauma 2.5 wyprodukowaną przez firmę Medatris. Płytka była umieszczone w ciele pacjenta przez okres 18 miesięcy. Implantacja tego typu implantem (Rysunek 5) wskazana jest w przypadkach:

[1] wiktoria.sapota@us.edu.pl, Instytut Informatyki, Wydział Informatyki i Nauki o Materiałach, Uniwersytet Śląski w Katowicach, www.us.edu.pl
[2] sebastian.stach@us.edu.pl, Instytut Informatyki, Wydział Informatyki i Nauki o Materiałach, Uniwersytet Śląski w Katowicach, www.us.edu.pl
[3] zygmunt.wrobel@us.edu.pl, Instytut Informatyki, Wydział Informatyki i Nauki o Materiałach, Uniwersytet Śląski w Katowicach, www.us.edu.pl

- nieregularnego złamania żuchwy,
- złamania wieloodłamkowego,
- złamania kąta żuchwy,
- rekonstrukcji żuchwy [3].

Rysunek 5. Płytka tytanowa MODUS Trauma 2.5, usunięta z ciała pacjenta po 18 miesiącach [źródło własne]

Wszystkie implanty MODUS wykonane są z tytanu niestopowego (ASTM F67, ISO 5832-2) lub stopu tytanu (ASTM F136, ISO 5832-3). Z założenia wszystkie użyte materiały tytanowe są biokompatybilne, odporne na korozję i nietoksyczne w środowisku biologicznym [1].

2.2. Aparatura pomiarowa i wykorzystane oprogramowanie

2.2.1. Mikroskop Olympus LEXT OLS4000

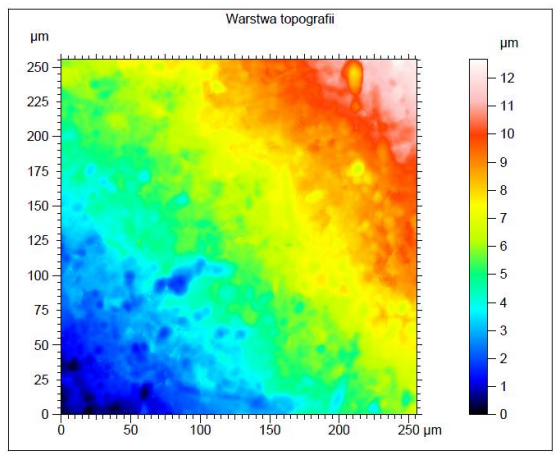
Narzędziem pomiarowym będącym fundamentalnym źródłem danych wykorzystywanych do dalszych badań był laserowy mikroskop konfokalny Olympus LEXT OLS4000. Mikroskop ten posiada możliwość skanowania powierzchni materiałów pełniąc dwie funkcje: mikroskopu świetlnego oraz mikroskopu konfokalnego. Ma on możliwość pracy z dwoma rodzajami źródła światła, dzięki czemu może być przeznaczony do wykonywania precyzyjnych pomiarów oraz trójwymiarowej analizy powierzchni. Mikroskop pozwala na uzyskanie powiększeń powierzchni w zakresie 108x-17280x oraz pomiarów z dokładnością od 1 nm w osi Z i 120 nm w płaszczyźnie XY. Źródłem światła dla pomiarów przestrzennych jest światło lasera, którego długość wiązki wynosi 405 nm [4]. Do badań w niniejszej pracy wykorzystano obiektywy MPLAPONLEXT o powiększeniu 50x.

2.2.2. Narzędzia informatyczne

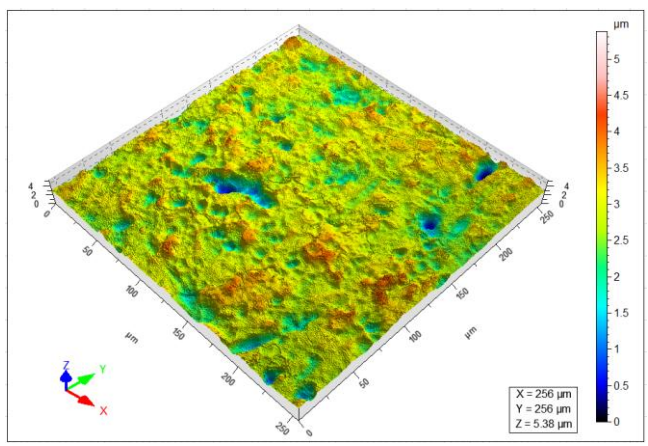
Obrazowanie i analiza struktur powierzchni niejednokrotnie wymaga użycia kilku specjalistycznych programów wzajemnie uzupełniających swoje funkcjonalności. W niniejszych badaniach oprogramowanie MountainsMap® Premium firmy Digital Surf w wersji 6.2.2700 umożliwiło wzbogacenie funkcji mikroskopu konfokalnego: zwiększenie pola widzenia; kompleksową analizę struktury powierzchni i jej geometrii (pomiary geometryczne, obliczanie odległości, kątów, objętości, analiza ziaren lub cząstek); analiza poziomu falistości powierzchni; analiza składowych powierzchni (segmentacja, progowanie) [2]. Do szczegółowych badań powierzchni płytki kostnej zostało wykorzystane oprogramowanie Matlab R2012b. Niniejsze środowisko programistyczne wyposażone w całą gamę funkcji przeznaczonych do rozwijania algorytmów, wizualizacji i analizy danych oraz prowadzenia obliczeń numerycznych idealnie wpisuje się w założoną na początku badań filozofię analizy powierzchni implantów.

3. Metoda badawcza

Prace badawcze rozpoczęto od analizy powierzchni implantu mikroskopem konfokalnym. Po wczytaniu do programu MountainsMap pliku zawierającego dane powierzchni uzyskano widok warstwy topografii (Rysunek 6) oraz obraz powierzchni 3D (Rysunek 7) próbki. Niniejsze obrazy mimo, iż są czytelne nie wnoszą istotnych informacji o nieprawidłowościach występujących w materiale. W celu przeprowadzenia analiz i wyodrębnienia potencjalnych ubytków na powierzchni w programie Matlab, konieczne było właściwe przygotowanie danych stereometrycznych badanych obrazów i ich zapis w postaci pliku .mat, czego dokonano za pomocą programu MountainsMap ® Premium.



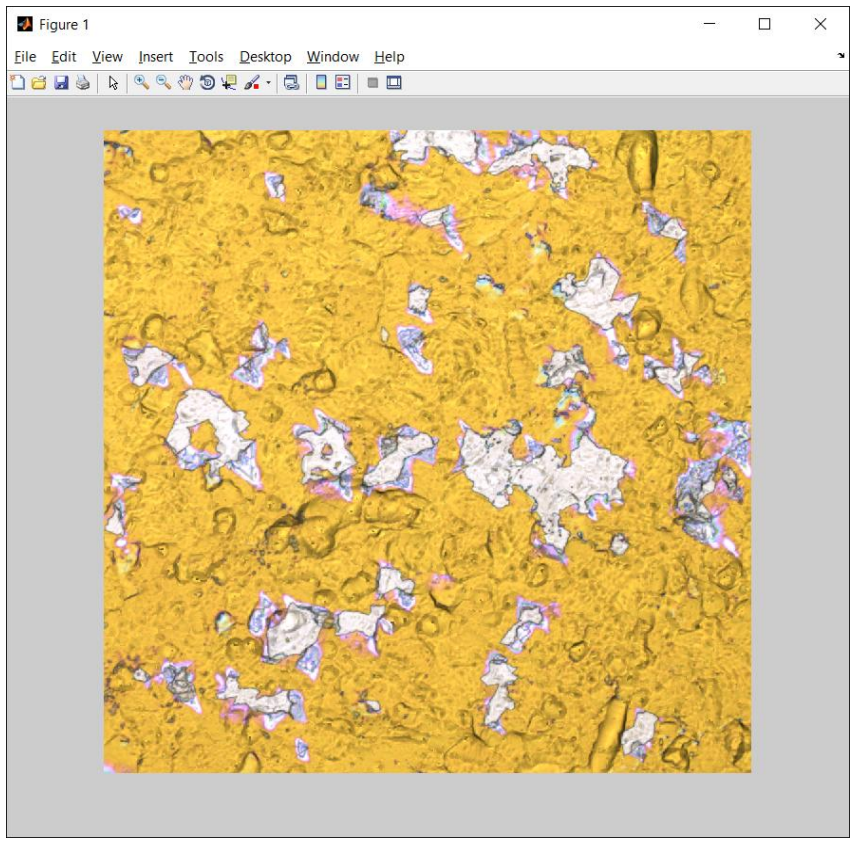
Rysunek 6. Warstwa topografii powierzchni badanej płytki wyświetlona w programie MountainsMap® Premium [opracowanie własne]



Rysunek 7. Widok 3D powierzchni badanej płytki, wyświetlony w programie MountainsMap® Premium [opracowanie własne]

3.1. Analiza powierzchni obrazu 2D

Niewielka ilość informacji o ubytkach wymusza przeprowadzenie działań na obrazie prezentującym topografię próbki. Prezentuje ona obecność ubytków, jednak ich wielkość i granice nie są jasno określone.



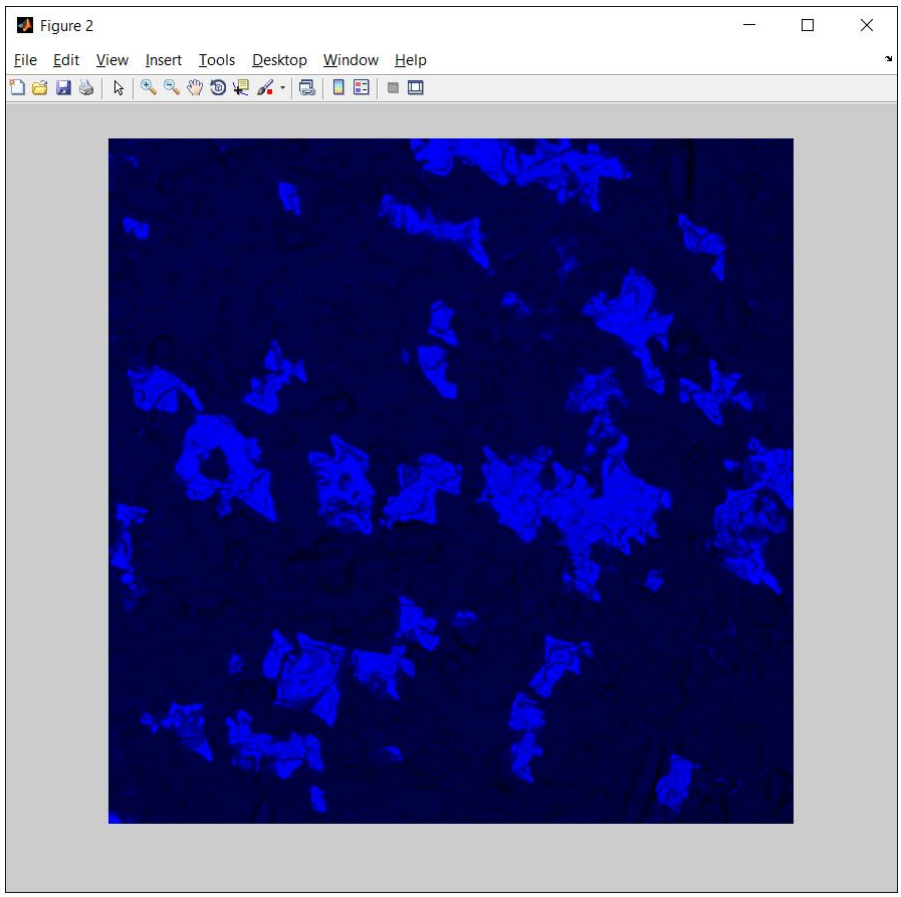
Rysunek 8. Obraz powierzchni badanej płytki uzyskany w trybie 2D za pomocą mikroskopu konfokalnego, wyświetlony w programie Matlab [opracowanie własne]

Aby uzyskać dokładniejsze obszary ubytków materiału konieczne było wyodrębnienie kanału niebieskiego (

Rysunek 9) ze składowej RGB. Operacja ta nazywana jest ekstrakcją kanału koloru i pozwala na wyświetlenie uzyskanego kanału jako obrazu monochromatycznego (w skali szarości) lub w kolorze wybranego kanału.

Następnie utworzono macierz zerową o wymiarach identycznych jak macierz blueChannel i połączono je, celem wyświetlenia obrazu w kolorze.

```
blueChannel = image(:, :, 3);
z = zeros(size(blueChannel));
```



Rysunek 9. Wyodrębnienie kanału niebieskiego z oryginalnego obrazu powierzchni badanej płytki [opracowanie własne]

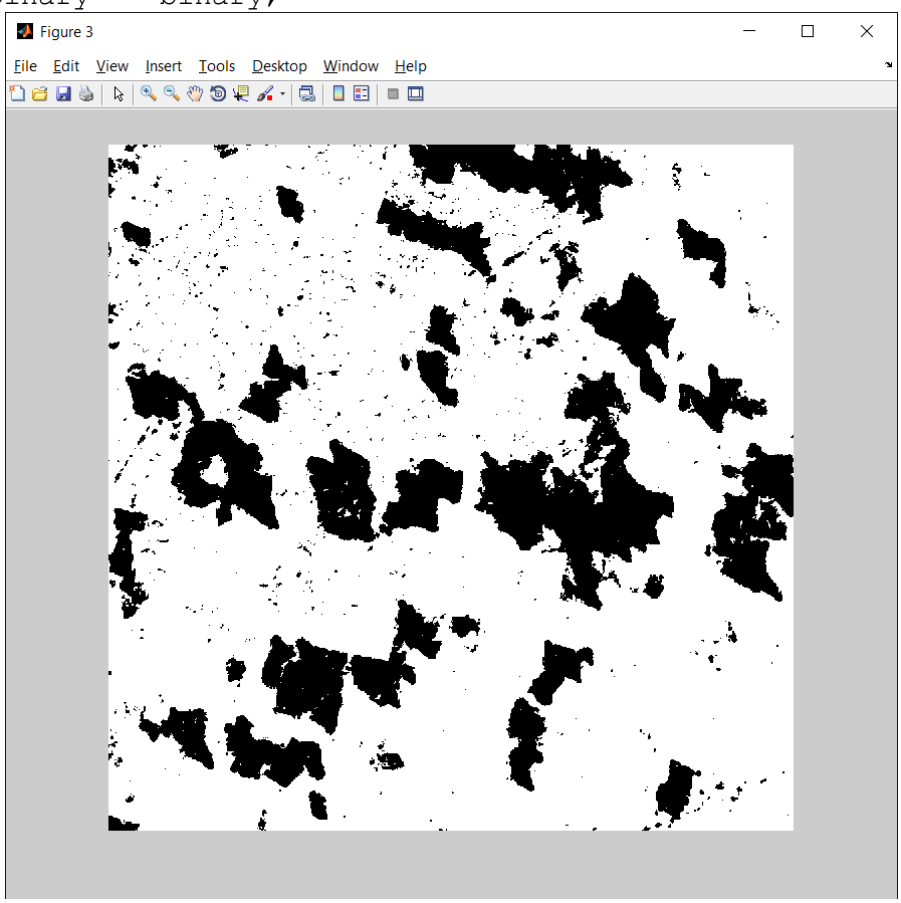
Dzięki binaryzacji z górnym progiem udało się uzyskać obraz bez szumów oraz wykluczyć zbędne dane z badanego obszaru.

Binaryzacja jest jedną z ważniejszych czynności punktowego przetwarzania obrazów. Przeprowadzenie procesu binaryzacji polega na zamianie obrazu zawierającego wiele poziomów szarości na obraz, którego piksele mają wyłącznie wartości 0 lub 1. Oznacza to, iż zachodzi proces konwersji obrazów kolorowych lub monochromatycznych do obrazu binarnego. Dopiero dla obrazów binarnych można przeprowadzić większość pomiarów oraz niektóre złożone przekształcenia. Binaryzacja prawie zawsze poprzedza szczegółową analizę, jest także bardzo przydatna w procesie rozpoznawania. Przeprowadzenie tej operacji znacząco redukuje ilość zawartej w nim informacji. Binaryzacja najczęściej realizowana jest przez progowanie (thresholding), polegające na ustaleniu wartości progowej (threshold), dla której piksele obrazu klasyfikowane są jako piksele obiektu lub jako piksele tła. Podstawowym problemem przy binaryzacji jest znalezienie odpowiedniego progu binaryzacji. Najczęściej dla

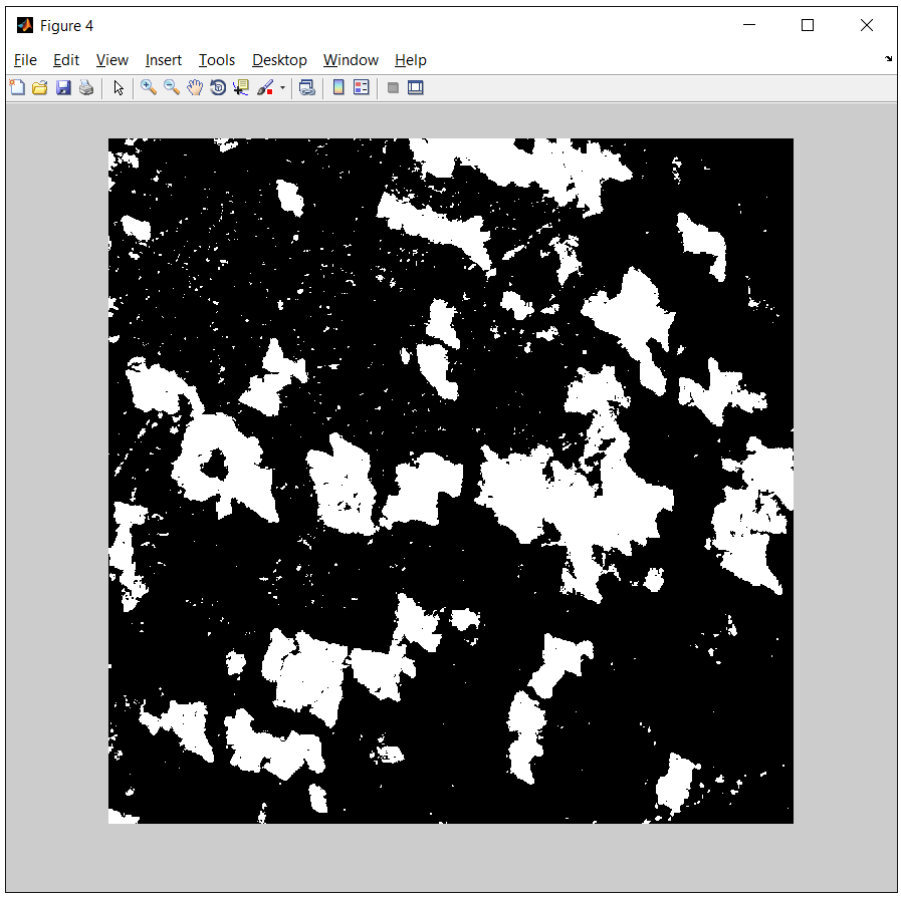
znalezienia właściwej wartości progu tworzy się histogram obrazu, a następnie próg binaryzacji ustala się w ewentualnej „dolinie histogramu" [**Błąd! Nie można odnaleźć źródła odwołania.**].

Za pomocą polecenia konwersji, poprzez usunięcie informacji o barwie i nasyceniu przekształcono kolorowy obraz typu RGB na szary. Ustalono poziom progu górnego oraz określono warunek przypisywania wartości binarnej uzyskując zbinaryzowany obraz (Rysunek 10), który następnie poddano operacji inwersji (Rysunek 11).

```
imagetobin = imread('rgb_blue','tif');
imagetobin = rgb2gray(imagetobin);
upper_threshold = 10;
binary = imagetobin < upper_threshold;
binary = im2bw(binary, 0.5);
binary = ~binary;
```



Rysunek 10. Obraz powierzchni badanej płytki uzyskany w wyniku operacji binaryzacji z progiem górnym [opracowanie własne]



Rysunek 11. Obraz powierzchni po inwersji [opracowanie własne]

Choć obszary ubytków stały się już znacznie lepiej widoczne, ich granice wciąż nie były jednoznaczne. Aby zwiększyć identyfikację obszaru ubytków i usunąć zbędne artefakty wykonano operację zamknięcia obrazu.

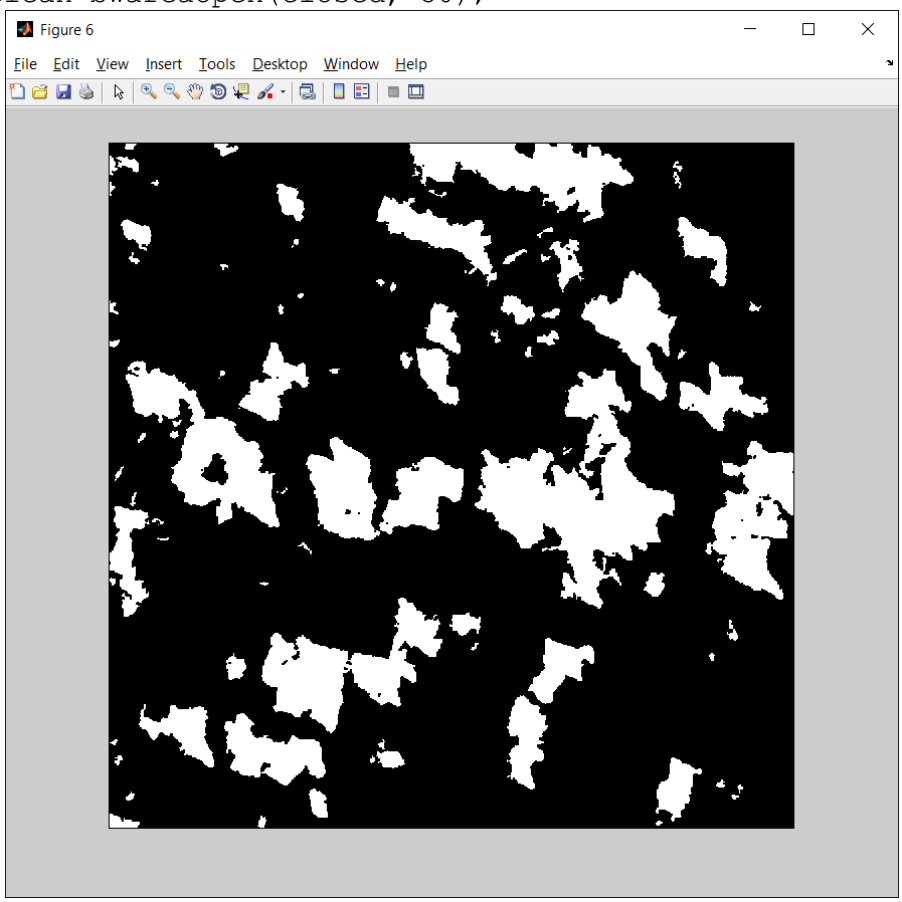
Zamknięcie to jedna z kilku operacji morfologicznych, w wyniku których struktura lub forma obiektu na obrazie zostaje zmieniona. Generalnie, wyróżnia się trzy podstawowe przekształcenia morfologiczne wykorzystywane w obróbce obrazów cyfrowych to: erozja, dylatacja oraz szkieletyzacja. Operacje te mogą być kombinowane w zestawy, w celu stworzenia operacji bardziej złożonych, dających możliwość analizy kształtu elementów oraz określania relacji pomiędzy obiektami zawartymi na obrazie. Definiując tę operacje należy zdefiniować obszar X i element strukturalny wraz z punktem środkowym. Figura zerodowana to zbiór środków wszystkich elementów strukturalnych, które w całości zawarte są we wnętrzu obszaru X. Erozja posiada następujące cechy:

- jest addytywna – erozja wielokrotna daje wynik odpowiadający wielu powtórzeniom erozji z tym samym elementem strukturalnym,

- ma zdolność eliminacji najmniejszych elementów i wygładzania brzegów figury,
- erozja elementem strukturalnym o podłużnym kształcie pozwala uwypuklić cechy obrazu zorientowane w tym samym kierunku, co element,
- dokonuje generalizacji obrazu, czyli pozostawia większe obiekty na obrazie, usuwając szczegóły [6].

Dobór wartości metodą prób i testowania rezultatów przyczynił się do zastosowania wartości progowej 50 oraz eliminacji obiektów niespełniających warunku określonej liczby pikseli:

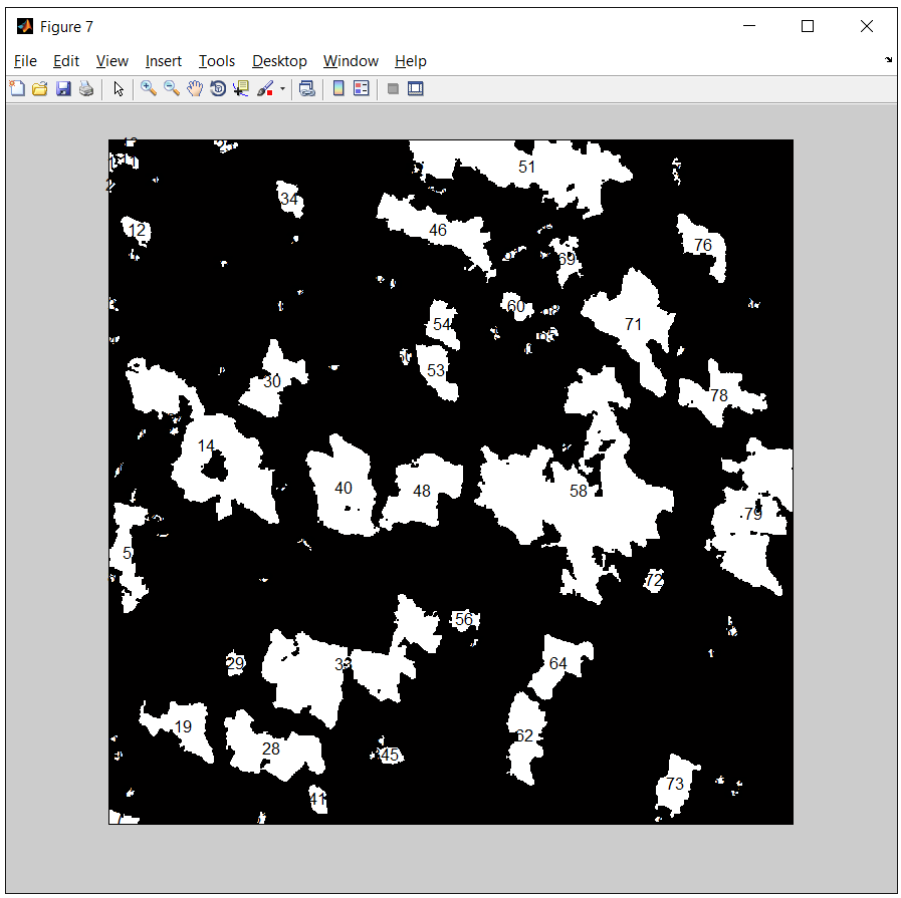
```
closed = bwmorph(binary,'close',50);
clean=bwareaopen(closed, 50);
```



Rysunek 12. Zbinaryzowany obraz powierzchni po operacji zamknięcia z wyeliminowanymi obszarami niespełniającymi warunku [opracowanie własne]

Na tym etapie uzyskany obraz prezentuje wizualnie dużo korzystniejszą jakość, która daje szanse otrzymywania kolejnych wartościowych informacji dotyczących ubytków w powierzchni implantu. Chcąc wyszczególnić pojedyncze obszary przeprowadzono proces ich etykietyzacji oraz numeracji:

```
label=bwlabel(clean);
max(max(label));
s = regionprops(label, 'Centroid');
imshow(label)
hold on
for k = 1:numel(s)
 c = s(k).Centroid;
 text(c(1), c(2), sprintf('%d', k), ...
 'HorizontalAlignment', 'center', ...
 'VerticalAlignment', 'middle');
end
hold off
```



Rysunek 13. Obraz powierzchni próbki z ponumerowanymi i wyróżnionymi obszarami [opracowanie własne]

Polecenie bwlabel (obraz, sąsiedztwo) – służy do etykietowania, tzn. do przypisywania jednakowej wartości pikselom wewnątrz obszarów jednorodnych i rozłącznych z innymi obiektami. Sąsiedztwo przyjmuje wartość 4 lub 8. Każdy jednorodny obiekt ma unikalny numer, będący kolejnymi liczbami całkowitymi. Składnia formuły:

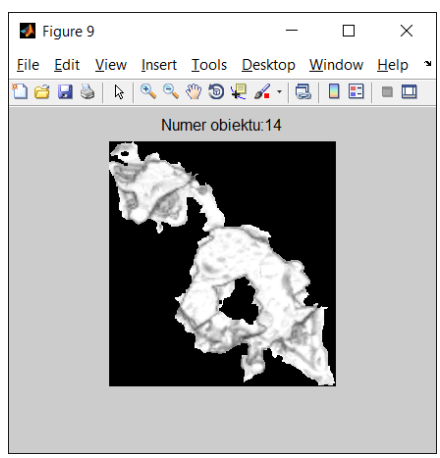
```
[L, n] = bwlabel(BW,s)
```

Wynikiem jest macierz o rozmiarach takich samych jak BW, w której liczby wyrażające jasność pikseli zastąpiono numerami kolejnych obiektów, do których te piksele należą. Tło jest obiektem o numerze zerowym, a pozostałe obiekty numerowane są kolejnymi liczbami naturalnymi od 1 do n [7].

Ponumerowanie obiektów umożliwiło odnoszenie się do dowolnego z nich. Program Matlab, za pośrednictwem odpowiedniego algorytmu pozwolił także na wyświetlenie pojedynczych obszarów – ubytków w materiale (Rysunek 14):

```
for j=14;
 [row,col] = find(label==j);
 len=max(row)-min(row)+2;
 breadth=max(col)-min(col)+2;
 target=uint8(zeros([len breadth] ));
 sy=min(col)-1;
 sx=min(row)-1;
for i=1:size(row,1);
 x=row(i,1)-sx;
 y=col(i,1)-sy;
 target(x,y)=image(row(i,1),col(i,1));
end

mytitle=strcat('Numer obiektu:', num2str(j));
figure,imshow(target);title(mytitle);
end
```

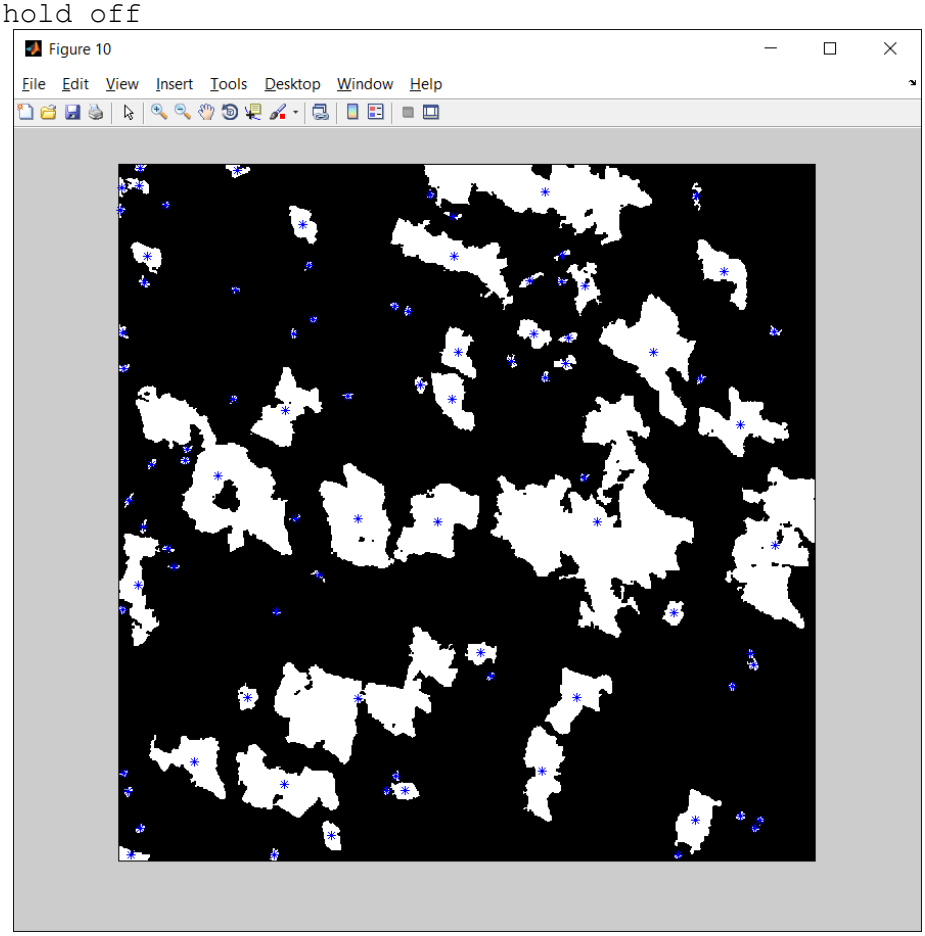


Rysunek 14. Wyodrębniony obiekt nr 14 [opracowanie własne]

Dodatkową możliwością jest także utworzenie wizualizacji szczególnych cech wydzielonych obszarów. W badaniach wyliczono środki ciężkości dla wszystkich ubytków powierzchni (Rysunek 15):

```
centroids = cat(1, s.Centroid);

imshow(cut_image)
hold on
plot(centroids(:,1),centroids(:,2), 'b*')
hold off
```



Rysunek 15. Graficzna prezentacja środków ciężkości wszystkich obiektów badanego obrazu [opracowanie własne]

3.2. Analiza obrazu powierzchni 3D

Wizualizacja danych stereometrycznych, ze względu na dużą ilość punktów pomiarowych i wysoką rozdzielczość prezentowała jedynie siatkę punktów, bez obszarów reprezentujących wysokość powierzchni próbki. Konieczne było utworzenie

zestawu każdorazowo powielanych ustawień zwiększających jakość wyświetlanych analiz.

```
  xlabel ('px')
  ylabel ('px')
  zlabel ('px')
  camlight left;
  lighting phong;
  axis([0, 1024, 0, 1024, -0.007, 0.007]); %określenie limitu wartości na osiach
  view(55,72);      %punkt      odniesienia      ustawienie kamery:(azymut, wysokość)
```

Aby zniwelować pochyłości płaszczyzny wykonano proces poziomowania, składający się z podrzędnych operacji:

Zamiana macierzy na tablicę – przeliczenie jednego wiersza macierzy na współrzędne jednego punktu w przestrzeni:

```
  licznik=0;
  [x,y]=size(SOURCE);
  XYZ = zeros(x*y,3,'double');
  for i=1:x
   for j=1:y
   licznik=licznik+1;
   wsp_pkt=[j i SOURCE(i,j)];
   XYZ(licznik,:)=wsp_pkt;
   end
  end
```

Utworzenie płaszczyzny dopasowanej do utworzonej chmury punktów:

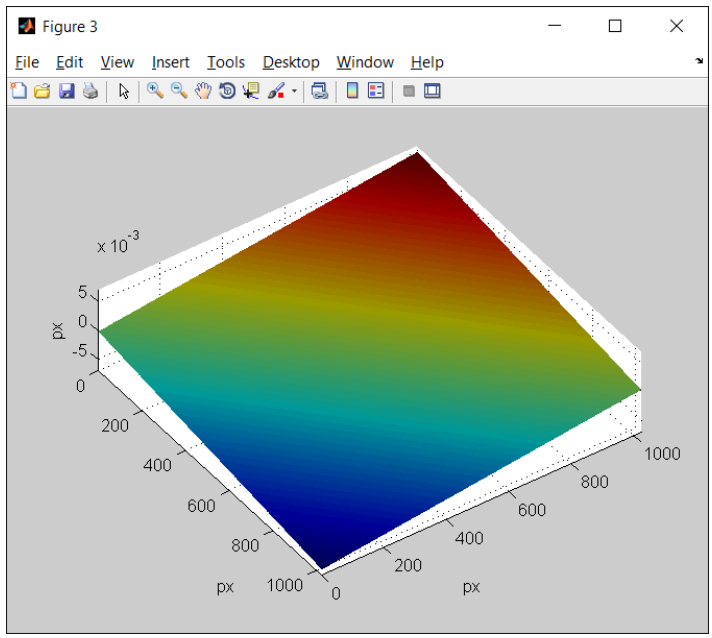
```
C = planefit(XYZ(:,1),XYZ(:,2),XYZ(:,3));
zfit = C(1)*XYZ(:,1)+C(2)*XYZ(:,2) + C(3);
```

Rysowanie płaszczyzny dopasowanej do chmury punktów (Rysunek 16)

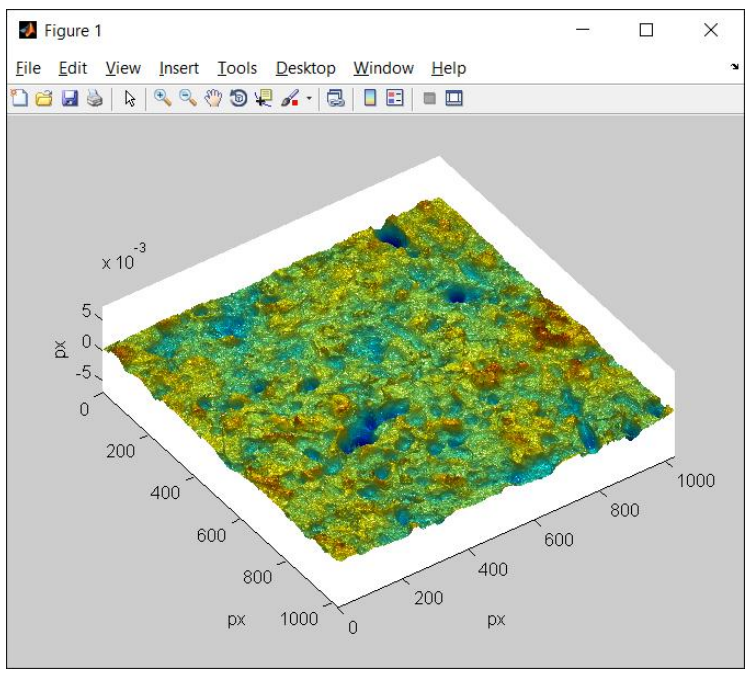
```
plane=accumarray(XYZ(:,[2 1]), zfit);
```

Odejmowanie płaszczyzny od chmury punktów – ostateczny wynik poziomowania (Rysunek 17)

```
NewSource=SOURCE-plane;
```



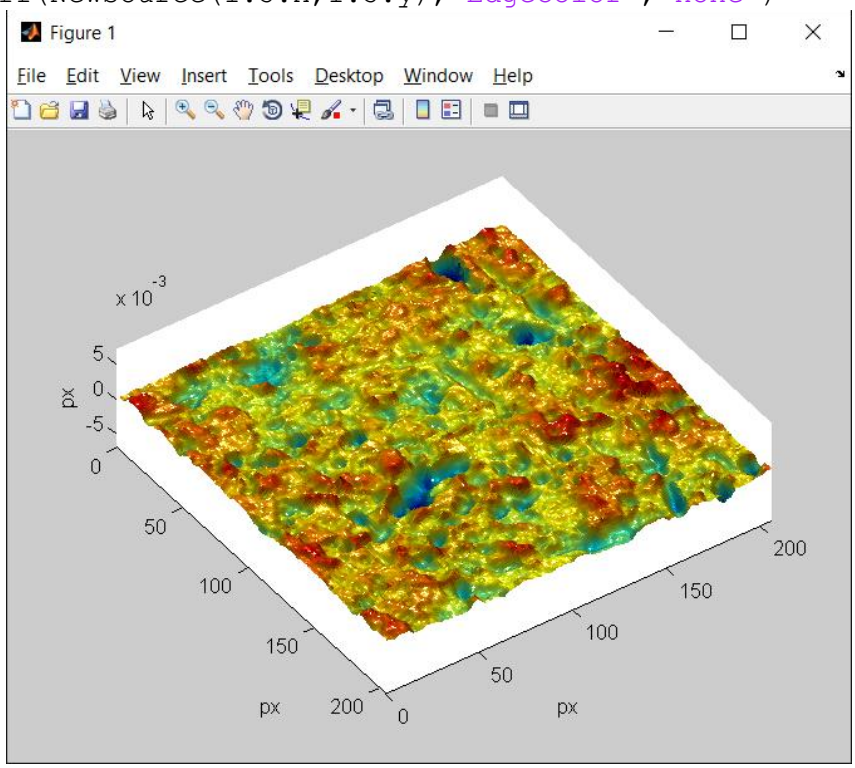
Rysunek 16. Płaszczyzna dopasowana do chmury punktów [opracowanie własne]



Rysunek 17. Wypoziomowana powierzchnia próbki [opracowanie własne]

Zwiększenie czytelności obrazu powierzchni (Rysunek 18) osiągnięto poprzez wyświetlenie co 5 punktu macierzy:

```
[x,y]=size(NewSource);
surf(NewSource(1:5:x,1:5:y),'EdgeColor','none')
```

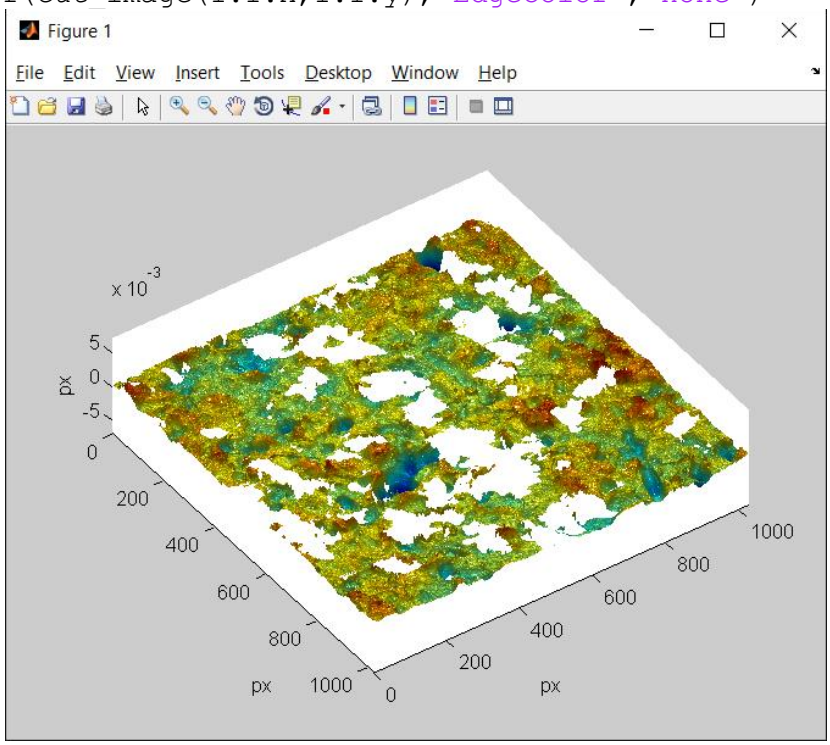


Rysunek 18. Powierzchnia próbki z ograniczeniem liczby wyświetlanych punktów, bez siatki punktów [opracowanie własne]

Ostatnim etapem analizy było połączenie uzyskanych powierzchni 2D i 3D celem wyświetlenia wizualizacji próbki z usunięciem fragmentów zawierających defekty (Rysunek 19):

```
cut_image=NewSource;
[x,y]=size(cut_image);
for i=1:y
 for j=1:x
 if clean(i,j)==1
 cut_image(i,j)=NaN;
 end
 end
end
hold on;
xlabel ('px') %opis jednostek na osiach
ylabel ('px')
zlabel ('px')
```

```
camlight left;
lighting phong;
axis([0, 205, 0, 205, -0.007, 0.007]);
view(55,72);
surf(cut_image(1:1:x,1:1:y),'EdgeColor','none')
```



Rysunek 19. Izometryczny widok powierzchni analizowanej płytki ze zwizualizowanymi ubytkami materiału warstwy wierzchniej [opracowanie własne]

3.3. Szacowanie pola powierzchni ubytków

Uzupełnieniem otrzymanych analiz jest zestaw obliczeń całkowitego pola powierzchni w μm^2, a także ubytków materiału na powierzchni próbki.

Suma pola powierzchni wszystkich wyodrębnionych obiektów

```
AreaOfCavity=sum([stats.Area]);
```

Pole powierzchni analizowanej powierzchni w pikselach^2

```
AreaInPx=SOURCEXSIZE*SOURCEXSIZE;
```

Pole powierzchni analizowanej powierzchni 1 w μm^2

```
AreaInUm=SOURCEXSIZE*SOURCEXSPACING/1000*SOURCEXSIZE*SOURCEYSPACING/1000;
```

Szacowanie udziału procentowego ubytków dla całej analizowanej powierzchni

```
PercentOfLoss=AreaOfCavity/AreaInPx
```

Zastosowane rozwiązania umożliwiły uzyskanie parametrów określających wielkość wybranych obszarów na powierzchni badanej próbki.

Tabela 1. Zestawienie wielkości powierzchni ubytków dla badanej próbki [opracowanie własne]

Parametr	Jednostka	Wartość
pole powierzchni analizowane pod mikroskopem	[px2]	1048576
pole powierzchni analizowane pod mikroskopem	[µm2]	65536
sumaryczne pole powierzchni wszystkich wydetekowanych obiektów (ubytków)	[px2]	230028
sumaryczne pole powierzchni wszystkich wydetekowanych obiektów (ubytków)	[µm2]	14377
udział procentowy powierzchni ubytków w stosunku do pola powierzchni analizowanego pod mikroskopem	[%]	22

4. Analiza wyników

Przeprowadzone analizy potwierdziły przypuszczenia o powstawaniu ubytków na warstwie wierzchniej badanego implantu. Na podstawie wyliczonych pól obiektów (Tabela 1) zauważyć można, że nie są to symboliczne wartości, a znaczny ubytek materiału. Można przyjąć, że są to jedynie złuszczenia lub odpryski, jednak geneza występowania tego zjawiska wymaga podjęcia dalszych badań i przeprowadzenia szerszej analizy. Z praktycznego punktu widzenia niepokojący jest fakt występowania zmian w pozornie biozgodnym materiale, w czasie obecności w ciele pacjenta.

5. Podsumowanie

Środowisko programistyczne Matlab umożliwiło bliższe poznanie zasad realizacji elementarnych metod służących do analizy i przetwarzania obrazu. Mnogość funkcji, elastyczność w dostosowywaniu współczynników i parametrów niejednokrotnie pozwalały na zastosowanie niestandardowych operacji. Praca z oprogramowaniem niejednokrotnie wykazała, że zastosowanie z pozoru prostych operacji umożliwia kreowanie wielu różnorodnych rozwiązań poprzez odpowiednie łączenie funkcji i procedur. Oprogramowanie jest bardzo przydatnym narzędziem do przeprowadzenia bardzo szczegółowej analizy powierzchni biomateriału i uzyskania dokładnych danych, których pozyskanie było głównym celem badawczym.

Wydetekowane miejsca ubytków dają oczywisty dowód na występowanie uszkodzeń w implancie. Niejasną kwestią pozostaje jednak ich pochodzenie, którym może być zarówno wpływ działających sił fizycznych czy też czynników takich jak np.: kontakt z płynami fizjologicznymi. Istnieją badania, które wykazują, że na korozję warstwy wierzchniej implantu ma wpływ nie tylko środowisko jamy ustnej, ale także fluor zawarty w zdecydowanej większości past do zębów [8]. Nieregularność występowania uszkodzeń na powierzchni implantu może sugerować, że płytka zawiera domieszki innych materiałów lub posiada wykonaną z nich powłokę. Istnieje również możliwość, że całe zespolenie kostne nie jest wykonane z tytanu, lecz jest nim pokryta jedynie warstwa wierzchnia. Nie ma wątpliwości, że została ona uszkodzona i z pewnością w jakiejś części trafiła do organizmu pacjenta. Niniejsze badania są dobrym wstępem do analizy, czy takie ilości tytanu są szkodliwe i czy mogą wywoływać dotychczas nieznane skutki niepożądane.

Literatura

1. Zielińska-Bliźniewska H., Niewiadomski P., Foczpański J., Pietkiewicz P., Olszewski J. *Analiza obrażeń twarzoczaszki w materiale własnym*, Kwartalnik ortopedyczny, (1) 2012, s. 135.

2. Sapota W.*Zastosowanie programu Matlab w analizie powierzchni zespoleń kostnych wykorzystanych w chirurgii twarzowo-szczękowej*, Uniwersytet Śląski w Katowicach, 2016
3. URL: http://www.e-mikroskopy.pl/programy-do-analizy-obrazu/digital-surf-mountainsmap/mountainsmap-premium-1/mountainsmap-premium.html [dostęp: 11 sierpnia 2013]
4. URL: http://www.medartis.com/uploads/MODUS-00000011_v0.pdf [dostęp: 15.04.2016]
5. Sapota W. *Obiektywy długodystansowe zastosowane w mikroskopii konfokalnej*, Uniwersytet Śląski w Katowicach, 2014 .
6. Wróbel Z., Koprowski R.*Praktyka przetwarzania obrazów z zadaniami w programie Matlab*, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2012.
7. Kruk M., Automatyczny system rozpoznawania komórek na podstawie obrazu mikroskopowego wybranej tkanki ludzkiej dla potrzeby diagnostyki medycznej, Warszawa, 2008 [rozprawa doktorska].
8. Dwornik M., *Wprowadzenie do analizy obrazu,* Analiza i Przetwarzania Obrazów Cyrowych [materiał dydaktyczny dla studentów IIIr Informatyki Stosowanej 2015/2016].
9. Egusa H., Ko N., Shimazu T., Yatani H. *Suspected association of an allergic reaction with titanium dental implants: a clinical report,* Journal of Prosthetic Dentistry, (5/100) 2008, 344–347.

Możliwości programu Matlab w zakresie analizy i wizualizacji powierzchni zespoleń kostnych stosowanych w leczeniu urazów twarzoczaszki

Streszczenie:
Wizualna ocena tytanowego implantów stosowanego w chirurgii twarzowo-szczękowej przebywającego w ciele pacjenta przez okres 18 miesięcy, wykazała zmiany w strukturze powierzchniowej materiału. Obserwacja ta przyczyniła się do wysunięcia tezy zakładającej niszczenie materiału pod wpływem czasu oraz innych czynników na które narażony jest implant. Głównym źródłem danych pozwalających na rozwiązanie problemu badawczego są warstwy wierzchnie uzyskane za pomocą mikroskopu konfokalnego Olympus LEXT OLS4000. Za pomocą oprogramowania MountainsMap® Premium uzyskano stereometryczne dane pomiarowe, które zostały poddane kolejnym operacjom przetwarzania i analizy badanych materiałów. Do szczegółowej analizy powierzchni implantów wykorzystano specjalistyczne oprogramowanie Matlab. Implementacja autorskiego programu, obejmującego realizację szerokiego zakresu funkcji i działań matematycznych, umożliwiła dokładną wizualizację obrazów prezentujących strukturę powierzchni biomateriału w trybie 2D i 3D. Metodyka oparta o operacje na elementarnych obrazach: binaryzację, różnego rodzaju konwersje, eliminowanie szumów i określanie granic obszarów pozwoliły na otrzymanie dokładnych danych dotyczących ubytków w materiale. Uzyskane rezultaty potwierdziły założoną na początku procesu badawczego tezę o uszkodzeniu warstwy powierzchniowej materiału.
Słowa kluczowe: Matlab, analiza obrazu, wizualizacja, osteosynteza, implant.

The capabilities of MATLAB for analysis and visualization of the surface of bone anastomoses used in the treatment of craniofacial injuries

Abstract:
Visual assessment of titanium implants used in maxillofacial surgery residing in the patient's body over a period of 18 months showed changes in the surface structure of the material. This observation led to advancing the thesis that the structure is damaged under the influence of time and other factors to which the implant is exposed. The main sources of data to solve the research problem were the surface layers obtained using the Olympus LEXT OLS4000. MountainMap® Premium provided stereometric measurement data that were subjected to further processing and analysis of the test materials. Matlab was used for a detailed analysis of the implant surfaces. Implementation of the authors' program, which includes a wide range of functions and mathematical operations, allowed for accurate visualization of images showing the structure of the biomaterial surface in 2D and 3D. The methodology based on operations on elementary images, such as binarization, different kinds of conversion, noise elimination and determination of the boundaries, provided accurate data on defects in the material. The results confirmed the thesis set forth at the beginning of the research about the damage to the material surface layer.
Keywords: Matlab, image analysis, visualisation, osteosynthesis, implant.

Katarzyna Ignatiuk[1]

Reguły akcji jako narzędzie wspomagające klasyfikację chorób

1. Wstęp

W przeciągu ostatnich kilkudziesięciu lat obserwuje się dynamiczny wzrost informacji przetwarzanych przez systemy komputerowe. Akumulacja informacji w bazach danych, potrzeba ich szybkiego przetwarzania i wydobywania z nich znaczących informacji miała wpływ na rozwój technik eksploracji wiedzy (ang. *data mining*).

Początki technik *data mining* sięgają roku 1989, kiedy to Gregory Piatetsky-Shapiro zorganizował pierwsze warsztaty technik odkrywania wiedzy z baz danych. Termin *data mining* został zaakceptowany przez środowisko sztucznej inteligencji i uczenia maszynowego, a techniki zostały sprawnie wykorzystane w biznesie przy opracowywaniu strategii marketingowych. Wydzielone nowe zagadnienie miało bardzo solidne podwaliny programistyczne. Wśród pierwszych prac tego okresu można wymienić prace [1-3] oraz szeroką działalność I.H. Wittena, który rozpoczął opracowywanie oprogramowania WEKA wykorzystującego techniki *data mining* [4, 5]. Szerszy rozwój tej dziedziny nastąpił kilka lat później, po utworzeniu w 1996 roku czasopisma *Data Mining and Knowledge Discovery*. W 2002 r. techniki *data mining* zostały wykorzystane do zdobycia informacji na temat terrorystów biorących udział w zamachu z 11 września 2001 [6]. W 2003 roku wykorzystano je również do polepszenia efektów leczenia guza mózgu u dzieci. Eric Bremer, dyrektor badań nad rakiem w szpitalu Children's Memorial Hospital w Chicago, podjął próbę budowy bazy danych kodów genetycznych dla przypadków zachorowań na raka mózgu wśród dzieci [7].

Prac poświęconych zagadnieniom *data mining* jest wiele. Obejmują one takie techniki eksploracji wiedzy jak metody statystyczne, sieci neuronowe, metody uczenia maszynowego, metody ewolucyjne, zbiory przybliżone czy reguły akcji. Szczególna uwaga w tej pracy została zwrócona na ostatni rodzaj technik odkrywania wiedzy z baz danych – reguły akcji.

Reguły akcji mogą być wydobywane z systemu informacyjnego, który budują trzy, niepuste zbiory: zbiór obiektów, zbiór atrybutów oraz zbiór wartości tych atrybutów. Atrybuty można podzielić na atrybuty stałe oraz elastyczne. Zmianom w systemie informacyjnym poddaje się jedynie atrybuty elastyczne i dzięki tym zmianom można przekwalifikować obiekt z jednego stanu do innego, najczęściej bardziej pożądanego.

W związku z powyższym, reguły akcji są przydatnym narzędziem do odkrywania „ścieżek chorób", bazując na wiedzy zgromadzonej w medycznych bazach danych. Ze względu na charakter informacji zgromadzonych w bazie danych, opracowywane są rożne algorytmy pozwalające wyselekcjonować reguły akcji. Tak więc, do wydobycia reguł z systemu HEPAR w pracy [8] został wykorzystany algorytm reguł asocja-

[1] ignatiukkatarzyna@gmail.com, Zakład Biocybernetyki i Inżynierii Biomedycznej, Wydział Mechaniczny, Politechnika Białostocka

cyjnych. System analizował bazę chorób wątroby i pozwalał na podstawie badań pacjenta ocenić, jakie leczenie w danym przypadku ma być podjęte. Reguły akcji zostały również wykorzystane w pracy [9] przy projektowaniu systemu MIRAI, który oceniał wpływ wybranego gatunku muzyki na emocje człowieka. W pracy [10] algorytm ADReD wspomógł diagnozę pacjentów z rakiem piersi, algorytm ARAS w pracy [11] zaproponował działania dla poprawy stanu zdrowia osób z problemami laryngologicznymi, zaś algorytm ERID w publikacjach [12] i [13] pomógł zrozumieć związek miedzy leczeniem płaskostopia u dzieci a pomiarami tej dysfunkcji. Wymienione przykłady prac wykorzystujących algorytmy *data mining* świadczą o różnorodności tematów medycznych, w obrębie których można wykorzystać reguły akcji

W poniższej pracy zostały zaprezentowane wybrane algorytmy wydobywające reguły z systemu informacyjnego. W pierwszej kolejności przestawiono pojęcie systemu informacyjnego oraz reguł akcji. Stanowiły one bazę dla prezentowanych w pracy algorytmów klasyfikujących LEM1, LEM2 oraz ERID. Każdy z tych algorytmów został zaprezentowany na podstawie uproszczonego, medycznego systemu informacyjnego chorób tarczycy.

2. Reguły akcji w systemie informacyjnym

2.1. System informacyjny

System informacyjnym S można zdefiniować jako trójelementowy zbiór $S = (X, A, V)$, w którym [14, 15, 16]:

X jest skończonym, niepustym zbiorem obiektów,

A jest skończonym, niepustym zbiorem atrybutów,

V jest zbiorem wartości atrybutów A.

Zależność $X \rightarrow V_a$ nazywana jest funkcją dla atrybutu $a \in A$, która zwraca wartość v atrybutu a wybranego obiektu x [14].

Tab.1. przedstawia przykładowy kompletny system informacyjny S utworzony na podstawie bazy chorób tarczycy. Zawarte w nim zostały wartości hormonów tarczycy pacjentów x, które mają największy wpływ na zdiagnozowanie niedoczynności tarczycy, nadczynności czy eutyreozy (prawidłowej funkcji hormonalnej tarczycy). W systemie S, konkretne wartości atrybutów (wiek, TSH, T3, rodzaj choroby tarczycy) i obiektów (pacjentów) zostały zastąpione zmiennymi dla obrazowego przedstawienia algorytmów reguł klasyfikujących.

Tabela 1. Fragment kompletnego systemu informacyjnego S

Pacjent (obiekt)	Wiek (atrybut a)	TSH (atrybut b)	T3 (atrybut c)	Rodzaj choroby tarczycy (atrybut d)
x_1	a_3	b_1	c_2	d_2
x_2	a_2	b_1	c_3	d_2
x_3	a_1	b_2	c_1	d_1
x_4	a_3	b_2	c_2	d_2
x_5	a_1	b_1	c_1	d_2

Źródło: Opracowanie własne

Atrybuty wykorzystywane do opisu obiektów systemu informacyjnego można podzielić na stałe, elastyczne oraz decyzyjne [14]. Atrybutem stałym jest wiek, atrybutem elastycznym TSH oraz T3, zaś decyzyjnym rodzaj choroby tarczycy. Jeżeli wszystkie atrybuty w systemie S są funkcjami, czyli jeden atrybut jest opisywany przez jedną wartość atrybutu, to wówczas system S jest pełny. Niepełnym systemem jest system, w którym można odnaleźć nie w pełni zdefiniowane, poprzez wartości atrybutów, obiekty i jest to system z „lukami" informacyjnymi [14, 15]. Użytkownik systemu może manipulować wartościami atrybutów elastycznych, proponując zmiany ich wartości, które wpłyną na zmianę wartości atrybutów decyzyjnych.

2.2. Reguły akcji

Regułą akcji jest reguła wydobyta z systemu informacyjnego, która opisuje możliwą transformację stanu atrybutu decyzyjnego, poprzez zmianę wartości atrybutów elastycznych. Regułę akcji definiuje wyrażenie [14]:

$$[(\omega)*(\alpha \rightarrow \beta)] \Rightarrow (\varphi \rightarrow \psi) \qquad (1)$$

Atrybut ω jest połączeniem uzależnionych od siebie cech obu grup, (α→β) jest proponowaną zmianą elastycznych cech obiektu, zaś (φ→ψ) jest rezultatem akcji [14].

Przykład

Dla systemu decyzyjnego S przedstawionego w tab. 1., atrybutami elastycznymi, a tym samym klasyfikującymi, są atrybuty a, b oraz c. Dla obiektu x_3 w systemie S odpowiednią decyzję można przedstawić opisowo:

- Jeżeli wartość atrybutu a jest równa a_1,
- Jeżeli wartość atrybutu b jest równa b_2,
- Jeżeli wartość atrybutu c jest równa c_1,
- Wtedy wartość atrybutu d jest równa d_1.
- Taką regułę opisową można przedstawić w bardziej formalny sposób [14]:

$$[(a(x_3) = a_1)*(b(x_3) = b_2)*(c(x_3) = c_1)] \Rightarrow [d(x_3) = d_1] \qquad (2)$$

W systemie S nie ma żadnego innego obiektu, który opisywany jest przez te same wartości atrybutów a, b i c, dlatego też regułę (2) można uogólnić dla całego system informacyjnego S [14]:

$$[(a(x) = a_1)*(b(x) = b_2)*(c(x) = c_1)] \Rightarrow [d(x) = d_1] \qquad (3)$$

Ponadto, regułę można uprościć. W systemie S można zdefiniować podobną regułę dla obiektu x_5. Obiekty x_3 i x_5 różnią się wartościami atrybutu b, atrybut a jest atrybutem stałym. Można uprościć regułę pomijając atrybut c [14]:

$$[(a(x) = a_1)*(b(x) = b_2)] \Rightarrow [d(x) = d_1] \qquad (4)$$

Dla pozostałych obiektów reguły wyglądają w następujący sposób [14]:

$$[(a(x_1)=a_3)*(b(x_1)=b_1)*(c(x_1)=c_2)]\Rightarrow[d(x_1)=d_2] \quad (5)$$

$$[(a(x_2)=a_2)*(b(x_2)=b_1)*(c(x_2)=c_3)]\Rightarrow[d(x_2)=d_2] \quad (6)$$

$$[(a(x_4)=a_3)*(b(x_4)=b_2)*(c(x_4)=c_2)]\Rightarrow[d(x_4)=d_2] \quad (7)$$

$$[(a(x_5)=a_1)*(b(x_5)=b_1)*(c(x_5)=c_1)]\Rightarrow[d(x_5)=d_2] \quad (8)$$

Reguły (5), (6), (7) można zastąpić jedną regułą, gdyż decyzja ponownie nie zależy od wartości atrybutu c [14]:

$$[(a(x)=a_1)*(b(x)=b_1)\Rightarrow[d(x)=d_2] \quad (9)$$

Ostatecznie, uproszczona reguła dla systemu S ma postać [14]:

$$[a_1*(b_2\rightarrow b_1)\Rightarrow[d_1\rightarrow d_2] \quad (10)$$

Reguły akcji mogą być odkrywane z systemu informacyjnego na różne sposoby, zależnie od rodzaju analizowanego systemu informacyjnego. W kolejnym rozdziale zostaną zaprezentowane algorytmy klasyfikujące obiekty do klas decyzyjnych, które należą do systemu LERS oraz algorytm ERID, typowy algorytm dla niepełnych systemów informacyjnych.

3. Algorytmy reguł klasyfikujących

Wśród technik *data mining* spotyka się wiele algorytmów klasyfikujących obiekty baz danych. Należą do nich algorytmy drzew decyzyjnych, k - najbliższego sąsiada, sztuczne sieci neuronowe oraz również reguły akcji. Ostatnia z wymienionych metod nie tylko pozwala sklasyfikować obiekt względem atrybutów decyzyjnych, ale również wydobyć z systemu informacje na temat brakujących danych, które generują wiele problemów w odkrywaniu wiedzy innymi metodami. Bardzo często brakujące dane są traktowane jako wartość pośrednia miedzy nieznaną wartością a dokładnie określoną.

W rozdziale tym zostaną zaprezentowane trzy algorytmy wydobywające z systemu informacyjnego reguły klasyfikujące. W pierwszej kolejności zostaną opisane algorytmy systemu LERS, który składa się z algorytmu LEM1 dla kompletnego systemu decyzyjnego oraz algorytmu LEM2, dla systemu niekompletnego. LEM2 zostanie porównany z innym algorytmem dla systemów z „lukami informacyjnymi" – z algorytmem ERID.

3.1. Algorytmy systemu LERS

LERS (*Learning from Examples based on Rough Sets*) jest systemem, który wyłania zbiór reguł na podstawie oryginalnych danych i klasyfikuje nowe przykłady używając zbioru reguł wcześniej zdobytych. To podejście wykorzystuje teorię zbiorów przybliżonych [14, 15]. System bazujący na LERS posiada dane wejściowe

reprezentowane przez tabele decyzyjne. Przykłady są opisane przez wartości atrybutów i charakteryzowane przez wartości decyzji. Wszystkie przykłady z taką samą wartością decyzji należą do tej samej grupy. LERS poszukuje regularności w tabeli decyzyjnej i wyłania dwa zestawy reguł: pewne i możliwe [14]. System LERS posiada dwa główne podejścia do tworzenia reguł: LEM1 (praca nad całymi atrybutami) i LEM2 (praca nad parami wartości atrybutów) [14, 15].

3.1.1. Algorytm LEM1

System S jest definiowany przez zbiory obiektów - *X*, atrybutów - *A* i wartości tych atrybutów - *V*. Głównym celem algorytmu przedstawionego poniżej jest znalezienie zbioru *K* ze wszystkich pokryć *C* w *D*, gdzie *C* i *D* zawierają się w *A* [14]. Moc zbioru *X* zdefiniowana jest, jako card(X). Zbiór wszystkich podzbiorów o tej samej mocy *k* zbioru *D* jest oznaczony następująco:

$$B_k = \{\{x_{i_1}, x_{i_2}, ..., x_{i_k}\} : (\forall_j \le k)[x_{i_j} \in D] \qquad (11)$$

gdzie *k* jest dodatnią liczbą całkowitą [14].

Dla takich warunków algorytm LEM1 ma postać [14]:

Start
k:=0
Dla każdego *a* w zbiorze *D* wykonaj:
oblicz podział {a}*
oblicz podział C*
k:=1;
jeżeli k≤card(D) wykonaj:
dla każdego zbioru B w B_k wykonaj
jeżeli (B nie jest nadzbiorem z członkami zbioru K) i ($\prod_{x \in B}\{a\}^* \le C^*$)
wtedy dodaj B do K;
k:=k+1
zakończ
Koniec

Przykład

Tab. 2. reprezentuje system $S = (X, A, V)$. Atrybuty $\{a, b, c\}$ są atrybutami klasyfikującymi, $\{d\}$ jest atrybutem decyzyjnym. $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ jest zbiorem obiektów.

Tabela 2. Pełny system informacyjny

Pacjent (obiekt)	Wiek (atrybut a)	TSH (atrybut b)	T3 (atrybut c)	Rodzaj choroby tarczycy (atrybut d)
x_1	a_1	b_2	c_2	d_2
x_2	a_1	b_1	c_1	d_1
x_3	a_2	b_2	c_2	d_2
x_4	a_2	b_1	c_1	d_1
x_5	a_3	b_1	c_2	d_3
x_6	a_3	b_1	c_2	d_3

Źródło: Opracowanie własne

Algorytm LERS na przykładzie systemu decyzyjnego z tab. 2. można opisać w kilku krokach [14, 15]:

Krok pierwszy: Zbiór X jest dzielony na podzbiory dla pojedynczych atrybutów:

$\{a\}^* = \{\{x_1, x_2\},\{x_3, x_4\},\{x_5, x_6\}\}$,

$\{b\}^* = \{\{x_2, x_4, x_5, x_6\},\{x_1, x_3\}\}$,

$\{c\}^* = \{\{x_2, x_4\},\{x_1, x_3, x_5, x_6\}\}$,

$\{d\}^* = \{\{x_2, x_4\},\{x_1, x_3\},\{x_5, x_6\}\}$.

Krok drugi: Żaden podzbiór nie jest podzbiorem $\{d\}^*$ i podzbiorem nieoznaczonym, więc generowane są dwuelementowe podzbiory:

$\{a,b\}^* = \{\{x_1\},\{x_3\},\{x_2\},\{x_4\}, \{x_5, x_6\}\} \subseteq \{d\}^*$ - **oznaczony,**

$\{a,c\}^* = \{\{x_1\}, \{x_3\},\{x_2\},\{x_4\}, \{x_5, x_6\}\} \subseteq \{d\}^*$ - **oznaczony,**

$\{b,c\}^* = \{\{x_1, x_3\},\{x_2, x_4\}, \{x_5, x_6\}\} \subseteq \{d\}^*$ - **oznaczony.**

Krok trzeci: Wszystkie podzbiory są oznaczone, więc należy przekonwertować te podzbiory. Przekonwertowanie dla podzbioru $\{a, b\}$ wygląda następująco:

$(a,a_1)^* = \{x_1,x_2\}$,

$(a,a_2)^* = \{x_3, x_4\}$,

$(a,a_3)^* = \{x_5, x_6\} \subseteq \{(d, d_3)\}^*$ - **oznaczony,**

$(b,b_1)^* = \{x_2, x_4, x_5, x_6\}$,

$(b,b_2)^* = \{x_1, x_3\} \subseteq \{(d, d_2)\}^*$ - **oznaczony.**

Krok czwarty: Skupiamy się na podzbiorach $(a,a_1)^*$, $(a,a_2)^*$ oraz (b,b_1), ponieważ one są nieoznaczone**.**

$((a,a_1),(b,b_1))^* = \{x_2\} \subseteq \{(d, d_1)\}^*$ - **oznaczony,**

$((a,a_2),(b,b_1))^* = \{x_4\} \subseteq \{(d, d_1)\}^*$ - **oznaczony.**

Ponieważ z przekonwertowania zbiorów $\{a, b\}$ otrzymaliśmy oznaczenia, algorytm się zatrzymuje. Analogicznie postępujemy konwertując zbiory $\{a, c\}$ i $\{b, c\}$. Tab. 3. przedstawia reguły pewne dla zbioru S, które pochodzą ze wszystkich oznaczonych zbiorów.

Tabela 3. Reguły pewne systemu *S*

zbiór {a, b}	zbiór {a,c}	zbiór {b, c}
$(a,a_3) \Rightarrow (d, d_3)$	$(a,a_3) \Rightarrow (d, d_3)$	$(b,b_2) \Rightarrow (d, d_2)$
$(b,b_2) \Rightarrow (d, d_2)$	$(c,c_1) \Rightarrow (d, d_1)$	$(c,c_1) \Rightarrow (d, d_1)$
$(a,a_1) * (b,b_1) \Rightarrow (d, d_1)$	$(a,a_1) * (c,c_2) \Rightarrow (d, d_2)$	$(b,b_1) * (c,c_2) \Rightarrow (d, d_3)$
$(a,a_2) * (b,b_1) \Rightarrow (d, d_1)$	$(a,a_2) * (c,c_2) \Rightarrow (d, d_2)$	X

Źródło: Opracowanie własne

Na podstawie tych reguł można opisać zachowania w systemie. Reguły dla zbioru $\{b, c\}$ można opisać warunkami:

- Jeżeli TSH osoby wynosi b_2 to można jej przypisać chorobę d_2.
- Jeżeli hormon T3 osoby wynosi c_1 to można jej przypisać chorobę d_1.
- Jeżeli TSH osoby wynosi b_1 i T3 wynosi c_2 to można jej przypisać chorobę d_3.
- Reguły możliwe pochodzą z nieoznaczonych zbiorów i zostały zebrane w tab. 4.

Tabela 4. Reguły możliwe dla systemu *S*

zbiór {a, b}	zbiór {a, c}	zbiór {b, c}
$(a,a_1) \Rightarrow (d, d_1)$	$(a,a_1) \Rightarrow (d, d_1)$	$(b,b_1) \Rightarrow (d, d_1)$
$(a,a_1) \Rightarrow (d, d_2)$	$(a,a_1) \Rightarrow (d, d_2)$	$(b,b_1) \Rightarrow (d, d_3)$
$(a,a_2) \Rightarrow (d, d_1)$	$(a,a_2) \Rightarrow (d, d_1)$	$(c,c_2) \Rightarrow (d, d_2)$
$(a,a_2) \Rightarrow (d, d_2)$	$(a,a_2) \Rightarrow (d, d_2)$	$(c,c_2) \Rightarrow (d, d_3)$
$(b,b_1) \Rightarrow (d, d_1)$	$(c,c_2) \Rightarrow (d, d_2)$	X
$(b,b_1) \Rightarrow (d, d_3)$	$(c,c_2) \Rightarrow (d, d_3)$	X

Źródło: Opracowanie własne

Wszystkie reguły występują z prawdopodobieństwem 0.5. Dla zbioru $\{b, c\}$ reguły klasyfikujące można opisać następująco:

- Jeżeli TSH osoby wynosi b_1 to można jej przypisać chorobę d_1 z prawdopodobieństwem 0.5 lub chorobę d_3 również z prawdopodobieństwem 0.5.
- Jeżeli T3 osoby wynosi c_2 to można jej przypisać chorobę d_2 z prawdopodobieństwem 0.5 lub chorobę d_3 również z prawdopodobieństwem 0.5.

3.1.2. Algorytm LEM2

Drugi algorytm z systemu LERS-LEM2, został stworzony z myślą o wydobywaniu reguł z niepełnych systemów. Zajmuje się on znajdowaniem pewnych i możliwych reguł dla danych wartości atrybutów decyzyjnych. Należy założyć, że użytkownik systemu przyjmuje wartość progową dla minimalnego zaufania wstosunku do reguł, które mają być wydobyte z systemu S [14, 15, 17, 18].

Przykład

System S jest systemem niekompletnym $S = (X, A, V)$, gdzie $A = \{a, b, c\}$ są atrybutami klasyfikującymi, a atrybut $\{d\}$ jest atrybutem decyzyjnym. $X = \{ x_1, x_2, x_3, x_4, x_5, x_6 \}$ jest zbiorem obiektów, zaś V jest zbiorem wartości atrybutów ze zbioru A. Przykładowy niekompletny system decyzyjny S przedstawia tab. 5.

Tabela 5. Niepełny system decyzyjny *S*

Pacjent (obiekt)	**Wiek** (atrybut a)	**TSH** (atrybut b)	T3 (atrybut c)	Rodzaj choroby tarczycy (atrybut d)
x_1	a_1	b_2	c_2	d_2
x_2	a_1	b_1	c_1	d_1
x_3	a_2	b_2	c_2	d_2
x_4	a_2	X	c_1	d_1
x_5	a_3	b_1	X	d_3
x_6	a_3	b_1	c_2	d_3

Źródło: Opracowanie własne

Na potrzeby algorytmu przyjęto minimalny próg zaufania $\lambda = 0.75$.

Klasy związane z atrybutem decyzyjnym mają postać:

$d_1^* = \{x_2, x_4\}$,

$d_2^* = \{x_1, x_3\}$,

$d_3^* = \{x_5, x_6\}$.

Klasy związane z atrybutem decyzyjnym można wydobyć z systemu S w kilku krokach, które są prezentowane poniżej.

Krok pierwszy [17,18]:

$a_1^* = \{x_1, x_2\}$ $a_1^* \not\subseteq d_1^*$ nieoznaczone
$a_1^* \not\subseteq d_2^*$ nieoznaczone
$a_1^* \not\subseteq d_3^*$ nieoznaczone
$a_2^* = \{x_3, x_4\}$ $a_2^* \not\subseteq d_1^*$ nieoznaczone
$a_2^* \not\subseteq d_2^*$ nieoznaczone
$a_2^* \not\subseteq d_3^*$ nieoznaczone
$a_3^* = \{x_5, x_6\}$ $a_3^* \subseteq d_3^*$ **oznaczone**
$b_1^* = \{x_2, x_5, x_6\}$ $b_1^* \not\subseteq d_1^*$ nieoznaczone
$b_1^* \not\subseteq d_2^*$ nieoznaczone
$b_1^* \not\subseteq d_3^*$ nieoznaczone
$b_2^* = \{x_1, x_3\}$ $b_2^* \subseteq d_2^*$ **oznaczony**
$c_2^* = \{x_1, x_3, x_6\}$ $c_2^* \not\subseteq d_1^*$ nieoznaczone
$c_2^* \not\subseteq d_2^*$ nieoznaczony
$c_2^* \not\subseteq d_3^*$ nieoznaczone
$c_1^* = \{x_2, x_4\}$ $c_1^* \subseteq d_1^*$ **oznaczone**

Krok drugi: Należy zbudować podzbiory dwuelementowe z podzbiorów nieoznaczonych i jednoelementowych [17,18]:

$(a_1, b_1)^* = \{x_2\}$, $(a_1, b_1)^* \subseteq d_1^*$ **oznaczony**
$(a_1, c_2)^* = \{x_1\}$, $(a_1, c_2)^* \subseteq d_2^*$ **oznaczony**
$(a_2, b_1)^* = \emptyset$ oznaczony, ale nie jest regułą
$(a_2, c_2)^* = \{x_3\}$, $(a_2, c_2)^* \subseteq d_2^*$ **oznaczony**
$(b_1, c_2)^* = \{x_6\}$, $(b_1, c_2)^* \subseteq d_3^*$ **oznaczony**

Dla systemu informacyjnego z tab. 5. zostały po drugim kroku otrzymane wszystkie podzbiory oznaczone. W przypadku innych systemów, w kolejnych *i* krokach należy tworzyć zbiory *i*-elementowe, aż do uzyskania wszystkich podzbiorów oznaczonych w *i-tym* kroku.

Ze wszystkich podzbiorów nieoznaczonych należy wydobyć reguły możliwe przy przyjętym poziomie zaufania $\lambda = 0.75$:

$con\,f\,[a_1^* \subseteq d_1^*] = \frac{1}{2} < \lambda$ nie daje reguły
$con\,f\,[a_1^* \subseteq d_2^*] = \frac{1}{2} < \lambda$ nie daje reguły
$con\,f\,[a_1^* \subseteq d_3^*] = 0 < \lambda$ nie daje reguły
$con\,f\,[a_2^* \subseteq d_1^*] = \frac{1}{2} < \lambda$ nie daje reguły
$con\,f\,[a_2^* \subseteq d_2^*] = \frac{1}{2} < \lambda$ nie daje reguły
$con\,f\,[a_2^* \subseteq d_3^*] = 0 < \lambda$ nie daje reguły
$con\,f\,[b_1^* \subseteq d_1^*] = \frac{1}{3} < \lambda$ nie daje reguły
$con\,f\,[b_1^* \subseteq d_2^*] = 0 < \lambda$ nie daje reguły
$con\,f\,[b_1^* \subseteq d_3^*] = \frac{2}{3} < \lambda$ nie daje reguły
$con\,f\,[c_2^* \subseteq d_1^*] = 0 < \lambda$ nie daje reguły

$con\,f\,[c_2^* \subseteq d_2^*] = \frac{2}{3} < \lambda$ nie daje reguły

$con\,f\,[c_2^* \subseteq d_3^*] = \frac{1}{3} < \lambda$ nie daje reguły

Ostateczne reguły dla niepełnego systemu *S* prezentuje tab. 6.

Tabela 6. Atrybuty systemu informacyjnego chorób niedoczynności tarczycy

Reguła	Opis
$a_3 \rightarrow d_3$	Wiek a_3 ma wpływ na przynależność dolegliwości pacjenta do klasy chorób tarczycy d_3.
$b_2 \rightarrow d_2$	TSH o wartości b_2 ma wpływ na przynależność dolegliwości pacjenta do klasy chorób tarczycy d_2.
$c_1 \rightarrow d_1$	T3 o wartości c_1 ma wpływ na przynależność dolegliwości pacjenta do klasy chorób tarczycy d_1.
$a_1^* b_1 \rightarrow d_1$	Wiek a_1 oraz TSH o wartości b_1 ma wpływ na przynależność dolegliwości pacjenta do klasy chorób tarczycy d_1.
$a_1^* c_2 \rightarrow d_2$	Wiek a_1 lub a_2 oraz T3 o wartości c_2 ma wpływ na przynależność dolegliwości pacjenta do klasy chorób tarczycy d_2.
$a_2^* c_2 \rightarrow d_2$	
$b_1^* c_2 \rightarrow d_3$	TSH o wartości b_1 oraz T3 o wartości c_2 ma wpływ na przynależność dolegliwości pacjenta do klasy chorób tarczycy d_3.

Źródło: Opracowanie własne

3.2. ERID

ERID jest algorytmem służącym do wydobywania reguł klasyfikujących z niekompletnego systemu decyzyjnego [14,19]. W pracy został on przedstawiony na podstawie systemu informacyjnego z tab. 5, analizowanego również w opisie algorytmu LEM2.

W algorytmie ERID na początku należy określić λ_1 – minimalny poziom zaufania oraz λ_2 – minimalny poziom pewności. W poniższej analizie przyjęto $\lambda_{1,2} = 0.75$ [14,19]. Kroki algorytmu zostały przedstawione w tab.7.

Tabela 7. Kroki algorytmu ERID dla niepełnego systemu informacyjnego *S*

Krok pierwszy		
	Zaufanie dla reguły	oznaczenie
$a_1^* = \{x_1, x_2\}$	$\text{conf}(a_1^* \subseteq d_1^*) = \frac{1}{2} < \lambda$, $\text{conf}(a_1^* \subseteq d_2^*) = \frac{1}{2} < \lambda$	nie
$a_2^* = \{x_3, x_4\}$	$\text{conf}(a_2^* \subseteq d_1^*) = \frac{1}{2} < \lambda$, $\text{conf}(a_2^* \subseteq d_2^*) = \frac{1}{2} < \lambda$	nie
$a_3^* = \{x_5, x_6\}$	$\text{conf}(a_3^* \subseteq d_3^*) = 1 > \lambda$	tak
$b_1^* = \{x_2, x_5, x_6\}$	$\text{conf}(b_1^* \subseteq d_1^*) = \frac{1}{3} < \lambda$, $\text{conf}(b_1^* \subseteq d_3^*) = \frac{2}{3} < \lambda$	nie
$b_2^* = \{x_1, x_3\}$	$\text{conf}(b_2^* \subseteq d_2^*) = 1 > \lambda$	tak
$c_1^* = \{x_2, x_4\}$	$\text{conf}(c_1^* \subseteq d_1^*) = 1 > \lambda$	tak
$c_2^* = \{x_1, x_2, x_6\}$	$\text{conf}(c_2^* \subseteq d_1^*) = \frac{1}{3} < \lambda$, $\text{conf}(c_2^* \subseteq d_2^*) = \frac{1}{3} < \lambda$, $\text{conf}(c_2^* \subseteq d_3^*) = \frac{1}{3} < \lambda$	nie
Krok drugi: Zbudowanie zbiorów dwuelementowych ze zbiorów nieoznaczonych		
$(a_1, b_1)^* = \{x_2\}$	$\text{conf}\,[(a_1,b_1)^* \subseteq d_1^*] = 1 > \lambda$	tak
$(a_1, c_2)^* = \{x_1, x_2\}$	$\text{conf}\,[(a_1,c_2)^* \subseteq d_1^*] = \frac{1}{2} < \lambda$, $\text{conf}\,[(a_1,c_2)^* \subseteq d_2^*] = \frac{1}{2} < \lambda$	nie
$(a_2, b_1)^* = \{x_3, x_4\}$	$\text{conf}\,[(a_2,b_1)^* \subseteq d_1^*] = \frac{1}{2} < \lambda$,	nie

	conf $[(a_2,b_1)^* \subseteq d_2^*] = \frac{1}{2} < \lambda$	
$(b_1, c_2)^* = \{x_2, x_6\}$	conf $[(b_1,c_2)^* \subseteq d_1^*] = \frac{1}{2} < \lambda$, conf $[(b_1,c_2)^* \subseteq d_3^*] = \frac{1}{2} < \lambda$	nie
Krok trzeci: Zbudowanie zbiorów trójelementowych ze zbiorów jednoelementowych oraz dwuelementowych, które są nieoznaczone		
$(a_1, b_1, c_2)^* = \{x_2\}$	conf $[(a_1, b_1, c_2)^* \subseteq d_1^*] = 1 > \lambda$	tak

Źródło: Opracowanie własne

Na początku należy utworzyć klasy zależne od atrybutu decyzyjnego {d}:
$d_1^* = \{x_2, x_4\}$,
$d_2^* = \{x_1, x_3\}$,
$d_3^* = \{x_5, x_6\}$.

Następnie, podobnie jak w algorytmie LEM2, należy utworzyć klasy odnoszące się do klasyfikacji atrybutów i ocenić, w jakim stopniu zawierają się one w utworzonych klasach decyzyjnych. Wszystkie klasy, których poziom zaufania jest mniejszy niż przyjęty są nieoznaczone i są wykorzystywane w drugim kroku algorytmu do utworzenia klas dwuelementowych. Ponownie należy obliczyć zaufanie do reguły oraz porównać z przyjętą minimalną wartością. Algorytm ERID zatrzymuje się, gdy wszystkie n-elementowe klasy są oznaczone. Tab. 8. przedstawia reguły wydobyte podczas algorytmu ERID dla systemy decyzyjnego S.

Tabela 8. Reguły klasyfikujące wydobyte algorytmem ERID

Krok pierwszy		
reguła	zaufanie	wsparcie atrybutów
$a_3 \Rightarrow d_3$	1	2
$b_2 \Rightarrow d_2$	1	2
$c_1 \Rightarrow d_1$	1	2
Krok drugi		
$a_1 * b_1 \Rightarrow d_1$	1	1

Źródło: Opracowanie własne

4. Wnioski

Reguły opisują konkretne zależności między danymi w systemie informacyjnym, które pozwalają na sklasyfikowanie nowego pacjenta w bazie medycznej na podstawie wiedzy o innych pacjentach. Porównania tych algorytmów można wykonać dla dwóch algorytmów: LEM2 oraz ERID, które eksplorują wiedzę z systemów o niepełnej informacji. Opis reguł wydobytych algorytmem ERID dla systemu chorób tarczycy jest identyczny jak opis reguł w tab. 6. dla algorytmu LEM2, przy czym dla algorytmu systemu LERS otrzymano więcej reguł. Łącząc obserwacje z dwóch tych algorytmów, można wywnioskować, iż:

- pacjent, który posiada TSH o wartości b_2 z największym prawdopodobieństwem zostanie zaklasyfikowany do grupy pacjentów z chorobą d_2,
- pacjent, który posiada T3 o wartości c_1 z największym prawdopodobieństwem zostanie zaklasyfikowany do grupy pacjentów z chorobą d_1,

- pacjent w wieku a_1 oraz o TSH o wartości b_1 z największym prawdopodobieństwem zostanie zaklasyfikowany do grupy pacjentów z chorobą d_1.

Są to warunki klasyfikacji otrzymane dwoma algorytmami. Algorytm LEM2 wnioskuje szerzej i pozwala stwierdzić, że:

- pacjent w wieku a_1 lub a_2 oraz z T3 o wartości c_2 zakwalifikowany ma zostać do grupy d_2,
- pacjent, którego TSH wynosi b_1 oraz T3 wynosi c_2 będzie zakwalifikowany do grupy d_3.

Są to wnioskowania uogólnione i obowiązujące dla całego systemu decyzyjnego chorób tarczycy.

5. Podsumowanie

Potrzeba przetwarzania i analizowania dużej ilości informacji wymusiła opracowanie metod, które ułatwią pozyskanie, ze stosu wiedzy, tych najbardziej wartościowych informacji. Obok algorytmów drzew decyzyjnych, najbliższych sąsiadów, sztucznych sieci neuronowych czy metod statystycznych prężnie rozwijają się również reguły akcji. Dzięki nim możliwe jest opisanie działania, jakim powinny być poddane atrybuty, aby zmienić aktualny stan obiektu. Reguły akcji opisują nie tylko działanie i skutek, ale odgrywają również znaczącą rolę w klasyfikacji.

Zaprezentowane w pracy algorytmy są przydatnym narzędziem do generowania klasyfikujących reguł akcji na podstawie wiedzy zgromadzonej w medycznym systemie informacyjnym. Na podstawie zależności między atrybutami opisującymi pacjenta a atrybutami decyzyjnymi jesteśmy w stanie sklasyfikować tego pacjenta nie tylko w kompletnych systemach informacyjnych, ale również w tych z niepełną informacją. Na ogół medyczne bazy danych posiadają „luki informacyjne", a algorytmy takie jak LEM2 czy ERID są w stanie przewidzieć brakujące wartości i na tej podstawie przypisać nowy obiekt w bazie do odpowiedniej klasy.

Literatura

1. Wolberg W.H., Mangasarian O.L. *Multisurface method of pattern separation for medical diagnosis applied to breast cytology*, In Proceedings of the National Academy of Science (1987), pp. 9193-9196.
2. Frawley W.J., Piatetsky-Shapiro G., Matheus C. *Knowledge Discovery In Databases*, AAAIPress/MIT Press (1991), pp.1-30.
3. Quinlan J.R., *C4.5: Programs For Machine Learning*, CA: Morgan Kaufmann (1993), pp. 235-240.
4. Witten I.H., Cunningham S.J., Holmes G. *Intelligent data analysis using the WEKA workbench Tutorial Notes*, Conference on Artificial Neural Networks and Expert Systems (1995)
5. Witten I.H., Frank E. Data mining: Practical machine learning tools and techniques with Java implementations, Morgan Kaufmann (2000)
6. Clinton B., New York University speech, Salon.com (2002), http://www.calon.com/politics/feature/2002/12/06/clinton/print.html
7. *Mining Data to Save Children with Brain Tumors*, SPSS (2003), Inc., http://www.dynelytics.com/upload/1123069943_Bioinformatics%20with%20Clementine%202.pdf.
8. Ras Z., Dardzińska A., Tsay L., Wasyluk H. *Association Action Rules*, IEEE International Conference on Data Mining Workshops (2008), pp. 283-290.

9. Ras Z., Dardzińska A.,*Action Rules Discovery without Pre-existing Classification Rules*, Springer, LNAI 5306 (2008),pp. 181-190.
10. Ras Z., Dardzińska A., Liu X.,*System ADRed for discovering rules based on hyperplanes*, Elsevier, Engineering Applications of Artificial Intelligence 17 (2004), pp. 401-406.
11. Dardzińska- Głębocka A., Kosztyła-Hojna B., Łobaczuk-Sitnik A.,*Action Rules Mining in Laryngological Disorders*, Trends in Contemporary Computer Science, Podlasie 2014, pp. 187-193.
12. Dardzińska-Głębocka A., Pauk J.,*New Method for Finding Rules in Incomplete and Distributed Information Systems Controlled by Reducts*, Springer, AMLTA: Advanced Machine Learning Technologies and Applications (2012), pp. 43-51.
13. Dardzińska-Głębocka A., Pauk J., *Mining for Knowledge to Build Decision Support System for Treatment of Plano-Valgus*, Solid State Phenomena, Vol. 199 (2013), pp. 49-54.
14. Dardzińska-Głębocka A. *Action Rules Mining*, Springer (2013), pp. 5-16, 21-28.
15. Grzymala-Busse J., *A new version of the rule induction system LERS*, Fundamenta Informaticae, 31 (1997), Vol.1, pp. 27-39.
16. Ras Z., Dardzińska A.,From Data to Classification Rules and Actions, Rough Sets, Theory and Applications (2011).
17. Dardzińska A., Raś Z.*One Rules Discovery from Incomplete Information Systems*, Proceedings of ICDM'03 Workshop on Foundation and New Directions in Data Mining (2003), pp. 31-35.
18. Im S., Raś Z., Wasyluk H.,*Action rule discovery from incomplete data*, Knowledge Information Systems, 25(1) (2010), pp. 21-33.
19. Dardzińska A., Raś Z. W. *Extracting Rules from Incomplete Decision Systems: System ERID*, Foundations and Novel Approaches in Data Mining, Studies in Computational Intelligence, Springer 9 (2006), pp.143-154.

Reguły akcji jako narzędzie wspomagające klasyfikację chorób

Streszczenie

Dynamiczny wzrost informacji magazynowanych i przetwarzanych przez systemy komputerowe spowodował potrzebę ich szybkiego przetwarzania i ekstrahowania istotnych korelacji między danymi. Próbę rozwiązania tego problemu podejmują techniki eksploracji wiedzy, wśród których są reguły akcji.

W pracy zaprezentowano podstawowe algorytmy klasyfikujących reguł akcji LEM1, LEM2 oraz ERID, które są często podstawą dla wyłonienia z bazy danych już właściwych reguł wskazujących działanie i skutek. Każdy z zaprezentowanych algorytmów został poparty przykładem z medycznej bazy danych chorób tarczycy, a wyłonione reguły zostały przełożone na język opisowy dla wybranego systemu decyzyjnego. Na koniec porównano reguły decyzyjne dwóch algorytmów dla niekompletnego, medycznego systemu decyzyjnego.

Słowa kluczowe: reguły akcji, LEM1, LEM2, ERID

Action Rules as a tool for supporting classification of diseases

Abstract

The dynamic increase of information, collected and processed by computer systems, caused the need for their rapid processing and extraction of relevant correlations between data. Attempts to solve this problem are made by methods of knowledge exploration, and among them are action rules.

In this paper, the basic algorithms of classification action rules are presented. LEM1, LEM2 and ERID are the basis for identifying the right rules indicating action and effect. Each of the presented algorithms was supported by an example from the medical database of thyroid diseases, and selected rules was described using less formal language. At the end of this work, the extracted classification rules from medical, incomplete decision system are compared.

Keywords: Action rules, LEMI1, LEM2, ERID

Katarzyna Ignatiuk[1], Agnieszka Dardzińska-Głębocka[2]

Wybrane metody eksploracji wiedzy z systemu informacyjnego na bazie chorób tarczycy

1. Wstęp

We współczesnym świecie dostęp do informacji jest nieograniczony, ale nie każda informacja jest wartościowa. Nabiera ona znaczenia dopiero, gdy można określić jej korelacje w stosunku do innych informacji. Tym właśnie zajmują się algorytmy eksploracji danych (ang. *data mining*). Pozwalają odnaleźć w zbiorze danych nowe, znaczące i skorelowane informacje na podstawie licznych zbiorów danych. Do szacowania, przewidywania, klasyfikacji, grupowania czy odkrywania reguł podczas procesu odkrywania wiedzy z baz danych, standardowo wykorzystywane są metody statystyczne i matematyczne.

Medyczne systemy informacyjne posiadają ogromne zasoby danych na temat pacjentów, ich stanu zdrowia i procesów leczenia. Określenie relacji i stworzenie wzorców na podstawie zgromadzonych danych może dostarczyć użytkownikowi dodatkowej wiedzy medycznej. Dlatego też w ostatnich latach zauważa się gwałtowny rozwój metod *data mining* w medycynie. Są one wykorzystywane szczególnie w celu klasyfikacji chorób, pacjentów jak też w celach predykcyjnych.

Rosnące zapotrzebowanie na pozyskiwanie informacji z dużych baz danych wpłynęło na dynamiczny rozwój algorytmów eksploracji wiedzy oraz narzędzi, które te algorytmy wykorzystują. Biorąc to pod uwagę, na potrzeby niniejszej pracy dokonano oceny skuteczności klasyfikacji obiektów bazy chorób tarczycy algorytmami dostępnymi w oprogramowaniu WEKA. Analizy dokonano na podstawie danych pacjentów z niedoczynnością tarczycy. Szerzej opisane i poparte przykładem zostały dwa algorytmy klasyfikujące, dzięki którym otrzymano niemalże 100% poprawność klasyfikacji zbioru testowego.

2. System informacyjny na bazie chorób tarczycy

Aby zrozumieć istotę eksploracji wiedzy z medycznych baz danych, a w szczególności z bazy chorób tarczycy, w rozdziale tym zawarto najważniejsze informacje na temat hormonów tarczycy i ich wpływu na zdrowie człowieka. Nawiązano do tych atrybutów klasyfikujących (hormonów) i decyzyjnych (schorzeń), które odegrały kluczową rolę w opisie danych zaczerpniętych z repozytorium programu WEKA [1].

W rozdziale tym opisano również system informacyjny, jaki został utworzony na podstawie danych pacjentów badanych pod kątem niedoczynności tarczycy.

[1] ignatiukkatarzyna@gmail.com, Zakład Biocybernetyki i Inżynierii Biomedycznej, Wydział Mechaniczny, Politechnika Białostocka,
[2] a.dardzinska@pb.edu.pl, Zakład Biocybernetyki i Inżynierii Biomedycznej, Wydział Mechaniczny, Politechnika Białostocka

2.1. Hormony tarczycy i skutki zaburzeń ich produkcji

Z problemami związanymi ze schorzeniami tarczycy w Polsce zmaga się duża grupa osób. Niestety istnieje jednak niewiele dokładnych informacji na ten temat. Wiele osób w naszym kraju nie poddało się nigdy badaniom laboratoryjnym w kierunku chorób tarczycy. Nieleczone schorzenie tarczycy wiąże się z występowaniem licznych powikłań. W Stanach Zjednoczonych, jak wykazują badania, niepracująca prawidłowo tarczyca dotyczy ok. 3% dorosłych [2,3]. Ośmiokrotnie częściej zaburzenia w pracy tarczycy dotykają kobiet niż mężczyzn. Choroby tarczycy znacznie częściej występują również u osób, które obarczone są innymi chorobami autommunologicznymi [2] takimi jak cukrzyca typu 1., reumatoidalne zapalenie stawów, toczeń, jak też u osób starszych przyjmujących niektóre leki.

Diagnostyka chorób tarczycy często trwa latami z tego powodu, iż objawy dysfunkcji tarczycy są niespecyficzne. Drżenie, kołatanie serca, nerwowość występująca u osób z nadczynnością tarczycy można pomylić z innymi chorobami serca, zaś objawy typowe dla niedoczynności, takie jak osłabienie, zmęczenie, senność, zapominanie można wytłumaczyć szybkim tempem życia, przeziębieniem [2, 3]. Dodatkowo trudność w identyfikacji schorzeń tarczycy wynika z tego, że jest ona organem pracującym w niezauważalny dla człowieka sposób i nie jest on w stanie sam zauważyć pewnych anomalii w pracy gruczołu.

Do ogólnego podziału schorzeń tarczycy należą [2]:

- niedoczynność tarczycy – tarczyca nie jest w stanie wytwarzać i wydzielać odpowiedniej ilości hormonów tarczycy, aby utrzymać stan eutyreozy,
- nadczynność tarczycy – tarczyca wytwarza i wydziela zbyt dużą ilość hormonów tarczycy,
- guzki i wole – tarczyca może pracować prawidłowo, ale powiększa się (wole) lub wytwarza guzki.

Tyroksyna (T4), trójjodotyronina (T3) oraz kalcytonina wymieniane są wśród najważniejszych hormonów odpowiedzialnych za prawidłową pracę organizmu. Są to hormony produkowane przez tarczycę, które stymulują wzrost organizmu, gospodarkę wapniowo-fosforanową, mają istotne znaczenie w rozwoju układu nerwowego oraz wspomagają przemianę materii [2]. Zaburzenie ich produkcji wpływa negatywnie na fizyczne i psychiczne procesy zachodzące w organizmie człowieka. Osoby z zaburzoną gospodarką hormonalną tarczycy mogą cierpieć na ospałość, zaburzenia krążenia, zaburzenia pracy serca, problemy z wagą, stany depresyjne, marznięcie dłoni i stóp, wysuszenie skóry, czy też wypadanie włosów [2].

Prawidłowa praca metabolizmu człowieka zależna jest bezpośrednio od poziomu hormonu T3, który jest wytwarzany z hormonu T4 przy pomocy dejonaz [2]. Rola kalcytoniny w układzie hormonalnym jest mniej znacząca. Pracą tarczycy steruje przysadka mózgowa poprzez tyreotropinę (TSH) [2].

2.2. Baza chorób tarczycy

Do celów niniejszej pracy pobrano bazę chorób tarczycy utworzoną we współpracy Instytutu Badań Medycznych w Garvan w Sydney z J.Ross Quinlan'em. Została ona udostępniona w repozytorium oprogramowania WEKA [1] w formacie *arff*.

Pobrane z repozytorium dane utworzyły niekompletny system informacyjny. Zbudowały go 3772 obiekty, a każdy z tych obiektów był zdefiniowany przez 30 atrybutów, do których należały:

- atrybuty z wartościami numerycznymi: wiek, TSH, T3, TT4, T4U, FTI, TBG,
- atrybuty z odpowiedzią TAK/NIE: przeziębienie, ciąża, leczenie tyroksyną, '*Czy zapytano o leczenie tyroksyną*', leczenie przeciwtarczycowe, leczenie jodem radioaktywnym T131, operacja tarczycy, niedoczynność, nadczynność, przyjmowanie litu, guz, wole, niedoczynność przysadki, badania psychologiczne, '*Czy zmierzono TSH*', '*Czy zmierzono T3*', '*Czy zmierzono TT4*', '*Czy zmierzono T4U*', '*Czy zmierzono FTI*', '*Czy zmierzono TBG*',
- atrybut płeć – K/M,
- atrybut źródło skierowania – WEST/ STMW/ SVHC/ SVI/ SVHD/ inne,
- atrybut klasa obiektu – atrybut decyzyjny.

Atrybutem decyzyjnym względem, którego dokonano klasyfikacji danych, była klasa obiektu. W utworzonym systemie informacyjnym przyjmowała ona cztery wartości:

- negatywna – świadczyła o braku zdiagnozowanej niedoczynności tarczycy,
- wyrównywana niedoczynność tarczycy – nawiązywała do niedoczynności tarczycy kompensowanej lekami,
- pierwotna niedoczynność tarczycy – miała związek z niedoczynnością spowodowana uszkodzeniem gruczołu tarczycy np. na skutek przebytego zapalenia tarczycy, wycięcia gruczołu tarczycy, leczenia jodem radioaktywnym T131, napromieniowaniem, przedawkowaniem leków przeciw-tarczycowych, przyjmowaniem leków: soli litu, interferonu, sulfonamidów,
- wtórna niedoczynność tarczycy – wskazywała na nabytą niedoczynność tarczycy spowodowaną niedoborem TSH, związanym z niedoczynnością przysadki. W badaniach charakteryzowała się niskim lub prawidłowym stężeniem TSH oraz niskim T4 i T3.

3. Wybrane metody eksploracji wiedzy z systemu informacyjnego

Algorytmy eksploracji wiedzy mają coraz szersze zastosowanie w analizie danych medycznych. Wykorzystywane są głównie do celów predykcyjnych oraz do szybszego i bardziej precyzyjnego podejmowania decyzji szczególnie, gdy dotyczą one zdrowia i leczenia pacjenta [3-5].

Ogólnodostępnym narzędziem wykorzystywanym w eksploracji wiedzy z medycznych baz danych jest WEKA [4, 6]. Zostało one stworzone przez zespół badaczy z Uniwersytetu w Waikato w Nowej Zelandii. WEKA jest przyjaznym w obsłudze oprogramowaniem, posiadającym wbudowane algorytmy uczenia maszynowego z możliwością dostosowania algorytmów do charakteru importowanych danych,

automatyczną analizą oraz tworzeniem własnych algorytmów *data mining* [4, 6]. Narzędzie to zostało wykorzystane również na potrzeby niniejszej pracy. Dla wybranej bazy danych dokonano klasyfikacji obiektów zbioru testowego. Na podstawie wyników klasyfikacji wybrano i opisano dwa algorytmy eksploracji wiedzy, które niemalże bezbłędnie sklasyfikowały obiekty.

3.1. Algorytm J48

Algorytm J48 jest implementacją algorytmu drzew decyzyjnych C4.5, który buduje drzewa ze zbioru uczącego z wykorzystaniem entropii (teorii informacji (2)) [7-9]. Polega on na rekursywnym odwiedzaniu każdego węzła decyzyjnego i wybraniu możliwego podziału. Aby wybrać optymalny podział zbioru uczącego w algorytmie wyliczany jest zysk informacji, którego wzór reprezentuje zależność (3) [7].

Załóżmy, że *S* definiuje podział zbioru uczącego *T* na podzbiory T_1, T_2, T_3, ... T_k. Średnie zapotrzebowanie na informacje można obliczyć, jako ważoną sumę entropii $H_s(T_i)$ dla pojedynczych podzbiorów [7]:

$$H_s(T) = -\sum_{i=1}^{k} P_i H_s(T_i) \qquad (1)$$

gdzie P_i reprezentuje procent rekordów w i-tym podzbiorze.

Entropia jest najmniejszą liczbą bitów dla pojedynczego podzbioru T_i, która umożliwia przesłanie informacji na temat węzła. Węzeł może przyjąć *k* możliwych wartości, których wystąpienie określają prawdopodobieństwa p_1, p_2, ... p_k [7]. Wówczas entropię można określić wzorem:

$$H_s(T_i) = -\sum_j p_j \log_2(p_j) \qquad (2)$$

Na podstawie wzoru (1) i (2) można wyliczyć zysk informacyjny [7]:

$$zysk = H(T) - H_s(T) \qquad (3)$$

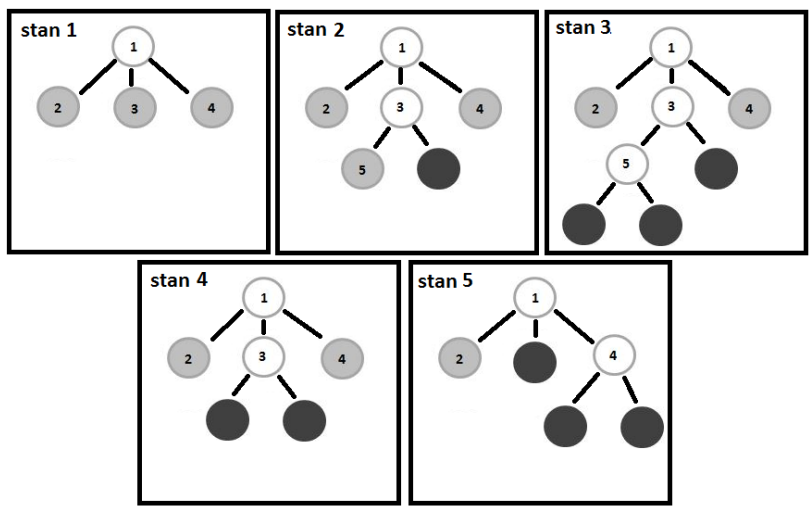
Dla każdego węzła decyzyjnego, algorytm selekcjonuje optymalny podział z najwyższym zyskiem informacyjnym.

Kroki algorytmu [7, 9]:

a) Obliczenie zysku informacyjnego dla każdego atrybutu.
b) Oznaczenie głównym węzłem atrybutu z największym zyskiem.
c) Wyprowadzenie gałęzi od głównego atrybutu (korzenia) do pozostałych atrybutów. Obliczenie znormalizowanej entropii dla każdego atrybutu i porównanie jej z wartością sprzed podziału.
d) Wybranie gałęzi, która prowadzi do najwyższego zysku informacyjnego.
e) Powtórzenie kroków **a) - d)** jeżeli różnorodność węzłów jest duża.
f) Zatrzymanie algorytmu, jeżeli wszystkie przypadki osiągną tę samą wartość.

3.2. Algorytm PART

Kluczem algorytmu PART jest budowa częściowego drzewa decyzyjnego, na podstawie którego jest odkrywana wiedza w postaci reguł. Drzewo częściowe jest zwykłym drzewem decyzyjnym, które podlega operacjom konstruowaniu i przycinania, aż do momentu znalezienia stabilnego poddrzewa, którego w dalszym etapie nie da się uprościć. Gdy tylko zostanie odnalezione drzewo częściowe, konstruowana jest reguła, zaś drzewo odrzucane (rys.1). Pozwala to uniknąć generalizacji reguł i nadmiernego rozbudowywania poddrzewa, jak dzieje się w przypadku budowania reguł naiwnymi metodami. Wykorzystując metodę rozdzielania i zwyciężania w drzewach decyzyjnych, zwiększa się wrażliwość i szybkość algorytmu wydobywającego regułę. Algorytm nie wymaga optymalizacji danych.



Rysunek 1. Przykład drzewa częściowego konstruowanego i przycinanego algorytmem PART. Kolejne etapy przedstawiają decyzje podejmowane przez algorytm PART [12]

Na rys. 2. przedstawiono algorytm budowania i przycinania drzewa decyzyjnego. Poniższe kroki opisują przykładowy algorytm działania dla drzewa z rys. 2. [10]:

a) Podział zbioru przypadków na odpowiednie podzbiory.
b) Wybór podzbioru o najmniejszej entropii – podzbiór 3. Szare węzły są jeszcze nierozszerzone. Czarne węzły są liśćmi. Podzbiór 3 ma dzieci w postaci węzła i liścia.
c) Wybór węzła 5 do rozszerzenia drzewa. Dziecko węzła 3 - czarny węzeł, ma niższą entropię niż jego rodzeństwo - węzeł 5, ale nie może być rozszerzony dopóki jest liściem.
d) Sprawdzenie czy korzystniejsze jest zastąpienie węzła 5 pojedynczym liściem. Ocena, czy błąd estymacji dla poddrzewa jest ≥ niż błąd estymacji dla korzenia tego poddrzewa. Jeśli tak, to trzeba zastąpić węzeł liściem.
e) Zastąpienie węzła 5 liściem.

f) Obliczenie błędu estymacji poddrzewa węzła 3 i samego węzła 3, w związku z tym, że dzieci są liśćmi.
g) Zastąpienie węzła 3 liściem.
h) Analiza rodzeństwa nowo zastąpionego węzła 3. Wybór węzła z najniższą entropią – węzła 4.
i) Wytypowanie węzła 4 do zastąpienia liściem ze względu na posiadanie dzieci liści.
j) Założenie, że węzeł 4 nie jest zamieniany na liść. Koniec algorytmu z drzewem częściowym z trzema liśćmi.

Celem implementacji tego algorytmu jest znalezienie najbardziej ogólnej reguły, wybierając drzewo częściowe - liść, który pokrywa największą liczbę przypadków [10].

4. Analiza wyników

Na potrzeby analizy zbiór danych został podzielony na zbiór uczący i zbiór testowy, który posłużył potwierdzeniu wiarygodności wyników otrzymanych na podstawie zbioru uczącego. W zbiorze testowym znalazło się 467 obiektów wybranych losowo z pełnego zbioru danych (3772 obiektów).

Dane podzielone na zbiór uczący i testowy poddano klasyfikacji z wykorzystaniem każdego algorytmu dostępnego w programie WEKA. Tab. 1 przedstawia zestawienie metod, dla których osiągnięto najlepsze rezultaty klasyfikacji zbioru testowego, biorąc pod uwagę liczbę przypadków poprawnie sklasyfikowanych, współczynnik Kappa oraz średni błąd bezwzględny.

Najlepsze rezultaty otrzymano dla klasyfikacji danych algorytmem J48. Z użyciem tego algorytmu sklasyfikowano nieprawidłowo tylko 1 obiekt. Klasyfikacja tą metodą pozwoliła osiągnąć wysoki współczynnik Kappa (1 świadczy o 100% zgodności danych) oraz niski średni błąd kwadratowy.

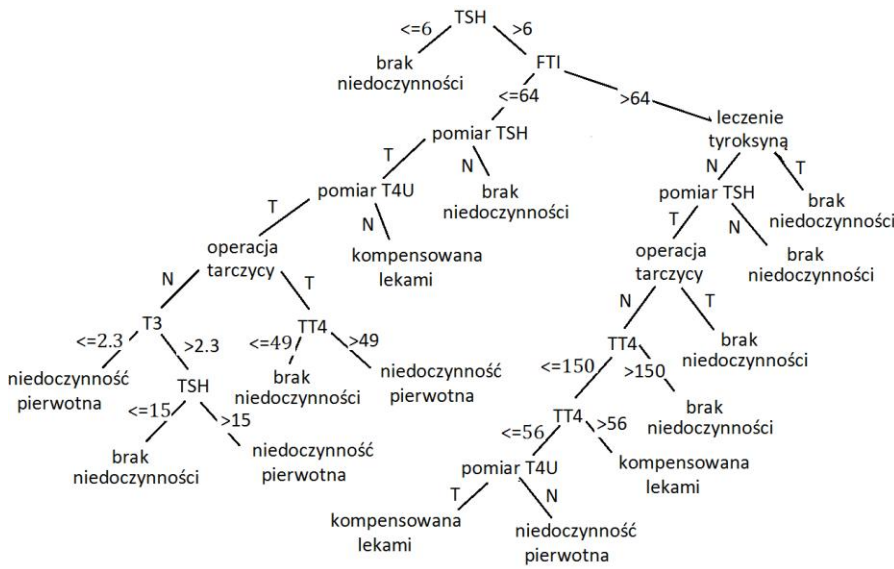
Równie skuteczny okazał się algorytm PART, dla którego osiągnięto porównawalne wyniki. Zestawienie przeprowadzonych klasyfikacji przedstawia tab.1.

Tabela 1. Atrybuty systemu informacyjnego chorób niedoczynności tarczycy

Algorytm	Odsetek poprawnie sklasyfikowanych obiektów [%]	Liczba niepoprawnych klasyfikacji	Współczynnik Kappa	Średni błąd bezwzględny
J48	99.7859	1	0.9865	0.0022
PART	99.7859	1	0.9863	0.0029
Random Committee	99.7859	1	0.9865	0.0116
BayesNet	98.7152	6	0.9208	0.0098
kNN dla k=6	94.0043	28	0.4243	0.0486

Źródło: Opracowanie własne

W tab. 1. przedstawiono również wyniki klasyfikacji dla trzech innych algorytmów. Każdy z nich pochodzi z innej rodziny algorytmów. Najsłabsze wyniki otrzymano w przypadku dość popularnego algorytmu *kNN* (*k najbliższych sąsiadów*) dla przyjętej wartości k=6. Dla wybranych danych doświadczalnych okazał się on najgorszą opcją. Wpływ na ten wynik mógł mieć fakt, iż między klasami obiektów nie ma wyraźniej granicy. Wyniki dwóch najdokładniejszych algorytmów prezentuje rys. 2. i tab. 2.



Rysunek 2. Drzewo decyzyjne utworzone dla wybranej bazy chorób tarczycy [opracowanie własne]

Tabela 2. Reguły i decyzje systemu informacyjnego chorób niedoczynności tarczycy

Reguła → Decyzja
[TSH ≤ 6] → brak niedoczynności
[FTI ≤ 64] * [zmierzone TSH = TAK] * [zmierzone T4U = TAK] * [operacja tarczycy = NIE] * [T3 ≤ 2.3] → niedoczynność pierwotna
[leczenie tyroksyną = NIE] * [zmierzone TSH = TAK] * [operacja tarczycy = NIE] * [48<TT4 ≤ 150] * [zmierzone TT4 = TAK] * [FTI > 65] → wyrównywana niedoczynność tarczycy
[TT4 > 66] * [zmierzone T4U = TAK] → brak niedoczynności tarczycy
[zmierzone TSH = NIE] → brak niedoczynności
[źródło skierowania = SVI] → pierwotna niedoczynność tarczycy
[źródło skierowania = inne] * [zmierzone TT4 = TAK] * [płeć =K] * [T3 ≤ 2.4] → brak niedoczynności tarczycy
[płeć = mężczyzna] → brak niedoczynności tarczycy

Źródło: Opracowanie własne

Tab.2. zawiera reguły utworzone algorytmem PART dla systemu informacyjnego niedoczynności tarczycy. Na ich podstawie można wysunąć następujące wnioski:

- Jeżeli FTI≤64 oraz T3≤2.3 to prawdopodobnie pacjent ma niedoczynność pierwotną,
- W przypadku, gdy pacjent ma TT4 w przedziale <48,150>, FTI>65 i nie jest leczony tyroksyną, należy wyrównać mu poziom hormonów tarczycy,
- Większość pacjentów, którzy otrzymali skierowanie z SVI ma pierwotną niedoczynność tarczycy,
- Niedoczynność tarczycy nie została zdiagnozowana u mężczyzn, u kobiet z T3 ≤ 2.4 i osób z TT4 > 66 lub TSH ≤ 6.

Dane testowe poddano również klasyfikacji z wykorzystaniem walidacji krzyżowej z podziałem zbioru danych na 10 podzbiorów. Do celów tej analizy połączono zbiór uczący i testowy. Zastosowana w klasyfikacji metoda walidacji krzyżowej polegała na podziale próby badawczej na 10 podzbiorów, następnie każdy z nich traktowany jest jako zbiór testowy, zaś pozostałe, połączone podzbiory, jako zbiór uczący. Analiza została powtórzona 10 razy, a wyniki 10-ciu rezultatów uśredniono. Tab. 3. przedstawia zestawienie wyników klasyfikacji z walidacją krzyżową.

Tabela 3. Zestawienie wyników walidacji krzyżowej systemu informacyjnego chorób niedoczynności tarczycy

Algorytm	Odsetek poprawnie sklasyfikowanych obiektów [%]	Liczba niepoprawnych klasyfikacji	Współczynnik Kappa	Średni błąd bezwzględny
J48	99.5758	16	0.9707	0.003
PART	99.4168	22	0.9598	0.0035
Random Committee	99.2577	28	0.9488	0.0126
BayesNet	98.5949	53	0.9028	0.011
kNN dla k=6	93.4783	246	0.2804	0.0494

Źródło: Opracowanie własne

Zastosowanie walidacji krzyżowej nieznacznie zmniejszyło odsetek poprawnie sklasyfikowanych obiektów przez każdy z przedstawionych w tab. 3. algorytmów. Zauważalnie zwiększyła się liczba nieprawidłowych klasyfikacji obiektów. Dla algorytmu J48 liczba ta wzrosła z 1 do 16 obiektów, zaś dla algorytmu PART z 1 do 22 obiektów.

5. Podsumowanie

Na podstawie przeprowadzonej analizy algorytmów klasyfikujących obiekty w bazie chorób tarczycy można stwierdzić, iż algorytmy eksploracji wiedzy są przydatnym narzędziem wspomagającym podejmowanie decyzji odnośnie metod leczenia pacjenta oraz przewidywań stanów zdrowia, jakie może on osiągnąć w przyszłości. Nie każdy jednak algorytm klasyfikujący posiada taką samą skuteczność. Zależy ona od rodzaju analizowanych danych, czy są to wartości numeryczne czy symboliczne, czy utworzony system informacyjny jest pełny czy niepełny, czy miedzy klasami danych istnieją wyraźne granice itp. Uwzględniając te wszystkie czynniki można dobrać dla konkretnych danych najbardziej optymalny algorytm eksploracji wiedzy, który w stosunkowo krótkim czasie dokona niemalże 100% prawidłowej klasyfikacji.

Dla wybranej bazy danych najdokładniejsze wyniki uzyskano dzięki zastosowaniu algorytmu J48 klasyfikując dane testowe w oparciu o wiedzę ze zbioru uczącego. Kolejny algorytm drzew decyzyjnych PART przyczynił się również do wyłonienia 8 pewnych reguł dla bazy chorób tarczycy. Na podstawie takiej wiedzy można określić kolejne etapy badań i leczenia pacjenta z niedoczynnością tarczycy. W przypadku diagnozy związanej z brakiem niedoczynności u pacjenta, należałoby pamiętać, iż przeprowadzona analiza nie świadczy o prawidłowym poziomie hormonów tarczycy u badanych osób. Analiza nie wyklucza nadczynności tarczycy oraz innych schorzeń, pod kątem których dane nie były analizowane. Tego typu wątpliwości może rozwiać konsultacja otrzymanych wyników z ekspertem (lekarzem).

Literatura:

1. http://repository.seasr.org/Datasets/UCI/arff/
2. Emanuel O. Brams *Thyroid Disease*, A Case-Based and Practical Guide for Primary Care, Humana Press (2005), s. 3-9.
3. Upadhayay A., Shukla S., Kumar S. *Empirical Comparison by data mining Classification algorithms (C4.5 & C5.0) for thyroid cancer data set*, International Journal of Computer Science & Communication Networks, Vol 3(1) (2013), s. 64-68.
4. Othman M.F., Yau T.M.S. *Comparison of Different Classification Techniques Using WEKA for Breast Cancer*, Biomed 06, IFMBE Proceedings 15 (2007), pp. 520-523.
5. Podgorelec V., Kokol P., Rozman I. *Decision Trees:An overview and their use in medicine*, Journal of Medical Systems, Vol. 26 (2002), s. 445-463.
6. Frank E., Hall M., Holmes G., Kirkby R., Pfahringer B., Witten I.H. *WEKA – A Machine Learning Workbench for Data Mining*, Data Mining and Knowledge Discovery Handbook:A Complete Guide for Practitioners and Researchers, Springer (2005), s. 1305-1314.
7. Larose D.T. *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, Wydawnictwo Naukowe PWN (2006), s. 118-120.
8. Hand D., Mannila H., Smyth P. *Eksploracja danych*, Wydawnictwo Naukowo-Techniczne, Warszawa (2005), s.181-200, 394-402.
9. Sharma J., Agrawal J., Sharma S.,*Classification Through Machine Learning Technique: C4.5 Algorithm based on various entropies*, International Journal of Computer Applications, Vol. 8 (2013), s. 20-27.
10. Frank E., Witten I.H.*Generating Accurate Rule Sets Without Global Optimization*,ICML'98 Proceedings of the Fifteenth International Conference on Machine Learning (1998), s.144-151.

Wybrane metody eksploracji wiedzy z system informacyjnego na bazie chorób tarczycy

Streszczenie

Medyczne bazy danych posiadają ogromne zasoby danych na temat pacjentów, ich stanu zdrowia oraz leczenia. Dane te nabierają znaczenia, dopiero, gdy można określić korelacje między nimi. Tym właśnie zajmują się algorytmy eksploracji wiedzy (*ang. data mining*). Na podstawie zgromadzonej wiedzy tworzą wzorce charakterystyczne dla danego systemu informacyjnego, które następnie wykorzystują do klasyfikacji nowych obiektów w systemie i prognozowania zachowań systemu.

W pracy zaprezentowane zostały dwie wybrane metody eksploracji wiedzy, dla których osiągnięto najlepsze wyniki klasyfikacji z wykorzystaniem programu WEKA. Klasyfikacji poddano dane z bazy chorób tarczycy, pobranej z internetowego repozytorium. Klasyfikację próby testowej przeprowadzono wszystkimi dostępnymi w WECE algorytmami, ale opisano i przeanalizowano dwa o niemalże 100% prawidłowej klasyfikacji oraz o niskim średnim błędzie bezwzględnym. Algorytm J48 sklasyfikował zestaw danych uczących, zaś algorytm PART wyłonił dodatkowo reguły klasyfikujące dane w oparciu o model drzewa decyzyjnego. Praca nad algorytmami *data mining* przyniosła również pomysły na własne implementacje reguł akcji w oprogramowaniu WEKA.

Słowa kluczowe: data mining, J48, PART

Selected methods of knowledge discovery from information system based on thyroid diseases

Abstract

Medical databases have accumulated large quantities of information about patients, their medical condition and treatment. This data become useful only when correlations between them exist. These relationships in databases are found by data mining algorithms. Based on accumulated knowledge, algorithms create specific patterns for information systems and they are useful to classification new objects in system and to predict behaviour of system.

The paper presents two selected methods of knowledge discovery. For them has been achieved the best classification results using WEKA. The classification was based on data from thyroid diseases database taken from the WEKA repository.

The classification of the training set was done for all available algorithms in WEKA, but only two had 100% correct classification with low mean absolute error. The J48 algorithm classified the training set, and PART algorithm extracted rules for classification based on decision tree model. The work with data mining algorithms provided on ideas to implement more action rules algorithms in WEKA.

Keywords: data mining, algorithm J48, algorithm PART

Ewelina Nadzieja[1], Michał Woś[1], Marian Jędrych[2,] Ewelina Firlej[3], Mariola Janiszewska[3]

Nowoczesne technologie medyczne w pracy z pacjentem

1. Wstęp

W dzisiejszych czasach nowoczesne technologie medyczne obecne są w wielu sferach życia. Wraz z postępem nauki wprowadzane są coraz bardziej nowoczesne rozwiązania, aby usprawnić proces diagnozy i leczenia. Wiąże się to ze sporymi nakładami finansowymi, które ponoszone są w celu opracowania i wdrożenia innowacyjnych technologii. Dzięki temu możliwe jest usprawnienie wielu czynności związanych z pracą lekarza. Istotną rolę w procesie leczenia odgrywa sprawna wymiana informacji, co pozwala zaoszczędzić czas, który z punktu widzenia personelu medycznego jest szczególne cenny.

Wymiana informacji i danych o pacjencie są kluczowymi elementami oraz zasobami każdego szpitala, wartością integrującą całą instytucję. Występujące systemy zarządzają informacją pochodzącą z różnych źródeł informacji, od ludzi, maszyn, z danych oraz procesów. Ich zadaniem jest optymalne tworzenie, organizowanie, przekształcanie oraz udostępnienie informacji w odpowiednim miejscu i czasie [1].

Nowoczesne technologie jak i wbudowane w infrastrukturę szpitali systemy informatyczne w znacznym stopniu ułatwiają wymianę informacji miedzy pacjentem a jednostkami służby zdrowia. Stosując takie systemy jak wysyłanie widomości SMS lub e-maili do przypomnienia wizyty czy podawanie numerków i obliczanie przez system przybliżonej możliwości wizyty, pacjent nie musi oczekiwać na wizytę w długich kolejkach w poczekalni.

Jednostkami które badają problem elektronicznych systemów gromadzenia informacji w służbie zdrowia od kilku lat zajmują się badacze ze Stanów Zjednoczonych oraz Wielkiej Brytanii. Jest to związane z faktem, że w tych państwach jest największa liczba szpitali w których są wdrożone elektroniczne rekordy pacjenta. Państwa te są także które wdrażają w swoją służbę zdrowia najnowsze technologie takie jak: „e-recepta", „zapisy elektroniczne skanów ciała z TK", „automatyczne zapisy do karty pacjenta", „Wdrażanie elektronicznej karty pacjenta na systemy Android oraz IOS (urządzenia mobilne)" itp.. Obecnie w polskiej służbie zdrowia po wprowadzeniu ustawy z dnia 28 kwietnia 2011r., która została znowelizowana z dniem 11 lutego 2017r. O systemie informacji w ochronie zdrowia, która nakłada obowiązek

[1]ewelinanadzieja@wp.pl;m.wos.official@gmail.com,Studenckie Koło Naukowe przy Zakładzie Matematyki i Biostatystyki Medycznej, Wydział Nauk o Zdrowiu, Uniwersytet Medyczny w Lublinie
[2]marian.jedrych@umlub.pl, Zakład Matematyki i Biostatystyki Medycznej, Wydział Nauk o Zdrowiu, Uniwersytet Medyczny w Lublinie
[3]ewefir@wp.pl;mariola.janiszewska@gmail.com, Katedra Zdrowia Publicznego, Wydział Nauk o Zdrowiu, Uniwersytet Medyczny w Lublinie

pracownikom na obowiązek gromadzenia informacji o pacjencie w sposób elektroniczny. Dlatego przebieg leczenia, monitorowania oraz diagnostyki stanowi jeden z istotnych elementów codziennych zajęć pracowników służby zdrowia.

2. Korzyści wynikające ze stosowania nowoczesnych technologii w opiece zdrowotnej

Istnieje wiele korzyści związanych z wdrażaniem nowoczesnych technologii w opiece zdrowotnej oraz z ich powszechnym wykorzystaniem. Z punktu widzenia pacjentów i pracowników medycznych są to:

- Poprawa jakości usług medycznych ,
- Zwiększenie dostępności specjalistycznych usług medycznych,
- Możliwość wyboru lekarza oraz oceny jego kompetencji,
- Natychmiastowy dostęp do danych medycznych przez służby ratownicze,
- Natychmiastowy dostęp do istotnych danych medycznych z dowolnego miejsca na świecie,
- Przyspieszenie procesu diagnostyki,
- Wspomaganie procesów decyzyjnych związanych z diagnostyką i leczeniem,
- Możliwość stosowania elektronicznych recept i skierowań,
- Minimalizacja niepotrzebnych lub dublujących się usług medycznych[2],
- Redukcja liczby błędów medycznych,
- Stworzenie podstawy informacyjnej i technologicznej do rozwoju innych usług powiązanych z EKP: e-recept, e-skierowań, e-kolejek [3].

3. Rodzaje nowoczesnych technologii najczęściej stosowanych w opiece zdrowotnej

3.1. Elektroniczna dokumentacja medyczna

Elektroniczna dokumentacja medyczna zawiera dane personalne i informacje medyczne które odnoszą się do stanu zdrowia pacjenta takie jak historia choroby czy podawanych leków, badania diagnostyczne, zapisy z postępów rehabilitacji oraz udzielonych świadczeń zdrowotnych. Dane te muszą być opatrzone datą powstania oraz umożliwiać jednoznaczną identyfikacje osoby je wprowadzającej. Obecnie obowiązuje nas rozporządzenie Ministra Zdrowia z dnia 9 listopada 2015 r. w sprawie rodzajów, zakresów i wzorów dokumentacji medycznej oraz sposobu jej przetwarzania które dopuszcza na gromadzenie i przechowywanie indywidualnej dokumentacji medycznej w formie cyfrowej pod warunkiem zapewniania selektywnego dostępu do zgromadzonych informacji [4].

3.2. Elektroniczny rekord pacjenta

Elektroniczny rekord pacjenta jest elektroniczną formą zapisu danych na temat stanu zdrowia, przebiegu leczenia, wypisywania recept, skierowań czy zwolnień oraz wydawania leku dla konkretnego pacjenta spełniająca jednocześnie wymogi kliniczne, administracyjne czy prawne [5]. W ramach elektronicznego rekordu pacjenta

gromadzone są również obrazowe wyniki badań pochodzące z urządzeń diagnoztycznych takich jak: CT, CR, MR, USG, PET itp. Elektroniczny rekord pacjenta służy także do gromadzenia informacji na temat leczenia radioterapeutycznego w zakresie prezentacji planów leczenia i przebiegu naświetlania oraz na zapisywanie zgromadzonych danych na temat leków, terapii lekowej, czy świadczeń udzielanych w różnych jednostkach terapeutycznych [6].

Obecnie leczenie pacjenta przebiega wieloośrodkowo w wyniku czego ważne jest przekazywanie danych o stanie zdrowia oraz przebiegu terapii pomiędzy różnymi placówkami ochrony zdrowia aby zapewnić ciągłość terapii oraz zastosować wiele rekordów medycznych należących do tej samej osoby. Taki indywidualny zbiór danych nazywany jest Elektronicznym Rekordem Medycznym lub Elektronicznym Rekordem Zdrowotnym [7]. Gromadzenie i udostępnianie upoważnionym jednostkom opieki zdrowotnej kompleksowej informacji o stanie zdrowia obywateli należy do obowiązku państwa. W Europie systemy gromadzenia danych medycznych wdrożono w Danii i Wielkiej Brytanii. Natomiast w Polsce w ramach projektu P1[8] czyli Elektroniczna Platforma Gromadzenia, Analizy i Udostępniania zasobów cyfrowych o zdarzeniach medycznych trwają dalsze prace, które w przyszłości umożliwią swobodną wymianę pomiędzy wszystkimi systemami ochrony zdrowia. Założenia projektu zakładają wymianę elektronicznej dokumentacji medycznej pomiędzy lekarzami rodzinnymi, szpitalami, poradniami, aptekami, ZUS, NFZ oraz ratownictwem medycznym.

3.3. Aplikacje mobilne jako alternatywa dla komputerów

Jedną ze zmian jaką zamierza wprowadzić Ministerstwo Zdrowia ma być przeniesienie dokumentacji medycznej z komputera na urządzenia mobilne. Dzięki takiemu rozwiązaniu lekarz będzie mógł kontrolować pracę za pomocą urządzeń przenośnych. Przykładem będzie karta pacjenta na urządzenia mobilne, co pozwoli na sprawdzenie informacji na temat leczenia pacjenta przenośnie, w szpitalu bądź w domu chorego.

Duży nacisk kładzie się na systemy e-recepty, e-skierowania czy e-zwolnienia w zastosowaniu aplikacji mobilnych. Nowelizacja ustawy z dnia 28 kwietnia 2011r., która została znowelizowana z dniem 23 lipca 2014r. w ochronie zdrowia, dotyczą przesunięcia terminu na dzień 1 sierpnia 2017r. obowiązkowego prowadzenia dokumentacji medycznej tylko w postaci elektronicznej.

3.4. Zastosowanie systemów informatycznych przez lekarzy

Jednym z zaawansowanych obszarów zastosowań informatyki w medycynie jest telemedycyna. Jedną z funkcji telemedycyny jest możliwość komunikacji lekarza z pacjentem na odległość, świadczenie usług medycznych odbywa się wirtualnie. W tym celu wykorzystuje się urządzenia pomiarowe, rozmowę telefoniczną, wideofonferencje. Lekarz w tym przypadku ma do czynienia głównie z informacją o pacjencie uzyskaną z Elektronicznego Rekordu Pacjenta, natomiast pacjent jest monitorowany, badany i ewentualnie także konsultowany i instruowany w związku z profilaktyką i terapią. Zastosowanie telemedycyny jest bardzo przydatne w przypadku objęcia opieką medyczną pacjentów do których trudno dotrzeć z tradycyjnymi formami usług

medycznych np. wioski znacznie oddalone od szpitali. Ta forma medycyny jest także przydatna marynarzom znajdującym się na morzu, uczestnikom egzotycznych wypraw itp. [9]. Telemedycyna jest użyteczna także w odniesieniu do osób po zabiegach operacyjnych i rekonwalescentów, których stan zdrowia można monitorować w sposób zdalny, nie narażając ich na utratę zdrowia oraz dyskomfort związany z koniecznością wizyty u lekarza. W tym przypadku najbardziej ważna jest rola opieki telemedycznej dla pacjentów przewlekle chorych. Dobrze przemyślane rozwiązania telemedyczne pozwalają im na prawidłowe funkcjonowanie [10]

Zaleta zastosowania technik telemedycznych polega także na tym, że dzięki użyciu nowoczesnych technologii informatycznych można wstępnie analizować stan zdrowia pacjenta. W tej sytuacji najbardziej przydatną technologią stają się coraz częściej smartfony, a także smartwatche, za pomocą których można badać puls oraz natlenienie krwi. Istnieją także specjalne urządzenia podłączane do telefonu, za pomocą których można też badać np. poziom cukru we krwi. Informacje z wyżej wymienionych urządzeń automatycznie trafiają do Elektronicznego Rekordu Pacjenta [11].

W systemie teleinformatycznym istnieją także laboratoria analityczne, personel sprawujący opiekę domową oraz liczni konsultanci, z których rad mogą korzystać zarówno pacjenci, jak i opiekujący się nimi lekarze. System telemedycyny zawsze jest osadzony w konkretnym terenie i zawiera komponentę związaną z pacjentami oraz część odbiorczą za pomocą której personel medyczny analizuje dane [12].

4. Cel pracy

Celem niniejszej pracy było zbadanie wykorzystania nowoczesnych technologii przez lekarzy w pracy z pacjentem. Osiągnięcie tego celu było możliwe po uzyskaniu odpowiedzi na następujące pytania:

- Czy środowisko medyczne posługuje się w swojej pracy nowoczesnymi technologiami informatycznymi?
- Ile czasu na pracę z systemami informatycznymi poświęcają respondenci?
- Jakie systemy są najczęściej wykorzystywane?
- Czy zastosowanie telemedycyny ma wpływ na zmniejszenie ilości obowiązków lekarzy?
- Czy w opinii badanych, zastosowanie elektronicznej karty pacjenta na smartfony/tablety ułatwi wizyty z pacjentem?
- Czy zdaniem respondentów- e-recepta, e-skierowanie, e-zwolnienie wpływa na skrócenie czasu pracy lekarza?

5. Materiał i metoda badawcza

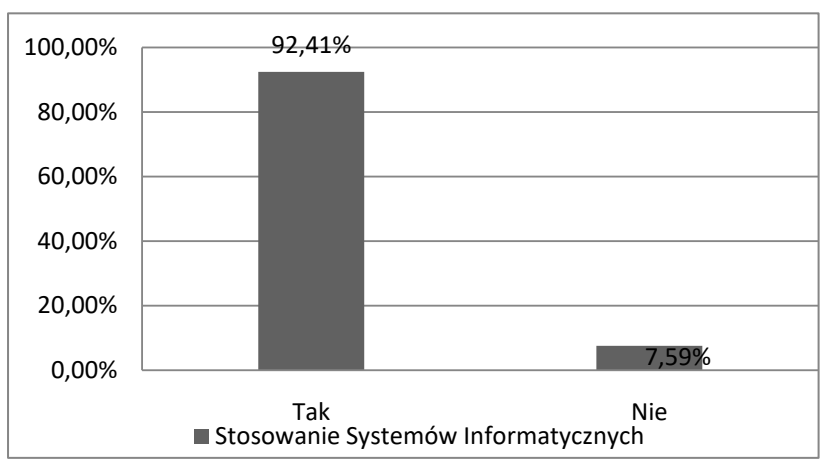
Badania były realizowane od listopada 2015 do maja 2016 roku. Grupę badaną stanowili czynni zawodowo lekarze, pracujący w wybranych placówkach medycznych na terenie całego kraju. W badaniu zastosowano metodę sondażu diagnostycznego, natomiast narzędziem badawczym był kwestionariusz ankiety własnego autorstwa. Uzyskany materiał badawczy poddano analizie statystycznej za pomocą programu Statistica w wersji 12 oraz Microsoft Excel w wersji 2013. W grupie osób badanych

znalazło się 158 respondentów, wiek badanych mieścił się w przedziale od 29 do 61 lat. W grupie badanej było 48% kobiet i 52% mężczyzn. Ankietowani lekarze posiadali różne specjalizacje. Największą grupa badanych byli lekarze o specjalizacji internetycznej(23,42%). Badania zostały przeprowadzone po uzyskaniu wcześniejszej zgody od następujących podmiotów w których zatrudnieni są lekarze uczestniczący w ankiecie:

- Samodzielny Publiczny Szpital Kliniczny nr 1 we Wrocławiu,
- Samodzielny Publiczny Centralny Szpital Kliniczny w Warszawie,
- Wojewódzki Szpital Zespolony w Toruniu,
- Samodzielny Publiczny Zakład Opieki Zdrowotnej Ministerstwa Spraw Wewnętrznych w Gdańsku,
- Samodzielny Publiczny Zakład Opieki Zdrowotnej Ministerstwa Spraw Wewnętrznych w Krakowie,
- Kliniczny Szpital Wojewódzki Nr 2 im. Św. Jadwigi Królowej w Rzeszowie,
- Samodzielny Publiczny Szpital Kliniczny Nr 2 PUM w Szczecinie,
- Samodzielny Publiczny Szpital Kliniczny Nr 4 w Lublinie,
- Samodzielny Publiczny Szpital Kliniczny Nr 1 w Lublinie,
- Szpital Kliniczny im. H. Święcickiego UM w Poznaniu,
- Uniwersyteckie Centrum Kliniczne im.prof. K. Gibińskego Śląskiego Uniwersytetu Medycznego w Katowicach.

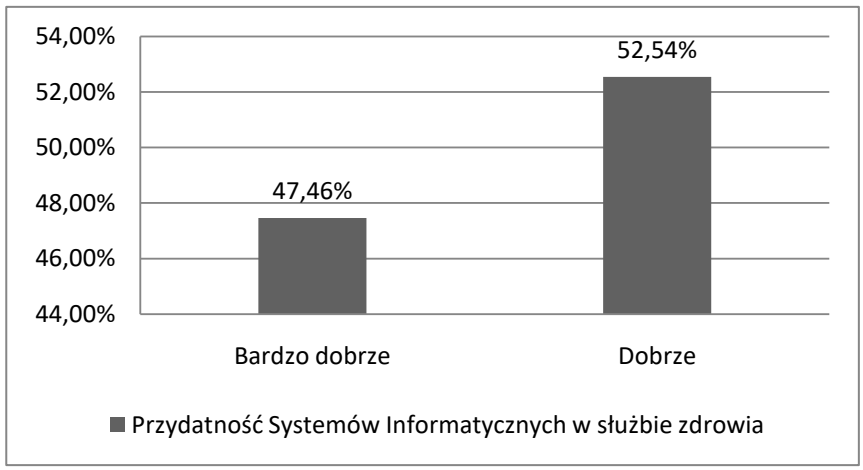
Na podstawie odpowiedzi respondentów określono zestawy czynników istotnych statystycznie takich jak: płeć, miejsce badań, specjalizacje czy podkreślenie czynników ważnych w analizie. Miejsca badań podzielono na województwa: pomorskie i wielkopolskie, lubelskie i podkarpackie, śląskie i małopolskie oraz mazowieckie. Lekarzy podzielono na zapobiegawczych oraz zabiegowych.

6. Wyniki badań i omówienie



Wykres 1. Stosowanie systemów informatycznych w opiece zdrowotnej Źródło: [Opracowanie własne]

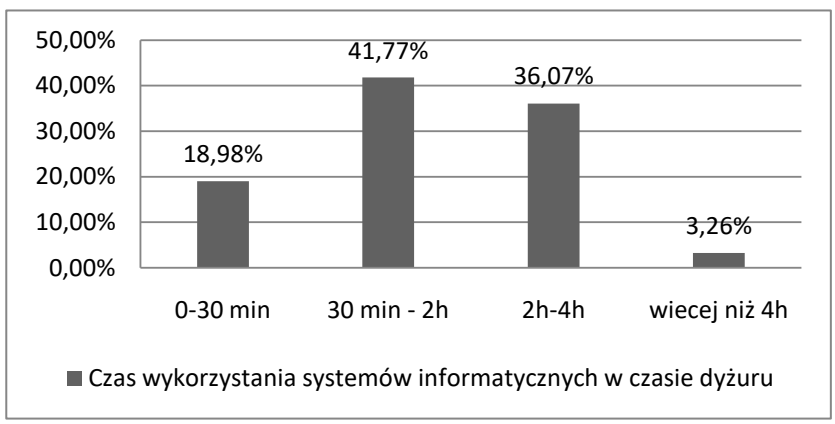
Badani stosują nowoczesne technologie, które są zaimplementowane w systemach informatycznych (92,41%) jedynie niewielka grupa (7,59%) nie używa systemu.



Wykres 2.Przydatność systemów w opiece zdrowotnej

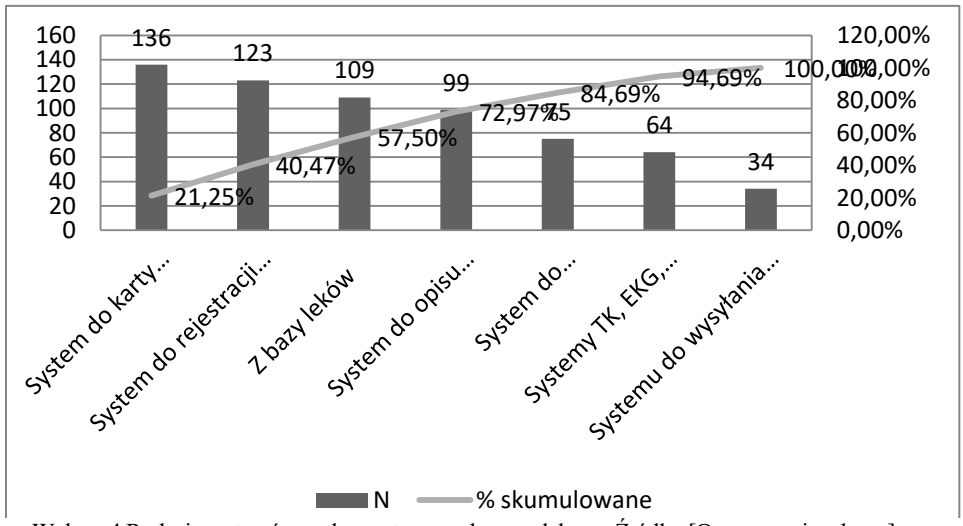
Źródło: [Opracowanie własne]

Badani określili przydatność systemów informatycznych na poziomie bardzo dobrym (52,54%). Z kolei (47,46%) uważa, że przydatność systemów informatycznych w opiece zdrowotnej jest na poziomie dobrym.



Wykres 3.Czas jaki poświęcają lekarze na korzystanie z systemów informatycznych w czasie dyżuru
Źródło: [Opracowanie własne]

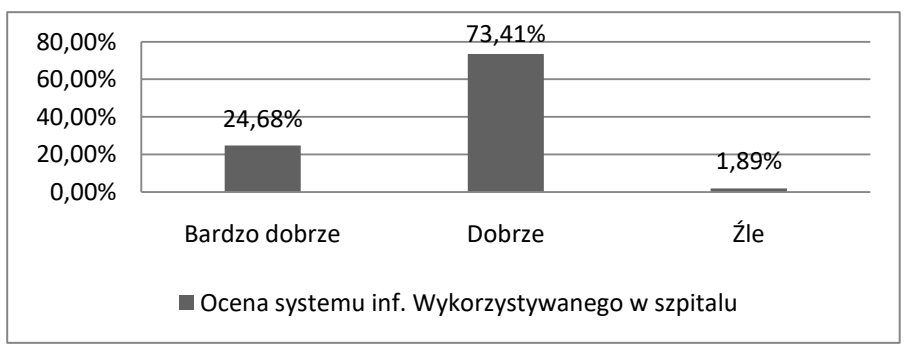
Większość ankietowanych (41,77%) korzysta z systemów informatycznych w czasie od 30 min do 4 godzin. Natomiast powyżej 2 godzin do 4 godzin wykorzystuje (36,07%) respondentów.



Wykres 4.Rodzaje systemów wykorzystywanych przez lekarzy Źródło: [Opracowanie własne]

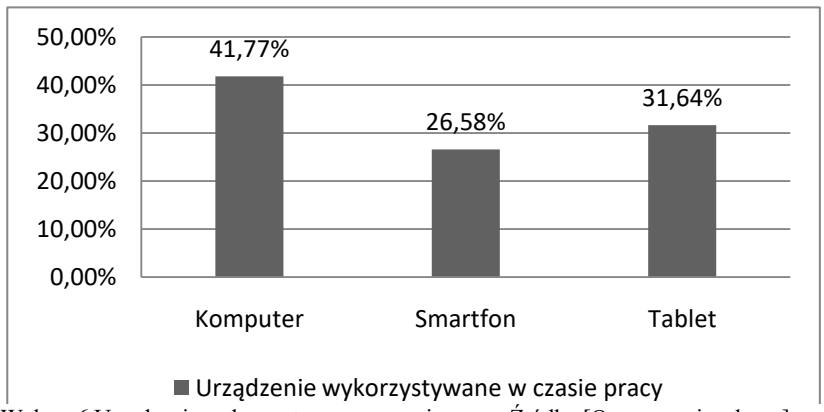
Jak wynika z wykresu lekarze najczęściej korzystają z systemu do kart pacjenta (21,25%). Pozostałe odpowiedzi ankietowanych rozkładają się następująco:

- System do rejestracji pacjenta(19,22%),
- Z bazy leków (17,03%),
- Systemu opisu przypadku(15,47%),
- Systemu do poszukiwania choroby (11,72%),
- Systemy TK, EKG oraz rezonansu magnetycznego(10,00%),
- System do wysyłania karetek (6,31%).



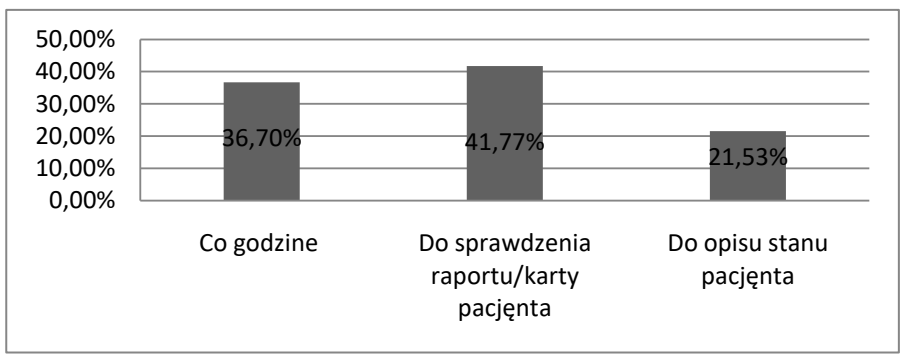
Wykres 5.Ocena systemu informatycznego wykorzystywanego w szpitalu. Źródło: [Opracowanie własne]

Badani stwierdzili, że systemy medyczne stosowane w ich miejscu pracy sprawują się bardzo dobrze (24,68%) oraz dobrze(73,41%). Tylko nieliczna grupa wskazała że system działa źle (1,89%).



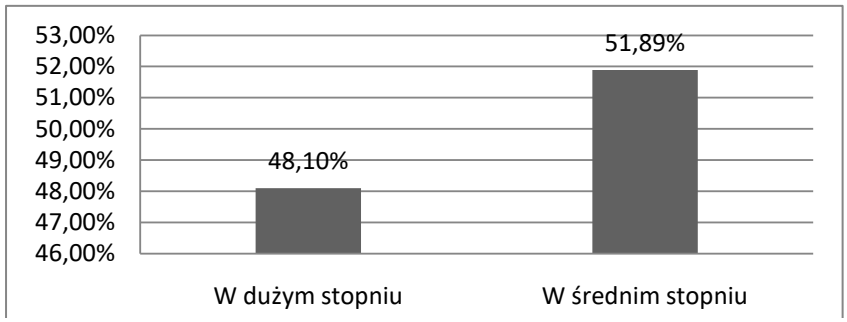
Wykres 6.Urządzenia wykorzystywane w czasie pracy. Źródło: [Opracowanie własne]

Komputer (41,77%)okazał się najczęściej stosowanym urządzeniem do zarządzania systemami informatycznymi w szpitalach. Następnymi urządzeniami jaki są stosowane w czasie pracy są tablety (31,64%) jak i smartfony(26,58%).



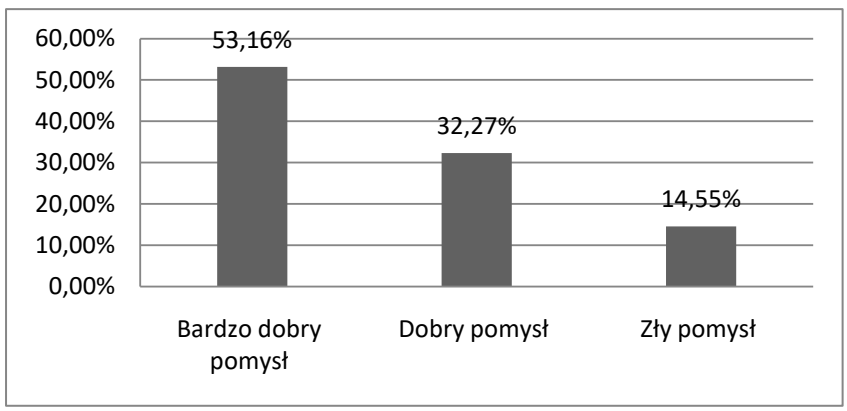
Wykres 7.Czas wykorzystywania nowoczesnych technologii w pracy z pacjentem. Źródło: [Opracowanie własne]

Czas jaki lekarze poświęcają na stosowanie nowoczesnych technologii w pracy z pacjentem jest zróżnicowany. Analiza wykazała, że najczęściej stosują te urządzenia do sprawdzenia raportu/karty pacjenta (41,77%).



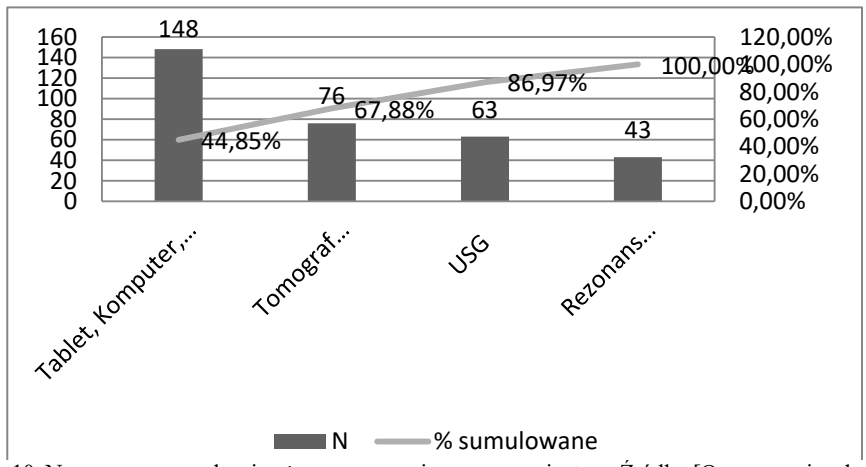
Wykres 8.Wykorzystanie telemedycyny w celu zwolnienia lekarza z niektórych obowiązków. Źródło: [Opracowanie własne]

Ponad połowa respondentów stwierdziła, że zastosowanie nowoczesnej telemedycyny do zwolnienia lekarza z niektórych obowiązków jest przydatne w średnim stopniu(51,89%). Natomiast 48,10% badanych uznało, że zastosowanie telemedycyny jest przydatne w dużym stopniu.



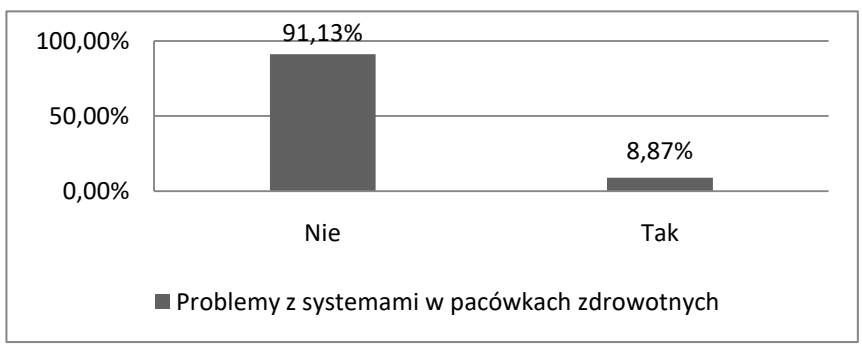
Wykres 9.Ocena wykorzystania nowoczesnej telemedycyny w codziennej pracy lekarza. Źródło: [Opracowanie własne]

Ponad połowa respondentów zapytana o ocenę wykorzystania nowoczesnej telemedycyny w codziennej pracy lekarza z pacjentem odpowiedziała, że jest to bardzo dobry pomysł(53,16%) lub dobry pomysł(32,27%). Tylko niewielka grupa (14,55%) stwierdziła, że jest to zły pomysł.



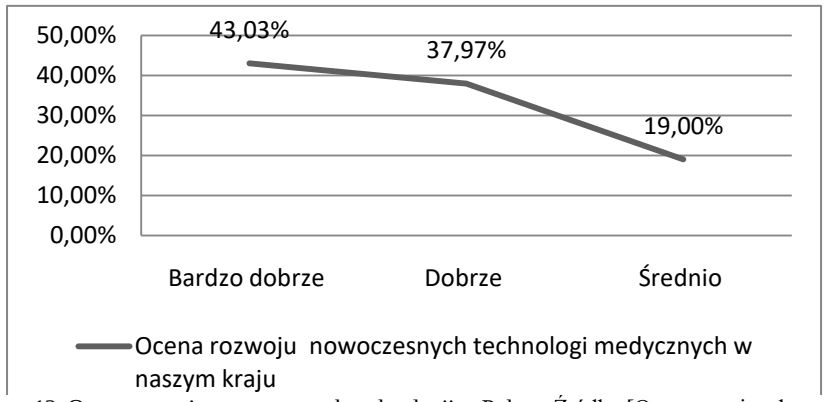
Wykres 10. Nowoczesne urządzenia używane w czasie pracy z pacjentem. Źródło: [Opracowanie własne]

Personel medyczny najczęściej korzysta z urządzeń typu komputer, smartfon oraz tablet (54,85%), natomiast inne urządzania takie jak tomograf komputerowy (23,03%), USG (19,09%) oraz rezonans magnetyczny (13,03%) są stosowane rzadziej.



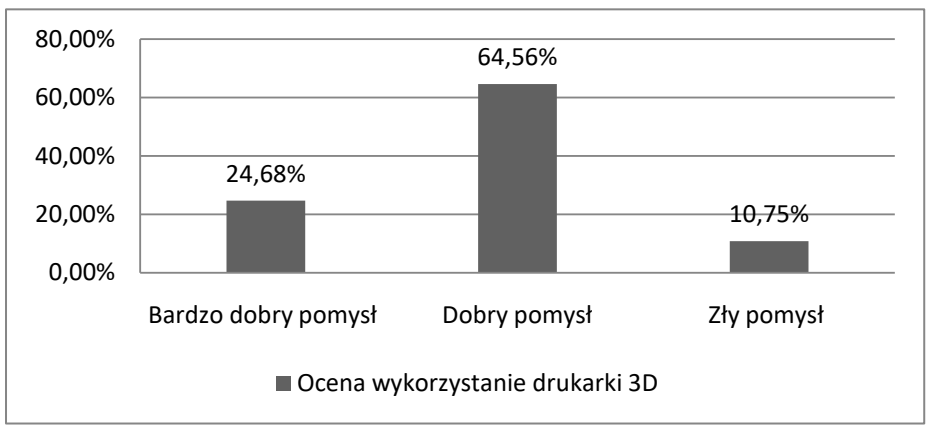
Wykres 11. Problemy z najnowszymi technologiami. Źródło: [Opracowanie własne]

Ankietowani nie wykazali problemu związanego ze stosowaniem nowoczesnych technologii w codziennej pracy(91,13%). Jedynie (8,87%) wykazuje problem.



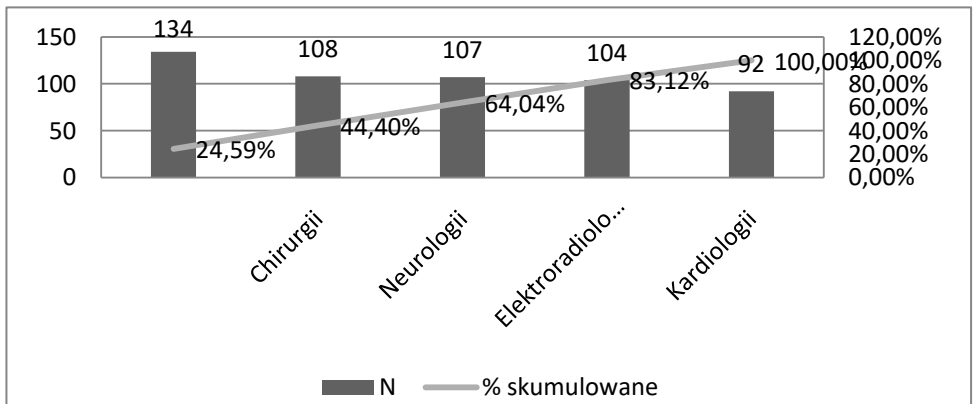
Wykres 12. Ocena rozwoju nowoczesnych technologii w Polsce. Źródło: [Opracowanie własne]

Większość ankietowanych ocenia rozwój nowoczesnych technologii medycznych w Polsce jako bardzo dobry(43,03%) oraz dobry(37,97%). Niewielka grupa badanych odpowiedziała, że ocenia rozwój nowoczesnych technologii na poziomie średnim (19,00%).



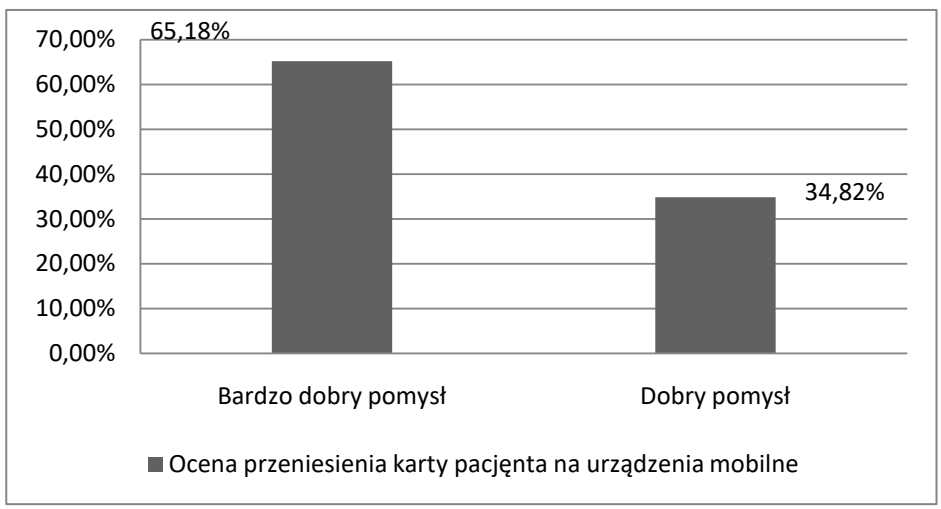
Wykres 13.Wykorzystanie nowoczesnych drukarek 3D w medycynie. Źródło: [Opracowanie własne]

Większość respondentów ocenia pozytywnie zastosowanie drukarek 3D w medycynie. Jako bardzo dobry pomysł uznało 24,65% respondentów oraz na dobry (64,56%). Jako zły pomysł oceniło (10,75%) badanych.



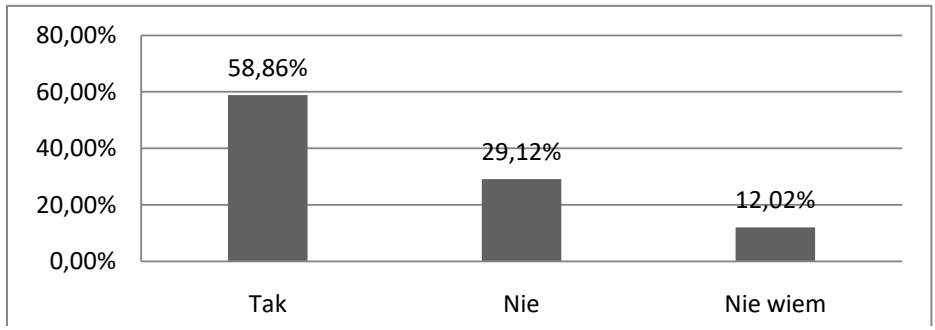
Wykres14.Kierunek rozwoju nowoczesnych technologii w specjalizacjach medycznych.
Źródło:[Opracowanie własne]

Badani stwierdzili, że wykorzystanie nowoczesnych technologii będzie stosowane w przyszłości częściej w ratownictwie medycznym (24,14%), elektrokardiologii (20,36%) oraz neurologii (19,28%).



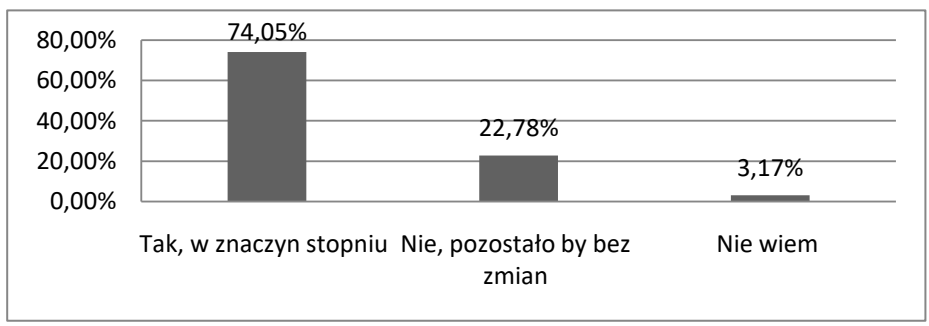
Wykres 15.Ocena przeniesienia karty pacjenta na urządzenie mobilne. Źródło: [Opracowanie własne]

Ponad połowa badanych oceniła przeniesienie karty pacjenta na urządzenia mobilne na poziomie bardzo dobrym (65,18%) oraz pozostali respondenci na poziomie dobrym (34,82%).



Wykres 16. Możliwość korzystania z mobilnej wersji e-recepty, e-skierowanie i e-zwolnienia.
Źródło:[Opracowanie własne]

Ponad połowa respondentów stwierdziło, że korzystałoby z mobilnej wersji e-recepty, e-skierowania oraz e-zwolnienia (58,86%). Natomiast nie korzystałoby z mobilnej wersji (29,12%).



Wykres 17. Ocena wykorzystania e-recepty do zmniejszenia kolejek do specjalistów.
Źródło: [Opracowanie własne]

Prwie trzy czwarte ankietowanych stwierdziło, że zastosowanie e-recepty w znacznym stopniu przyczyni się do zmniejszenia kolejek do specjalistów(74,05%).

Czynniki istotne statystycznie

Tabela 1. Płeć a rodzaj urządzeń wykorzystywanych przez badanych

Płeć	Z jakich urządzeń korzystają badani			
	Komputer	Smartfon	Tablet	Wiersze razem
Kobiety	39(51,31%)	22(28,94%)	15(19,73%)	76
Mężczyźni	27(32,92%)	20(24,39%)	35(42,68%)	82
Ogół	66	42	50	158
Chi^2 Pearsona 6,197056; p=0,045				

Źródło: [Opracowanie własne]

Analiza statystyczna wykazała istotny związek pomiędzy płcią a urządzeniem wykorzystywanym w pracy(p = 0,045). Kobiety(51,31%) częściej korzystają z komputera niż mężczyźni (32,92%). Kobiety (28,94%) również częściej używają smartfonów niż mężczyźni (42,68%) którzy bardziej przekonani są do tabletów.

Tabela 2. Grupy zawodowe a wykorzystanie drukarki 3D w medycynie

Podział na Grupy Zawodowe	Wykorzystanie drukarki 3D w medycynie			
	Bardzo dobry pomysł	Dobry Pomysł	Zły pomysł	Wiersz razem
Zabiegowe	21(19,09%)	74(67,27%)	15(13,64%)	110
Zachowawcze	20(41,67%)	28(58,33%)	0(0%)	48
Ogół	41	102	15	158
Chi^2 Pearsona 7,786983; p=0,02037				

Źródło: [Opracowanie własne]

Analiza statystyczna wykazała istotny związek między grupami zawodowymi a wykorzystaniem drukowania 3D w medycynie(p=0,02037). Badani którzy pracują w jednostkach zabiegowych i zachowawczych twierdzą, że zastosowanie drukarek 3D w czasie pracy jest dobrym pomysłem

Tabela 3. Miejsce zamieszkania w grupach a czas poświęcony na korzystanie z nowoczesnych technologii

Miejsce zamieszkania w grupach	Czas poświęcony na korzystanie nowoczesnych technologii			
	0-30 min	30 – 2h	2h-4h	Wiersz Razem
Dolnośląskie, Śląskie i Małopolska	15(29,41%)	17(33,33%)	20(39,21%)	51
Pomorskie, Kujawsko Pomorskie i Wielkopolskie	11(24,44%)	17(37,78%)	17(37,78%)	45
Lubelskie i Podkarpackie	5(14,71%)	18(52,94%)	11(32,35%)	34
Mazowieckie	2(7,14%)	14(50,00%)	12(42,86%)	28
Ogół	32	66	60	158
Chi^2 Pearsona 17,28767; p=0,04440				

Źródło: [Opracowanie własne]

Na podstawie powyższej tabeli można stwierdzić, że miejsce zamieszkania ma znaczący wpływ na poświęcany czas na korzystanie z nowoczesnych rozwiązań informatycznych. Okazało się bowiem, że ankietowani w województwach pomorskie, kujawsko-pomorskie i wielkopolska (37,78%), lubelskie i podkarpackie(52,94%) oraz

mazowieckie (50,00%) korzystają z urządzeń od 30 min do 2h. Natomiast w województwach śląskich oraz małopolsce badani korzystają z urządzeń od 2h do 4h. Analiza statystyczna wykazała znaczący poziom istotności(p=0,04440).

Tabela 4.Miejsce zamieszkania w grupach a zastosowanie telemedycyny do zwolnienia lekarza z niektórych obowiązków

Miejsce zamieszkania w grupach	Zastosowanie telemedycyny do zwolnienia lekarza z niektórych obowiązków		
	W dużym stopniu	W średnim stopniu	Wiersz ogółem
Dolnośląskie, Śląskie i Małopolska	25(49,02%)	26(50,98%)	51
Pomorskie, Kujawsko Pomorskie i Wielkopolskie	19(42,22%)	26(57,78%)	45
Lubelskie i Podkarpackie	10(29,41%)	24(70,59%)	34
Mazowieckie	22(78,57%)	6(21,43%)	28
Ogół	76	82	158
Chi^2 Pearsona 15,81101; p=0,00124			

Źródło: [Opracowanie własne]

Miejsce zamieszkania ma wpływ na postrzeganie zastosowania telemedycyny do zwolnienia lekarza z niektórych obowiązków. W województwach pomorskim, kujawsko-pomorskim i wielkopolskim(50,98%) lubelskim i podkarpackim(57,78%) oraz dolnośląskim, śląskim i małopolskim(50,98%) badani stwierdzili, że zastosowanie w przyszłości telemedycyny przez lekarzy zwolni ich z niektórych obowiązków w średnim stopniu. Natomiast ankietowani z województwa mazowieckiego(78,57%) twierdzą, że w rozwiązania telemedyczne w dużym stopniu zwolnią lekarzy z niektórych obowiązków wobec pacjentów.

Tabela 5.Możliwości wypisywania e-recepty a wykorzystanie e-recepty do zmniejszenia kolejek

Możliwości wypisywania e-recepty e-skierowania e-zwolnienia.	Wykorzystanie e-recepty, e-skierowania, e-zwolnienia do zmniejszenia kolejek do lekarzy		
	Tak, w znacznym stopniu	Nie, pozostało by bez zmian	Wiersz Razem
Tak	90(96,77%)	3(3,23%)	93
Nie	21(45,65%)	25(54,35%)	46
Nie wiem	8(42,10%)	11(57,90%)	19
Ogół	119	39	158
Chi^2 Pearsona 86,38133; p=0,0000001			

Źródło: [Opracowanie własne]

Analiza statystyczna wykazała bardzo istotny związek pomiędzy możliwością wypisywania e- recepty, e-skierowania i e –zwolnienia a wykorzystaniem e-recepty, e-skierowania oraz e-zwolnienia do zmniejszenia kolejek do lekarzy (p=0,0000001). Jak można zauważyć badani jednomyślnie odpowiedzieli, że jeżeli mieliby możliwość wypisywania e-recept, e-skierowań oraz e-zwolnień (96,77%) zmniejszyłoby to kolejki w znacznym stopniu. Natomiast osoby które zaznaczyły odpowiedź, że nie korzystałyby z udogodnień e-recepty, e-skierowania oraz e-zwolnienia(54,35%), stwierdziły, że kolejki pozostały by bez zmian.

7. Wnioski

- Jak można zauważyć, lekarze bardzo chętnie stosują nowoczesne rozwiązania informatyczne w swojej pracy,
- Według badanych czas jaki poświęcają w czasie dyżuru wynosi od 30 min do 2 godzin. Najczęściej korzystają z komputera, który według nich jest najlepszym urządzeniem w czasie pracy z pacjentem,
- Badani najczęściej korzystają z systemu do opisu stanu pacjenta, systemu do rejestracji pacjenta oraz z dostępnej bazy leków,
- Ankietowani wyrazili chęć stosowania e-recepty, e-skierowania oraz e-zwolnienia i uznali, że przyspieszy to czas ich pracy jak i powinno przyczynić się zmniejszenie kolejek do specjalistów,
- Mieszkańcy województwa mazowieckiego w przeważającej większości uznali za dobrą alternatywę wykorzystanie telemedycyny do stawiania diagnoz,
- Respondenci deklarują, że wykorzystanie drukarek 3D w medycynie będzie bardzo pomocne w pracy lekarza.

Literatura

1. Piętka E. *Zintegrowany System Informacyjny w pracy szpitalu* , PWN 2004 s.104-109.
2. Por. J. Chluska *Model rachunku kosztów świadczeń zdrowotnych szpitali*, Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2008. s.186.
3. Cała J., Czekierda Ł. *Internet Technologies in Medical Systems* 2006 s. 46-76.
4. Rozporządzenie Ministra Zdrowia z dnia 9 listopada 2015 r. w sprawie rodzajów, zakresu i wzorów dokumentacji medycznej oraz sposobu jej przetwarzania(Dz.U. 2015 poz. 2069).
5. Beolchi L. European Glossary of Concepts, Standards, Technologies and Users. European Commision 2001.
6. Farncombe T., Iniewski T. *Medical Imaging: Technology and Applications* 2014 s.256-267.
7. Rudowski R. *Informatyka Medyczna*, Wydawnictwo Naukowe PWN 2003.
8. *www.p1.csioz.gov.pl/11* dostęp 18.01.2016.
9. November J. *Biomedical Camputing: Digitizing Life in the United States*, Johns Hopkins University Pass, Baltimore 2012.
10. Haux R. *Medical informatics past,present, future*., International journal of medical information2010 s.596-614.
11. Richesson R.L., Andrews J.E.*Clinical research informatics*, Springer, Berlin 2012
12. Cieśla A. , Krzaszewski W., Tadeusiewicz R. *Magnetictele*therapy: part 2 – visualization of magnetotherapy, 2012.

Nowoczesne technologie medyczne w pracy z pacjentem

Streszczenie
W dzisiejszych czasach nowoczesne technologie medyczne obecne są we wszystkich sferach życia. Ich stosowanie przez lekarzy pozwala na szybszą diagnostykę oraz leczenie pacjentów. Celem pracy była ocena wykorzystania nowoczesnych technologii przez lekarzy w pracy z pacjentem. Badania prowadzono od listopada 2015 do maja 2016 roku. Grupę badaną stanowiło 158 lekarzy pracujący w wybranych placówkach medycznych na terenie całego kraju. W badaniu zastosowano metodę sondażu diagnoztycznego, natomiast narzędziem badawczym był kwestionariusz ankiety własnego autorstwa. Uzyskany materiał badawczy poddano analizie statystycznej za pomocą programu Statistica w wersji 12. Z przeprowadzonych badań wynika, że lekarze chętnie korzystają z nowoczesnych rozwiązań w pracy z pacjentem. Respondenci najczęściej wykorzystują system do opisu stanu pacjenta, system do rejestracji pacjenta oraz z dostępnej bazy leków. Ankietowani wyrazili chęć stosowania e-recepty, e-skierowania oraz e-zwolnienia (96,77%). Mieszkańcy województwa mazowieckiego (85,71%) w zdecydowanej większości uznali za dobrą alternatywę wykorzystanie telemedycyny do stawiania diagnoz. Na podstawie uzyskanych wyników można wywnioskować, iż lekarze są pozytywnie nastawieni do wdrażania nowoczesnych technologii w pracy zawodowej. W praktyce pozwoli to na szybką wymianę informacji o pacjencie, ale także skrócenie czasu oczekiwania na wizytę lekarską.
Słowa kluczowe: technologie medyczne, medycyna, pacjent, lekarz

Modern medical technologies in working with patient

Abstract
These days, modern medical technologies are present in all spheres of life. Their use by doctors allows for faster diagnosis and treatment of the patients. The aim of the study was to evaluate the use of modern technology by doctors working with the patients. The study was conducted from November 2015 to May 2016. The study group consisted of 158 doctors, working in selected medical institutions across the country. In a research was used the method of diagnostic survey, whereas the research tool was a questionnaire survey of his own authorship. The obtained research material were subjected to statistical analysis using Statistica 12. The study shows that doctors are willing to use modern solutions in work with the patients. The respondents most frequently use a system for describing patient's condition, a system for patient's registration and available databases of medicines. The respondents expressed willingness to use e-prescription, e-referrals and e-exempt (96,77%). Inhabitants of the mazowieckie voivodeship (85,71%) in the majority considered as a good alternative using a telemedicine for diagnoses. Based on the obtained results, it can be concluded, that doctors are positive about the implementation of modern technologies in their profession. In practice it will allow for quick exchange of informations about the patient, but it also will shorten the waiting time for a medical visit.
Key words: medical technologies, medicine, patient, doctor

Załącznik nr 1- Ankieta

Poniższa ankieta dotyczy badań na temat nowoczesnych technologii medycznych w pracy z pacjentem. Ankieta jest skierowana do personelu medycznego.

Celem ankiety jest zbadanie wiedzy na temat wykorzystania nowoczesnych technologii przez personel medyczny.

Zwracam się z uprzejmą prośbą do Pań i Panów o udzielenie odpowiedzi na poniższe pytania. Ankieta jest anonimowa. Jednocześnie informuje, że wyniki ankiety posłużą mi jedynie do badań. Z góry dziękuję za pomoc.

1.Czy Pan/Pani korzysta z systemów informatycznych w służbie zdrowia (Rejestracja Pacjenta, Baza leków, Cyfrowa karta pacjenta, itp) ?

a. Tak
b. Nie

2.Jak Pan/Pani określa przydatność systemów informatycznych w służbie zdrowia ?

a. Bardzo dobrze
b. Dobrze
c. Złe

d. Inne……

3.Jak Pan/Pani często korzysta z nowoczesnych technologii (Smartfony, programy komputerowe)z pracy z pacjentem w służbie zdrowia ?

a. Co godzinę
b. Do sprawdzenia raportu/karty chorób danego pacjenta.
c. Cały czas.
d. Do opisu stanu pacjenta.
e. Do wypisania recepty/zwolnienia/skierowania.
f. Inne…….. .

4.Ile czasu podczas dyżuru poświęca Pan/Pani na korzystanie z nowoczesnych technologii?

a. 0 – 30 min
b. 30 – 2h
c. powyżej 2h – 4h
d. Więcej niż 4h.

5.Z jakich systemów Pan/Pani korzysta?

a) System do rejestracji pacjenta
b) Z bazy leków
c) System do karty pacjenta
d) System do opisu przypadku choroby
e) System do poszukiwania choroby
f) Systemy TK, EKG, Rezonansu magnetycznego
g) Systemu do wysyłania karetek
h) inne…… .

6.Jak ocenia Pan/Pani prace systemu wykorzystywanego w Państwa szpitalu ?

a. Bardzo dobrze
b. Dobrze
c. Źle

7.Z jakiego urządzenie Pan/Pani korzysta z w/w systemu?

a. Komputer
b. Smartfon
c. Tablet
d. Inne jakie…… .

8.W jakim stopniu Pan/Pani sądzi, że wykorzystanie telemedycyny zwolni lekarza z niektórych obowiązków ?

a. W dużym stopniu
b. W średnim stopni
c. W małym stopniu

9.Co Pan/Pani sądzi o przeniesieniu karty pacjenta na tablet lub inne urządzenie przenośne?

a. Bardzo dobry pomysł
b. Dobry pomysł
c. Zły pomysł

10.Co Pan/Pani sądzi o większym wykorzystaniu telemedycyny w codziennej pracy lekarza ?

a. Bardzo dobry pomysł
b. Dobry pomysł
c. Zły pomysł

11.Czy Pan/Pani korzystała by z możliwości wypisywania e-recepty, e-skierowania, e-zwolnienia?

a. Tak
b. Nie
c. Nie wiem

12.Jak Pan/Pani sądzi czy wykorzystanie e-recepty, e-skierowania, e-zwolnienia zmniejszyło by kolejki do lekarzy specjalistów?

a. Tak, w znacznym stopniu
b. Nie, pozostało by to bez zmian.
c. Nie, wydłużyło by jeszcze kolejki
d. Nie wiem

13.Jakie urządzania Pan/Pani używa w czasie pracy z pacjentem ?
a. Tablet, Komputer, Smartfon.
b. Tomograf Komputerowy
c. USG
d. Rezonans magnetyczny
e. Inne

14.Czy Pan/Pani miała kiedykolwiek problemy z wykorzystaniem nowych technologiami w pracy ?
a. Tak
b. Nie

Jeżeli ”Tak” proszę opisać z jakim problemem Pan/Pani natrafił. .....

15.W jakim stopniu widzi Pan/Pani rozwój nowoczesnych technologii medycznych pracy z pacjentem w naszym kraju ?
5(Bardzo Dobrze) 4 3 2 1(Źle)

16.Co Pan/Pani sądzi o drukowaniu organów np. Serca za pomocą drukarki 3D (W pełnym wymiarze)?
a. Bardzo dobry pomysł
b. Dobry pomysł
c. Zły pomysł

17.Jak Pan/Pani sądzi czy za pomocą wykorzystania Smarfonowej kamerki i przepływu danych będzie możliwość zdiagnozowania niektórych chorób ?
a. Tak
b. Nie
c. Nie wiem

18.Jak Pan/Pani sądzi w jakich specjalizacjach najczęściej w przyszłości będziemy korzystać z nowoczesnych technologii.
a. Kardiologii
b. Elektrokardiologii
c. Neurologii
d. Chirurgii
e. Ratownictwie medycznym
f. Psychiatrii
g. Inne…….

Metryczka:

1.Płeć
a. Kobieta
b. Mężczyzna

2.Miejsce zamieszkania
a. Wieś
b. Miasto

3.Wiek…….

4.Wykształcenie (przy specjalistycznym proszę podać specjalizacje która Pan/Pani posiada)
a. Średnie
b. Zawodowe
c. Wyższe
d. Specjalistyczne……

Michał Lipiński[1]

Topologiczna analiza danych w neuroobrazowaniu funkcjonalnym

1. Wprowadzenie

Topologiczna analiza danych (TDA, ang. topological data analysis) jest relatywnie młodą dziedziną wywodzącą się z topologii algebraicznej. Głównym założeniem stojącym za TDA jest idea, że kształt danych również niesie w sobie informację. Przy czym kształt będziemy rozumieć jest w kategoriach niezmienników topologicznych, jakimi są np. grupy homologii. Gdzie natomiast w danych kształt może być zaszyty? W najbardziej oczywistym przypadku obserwowane obiekty mogą przejawiać nietrywialne kształty, przez co zbadanie ich cech topologicznych może umożliwić ich klasyfikację. W bardziej abstrakcyjnym przykładzie można utożsamić empiryczne gromadzenie informacji z próbkowaniem pewnej abstrakcyjnej przestrzeni możliwych stanów, im więcej obserwacji, tym lepsze jej przybliżenie. Różne warunki eksperymentalne oraz różne subpopulacje, prowadzą do innego rozłożenia prawdopodobieństw w przestrzeni stanów. Celem byłoby zatem wykorzystanie koncepcji topologicznych, aby uchwycić subtelności związane z różnicami między tymi przestrzeniami.

Obietnica identyfikacji nowego typu wzorców, spowodowała zwiększone zainteresowanie TDA, co można zaobserwować poprzez liczne próby jej zastosowania w wielu dziedzinach, m. in. w rozpoznawaniu obrazów [1], genetyce [2], analizie biznesowej [3], czy neuroobrazowaniu, zarówno strukturalnym [4-6], jak i funkcjonalnym [7-17].

W ostatniej z wymienionych tematyk, stanowiącej przedmiot tej pracy, analiza polega często na analizie korelacji wzorców aktywności przejawianych przez różne obszary mózgu. Najczęściej prowadzi to do konstrukcji macierzy korelacji lub jej równoważnej reprezentacji – grafu. Powszechnie, jego charakteryzacja następuje w oparciu o miary wywodzące się wprost z teorii grafów. Ich popularność wśród badaczy, przejawia się m.in. poprzez pojawiające się propozycje standaryzacji tych miar [18]. Jednak wiąże się z nimi parę problemów [19], takich jak np. zależność rezultatów od liczby wierzchołków/krawędzi w grafie, czy pojawiający się problem arbitralnego progowania danych. Narzędzia dostarczane przez topologię algebraiczną wydają się być skuteczne właśnie w sytuacjach, gdzie podejście grafowe może zawodzić. Stąd interesujące wydaje się porównanie obu metod oraz rozważenie komplementarnego ich użycia podczas analizy danych o charakterze grafowym. Narzędzia TDA mają również mocne ugruntowanie teoretycznie, dostarczające np. twierdzenie o stabilności [20], mówiące, że rozsądne zaburzenie danych nie spowoduje drastycznej zmiany w uzyskiwanych wynikach. Niemal zawsze pojawiający się problem szumu

[1] michal.lipinski@ii.uj.edu.pl, Katedra Matematyki Obliczeniowej, Instytut Informatyki i Matematyki Komputerowej, Wydział Matematyki i Informatyki, Uniwersytet Jagielloński

w danych pochodzących z neuroobrazowania stanowi kolejny argument przemawiający za użyciem TDA.

Poza głównym trendem badań skupiającym się na analizie grafów korelacji powstały również inne idee wydobycia informacji z danych funkcjonalnych opartych na TDA. Przykładem jest neuronalny model reprezentacji otoczenia lub analiza sygnału EEG wykorzystującego twierdzenie Takensa pochodzące z teorii układów dynamicznych.

2. Topologia

2.1. Przestrzenie topologiczne

Przed przejściem do samej charakterystyki kształtów badanych obiektów warto przywołać definicje pojawiających się dalej obiektów matematycznych.

Najbardziej podstawowym konstruktem jest sama przestrzeń topologiczna, którą można intuicyjnie rozumieć, jako zbiór punktów (oczywiście może być nieskończony, np. w $\mathbb{R}^n$) wraz ze zdefiniowanymi relacjami sąsiedztwa. Bardziej formalnie, *przestrzeń topologiczna* to para $\boldsymbol{X} = (X, \tau)$, gdzie X jest przestrzenią, natomiast τ – rodziną zbiorów otwartych spełniających następujące aksjomaty:

$$a, b \in \tau \Rightarrow a \cap b \in \tau \qquad (1)$$

$$\{a_i\}_{i \in I} \Rightarrow \bigcup_{i \in I} a_i \in \tau \qquad (2)$$

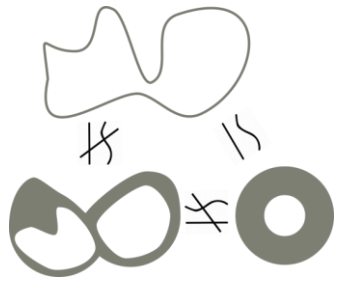
$$\emptyset, X \in \tau \qquad (3)$$

gdzie I jest dowolnym zbiorem indeksów.

Niech $\boldsymbol{X} = (X, \tau)$ i $\boldsymbol{Y} = (Y, \psi)$ będą przestrzeniami topologicznymi, wtedy funkcję $f: \boldsymbol{X} \rightarrow \boldsymbol{Y}$ nazywamy *ciągłą* jeśli, przeciwobraz każdego ze zbiorów otwartych $b \in \psi$ w $\boldsymbol{Y}$ jest zbiorem otwartym w $\boldsymbol{X}$, tzn. $f^{-1}(b) \in \tau$. Jeśli funkcja f jest dodatkowo bijekcją oraz jej odwrotność f^{-1} również ciągła, to f jest *homeomorfizmem*. Jeśli między $\boldsymbol{X}$ i $\boldsymbol{Y}$ istnieje homeomorfizm, to mówi się, że są topologicznie tego samego typu.

Słabszym od homeomorfizmu założeniem jest *homotopijna równoważność*, którą wyraża się następująco: niech $f: \boldsymbol{X} \rightarrow \boldsymbol{Y}$ będzie ciągła, jeśli istnieje $g: \boldsymbol{Y} \rightarrow \boldsymbol{X}$ ciągła, i taka, że $f \circ g \simeq \mathbf{1}_{\boldsymbol{Y}}$ oraz $g \circ f \simeq \mathbf{1}_{\boldsymbol{X}}$ to $\boldsymbol{X}$ i $\boldsymbol{Y}$ mają ten sam typ homotopii. Przy czym $\mathbf{1}_{\boldsymbol{X}}$ i $\mathbf{1}_{\boldsymbol{Y}}$ oznaczają funkcje indentycznościowe w odpowiednich przestrzeniach, tzn. $\mathbf{1}_{\boldsymbol{X}}(x) = x$, a $g \circ f \simeq \mathbf{1}_{\boldsymbol{X}}$ istnienie homotopii, tzn. takiej funkcji ciągłej $\boldsymbol{F}: \boldsymbol{X} \times \boldsymbol{X} \rightarrow \boldsymbol{Y}$, takiej, że $\boldsymbol{F}_0(x) = \boldsymbol{F}(x, 0) = (g \circ f)(x)$ oraz $\boldsymbol{F}_1(x) = \boldsymbol{F}(x, 1) = \mathbf{1}_{\boldsymbol{X}}$.

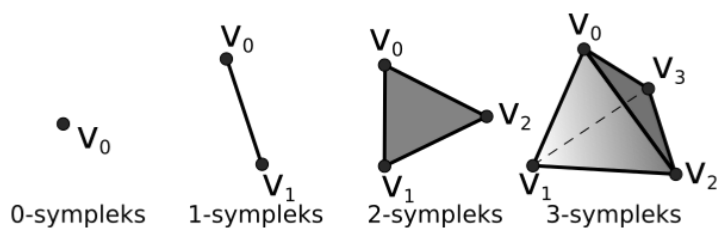
Powyższe definicje równoważności mówią intuicyjnie o możliwości przekształcenia (rozciągania, skręcania, ściskania) jednej przestrzeni w drugą, w sposób, który nie powoduje „rozerwania" lub „sklejenia" jej fragmentów. Homotopijność pozwala dodatkowo na ściskanie do zerowej szerokości, np. pierścienia do okręgu lub pętli (rys.1).



Rysunek 20 Homotopijność, zniekształcona ósemka nie jest homotopijna do pozostałych dwóch figur, ponieważ jej transformacja powodowałaby jej rozerwanie. Pierścień i pętlę da się przekształcić wzajemnie w siebie w sposób ciągły. Liczby Bettiego pierścienia: β_0=1, β_1=1, liczby Bettiego zniekształconej ósemki: β_0=1, β_1=2. [opracowanie własne]

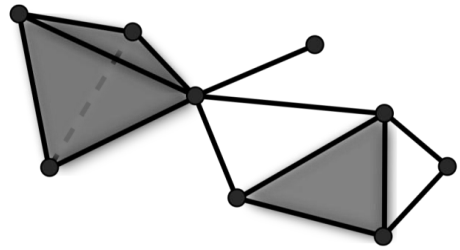
2.2. Kompleksy symplicjalne

Problem wyznaczania, choćby teoretycznie prostszego, typu homotopii jest zadaniem trudnym. Już charakteryzacja homotopii związanych z okręgiem wymaga technicznego dowodu [21]. Co więcej, niemal każdą parę przestrzeni należałoby wtedy rozważać indywidualnie. Jednak możliwe jest przyjęcie pewnych uproszczeń, powodujące, że powyższe zadanie staje się obliczalne w sposób zalgorytmizowany. Pierwsze z nich to przekształcenie danych do postaci kombinatorycznej. Ze względu na swoją prostotę, konstruuje się tzw. *kompleks symplicjalny*.



Rysunek 21 Sympleksy. [opracowanie własne]

Kompleks symplicjalny jest zbiorem sympleksów, połączonych zgodnie z pewnymi regułami. *Abstrakcyjnym n-sympleksem* jest zbiór n+1 punktów $[v_0 v_1 ... v_n]$. Jeśli punkty zanurzone są w przestrzeni euklidesowej, to *n-sympleksem* określa się otoczkę wypukłą tych punktów, co geometrycznie oznacza, że 0-sympleks to punkt, 1-sympleks – odcinek, 2-sympleks – trójkąt, 3-sympleks – wypełniony czworościan, itd. (rys. 2). Niech σ będzie sympleksem, wtedy sympleks μ rozpięty na dowolnym podzbiorze punktów określających σ, oznacza się mianem *ściany sympleksu σ*, co oznacza się przez zapis $\mu < \sigma$ (np. ścianą 2-sympleksu $[v_0 v_1 v_2]$ – trójkąta jest jego bok $[v_0 v_1]$).



Rysunek 22 Kompleks symplicjalny. [opracowanie własne]

Niech *K* będzie zbiorem sympleksów, *K* jest *kompleksem symplicjalnym* jeżeli spełnione są następujące warunki:

$$\sigma \in K \Rightarrow (\mu < \sigma \ \Rightarrow \mu \in K) \qquad (4)$$

$$(\sigma_1, \sigma_2 \in K, \mu = \sigma_1 \cap \sigma_2) \Rightarrow (\mu = \emptyset \vee (\mu < \sigma_1 \wedge \mu < \sigma_2)), \qquad (5)$$

gdzie pierwszy warunek mówi, że ściana sympleksu znajdującego się w *K* również należy do *K*, a drugi, że częścią wspólną dwóch sympleksów w *K* jest cała ściana – sympleks niższego wymiaru.

N-szkieletem kompleksu symplicjalnego jest podkompleks składający się jedynie z sympleksów o wymiarze nie większym niż *n* i oznacza się go przez K_n. *Wymiarem kompleksu* jest wymiar jego najwyżej wymiarowego sympleksu.

Kompleks symplicjalny (rys. 3) jest w szczególności przestrzenią topologiczną, w której zbiór wnętrz sympleksów stanowi rodzinę zbiorów otwartych.

2.3. Homologie

Drugim problemem, poza reprezentacją obiektów topologicznych, jest pytanie o ich homeomorficzność oraz sposób opisu różnych typów topologicznych. Aby uprościć osiągnięcie podobnego rezultatu, można ograniczyć się do analizy tzw. *niezmienników topologicznych*, których zachowanie można bardziej formalnie opisać następująco: niech $\boldsymbol{X}$ i $\boldsymbol{Y}$ będą przestrzeniami topologicznymi, a *L* – funkcją przypisującą przestrzeni topologicznej obiekt stanowiący opis niezmiennika topologicznego, wtedy:

$$\boldsymbol{X} \approx \boldsymbol{Y} \ \Rightarrow L(\boldsymbol{X}) = L(\boldsymbol{Y}) \qquad (6)$$

Zatem zgodność niezmienników topologicznych nie odpowiada na pytanie o homeomorfizm, ale daje gwarancję, że dwie różne przestrzenie topologiczne uda się rozróżnić (inkluzja tylko w jedną stronę). Powszechnie wykorzystywanym niezmiennikiem topologicznym są grupy homologii, które intuicyjnie opisują liczbę „dziur" lub „pętli" n-wymiarowych w przestrzeni. Aby zdefiniować je bardziej precyzyjnie potrzebujemy paru kolejnych definicji.

Niech *K* będzie kompleksem symplicjalnym. Zbiór:

$$C_n(K, \mathbb{Z}_2) := \{f : K_n \rightarrow \mathbb{Z}_2\}, \qquad (7)$$

oznacza *grupę n-łańcuchów* kompleksu *K* o współczynnikach w $\mathbb{Z}_2$ (w ogólności współczynniki mogą pochodzić z dowolnego pierścienia). Innymi słowy, każdemu sympleksowi przypisuje się pewną wartość (dla $\mathbb{Z}_2$ będzie to 0 lub 1) i wprowadza się operację dodawania sympleksów (w tym przypadku dwukrotne dodanie tego samego sympleksu zeruje stojący przy nim współczynnik). *Operator brzegu* ∂, zdefiniowany następująco:

$$\partial_n : C_n(K;\ \mathbb{Z}_2) \rightarrow\ C_{n-1}(K;\ \mathbb{Z}_2), \tag{8}$$

$$\partial_n(\sigma) = \sum_{i\in\{0,\dots,n\}}[v_0 \dots v_{i-1}v_{i+1} \dots v_n], \tag{9}$$

gdzie σ jest n-łańcuchem (w szczególności np. sympleks). Przypisuje on każdemu sympleksowi sumę jego ścian o wymiarze o 1 mniejszym.

Przykładowo, wartość operatora brzegu dla 2-sympleksu:

$$\partial_2([v_0v_1v_2]) = [v_0v_1] + [v_1v_2] + [v_0v_2] \tag{10}$$

oraz dla dwóch przystających krawędzi:

$$\partial_1([v_0v_1] + [v_1v_2]) = [v_0] + [v_1] + [v_1] + [v_2] = [v_0] + [v_2], \tag{11}$$

gdzie redukcja sympleksu $[v_1]$ wynika stąd, że 1+1=0 w $\mathbb{Z}_2$.

Grupą n-cykli określa się:

$$Z_n \coloneqq ker\ \partial_n \tag{12}$$

natomiast grupę n-brzegów:

$$B_n \coloneqq \mathrm{im}\ \partial_{n+1} \tag{13}$$

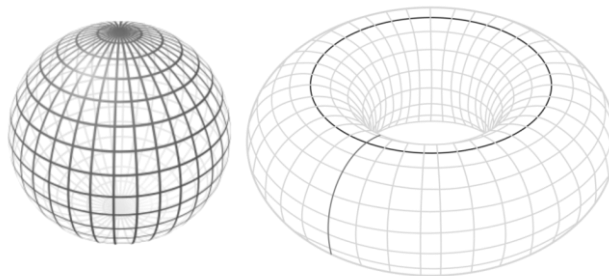
gdzie *ker* oznacza jądro odwzorowania, a *im* jego obraz.

Kompleksy symplicjalne posiadają pożądaną cechę, mianowicie wyrażenie $\partial_n\partial_{n+1}=0$ jest prawdziwe dla dowolnego podkompleksu, co prowadzi do natychmiastowego wniosku, że $B_n \subset Z_n$ i pozwala na poprawne zdefiniowanie grupy ilorazowej:

$$H_n \coloneqq {}^{Z_n}\!/_{B_n} \tag{14}$$

którą określa się mianem *n-tej grupa homologii*. Wymiar grupy homologii, czyli liczbę generatorów n-tej grupy nazywa się *n-tą liczbą Bettiego* lub symbolem β_n.

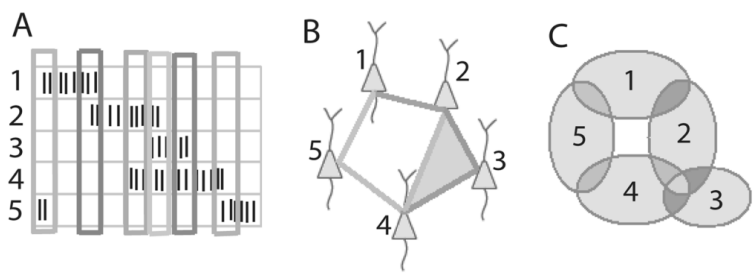
Zaskakującym rezultatem tej, wydawałoby się złożonej definicji, są obiekty, o których bardzo łatwo myśleć intuicyjnie jako o „dziurach n-wymiarowych". Generatory 0-homologii odpowiadają liczbie spójnych składowych w przestrzeni, 1-homologie – pętlom i tunelom, 2-homologie – zamkniętym przestrzeniom, ograniczonym powierzchnią 2D (jak np. przestrzeń wewnątrz sfery). Pomysł ten uogólnia się na dowolnie wysoki wymiar, z tym, że wyobraźnia geometryczna przestaje być tam pomocna, jednakże intuicje z niższych wymiarów uchwytują całą ideę. Przykładowe wartości liczb Bettiego wyszczególnione zostały dla obiektów znajdujących się na rysunkach 1 i 4.



Rysunek 23 Sfera i torus. Liczby Bettiego sfery: $\beta_0=1$, $\beta_1=0$, $\beta_2=1$. Liczby Bettiego torusa: $\beta_0=1$, $\beta_1=2$, $\beta_2=1$. [źródło: wikipedia.org]

3. Neuronalny model otoczenia

Pierwszym z przytoczonych przykładów i przy okazji jedno z pierwszych zastosowań TDA w badaniu aktywności mózgu dotyczy neuronalnego modelu reprezentacji otoczenia przestrzennego przez komórki miejsca (ang. place cells) znajdujące się w hipokampie [22]. Komórki te zaliczane są do grupy neuronów odpowiadających za kodowanie allocentryczne (tzn. globalnie, w stosunku do informacji przestrzennej, np. w centrum labiryntu, przeciwieństwo – egocentryczna reprezentacja, np. „na prawo ode mnie") położenia podmiotu w przestrzeni. Mózg zapoznając się z nowym otoczeniem, tworzy neuronalną mapę, przypisując komórce miejsca charakterystyczny obszar przestrzeni w której jest ona aktywna. Znając zatem odpowiadający danemu neuronowi obszar i obserwując jego aktywność można stwierdzić czy podmiot znajduje się w tym położeniu. Problem można przeformułować i zadać pytanie, jakie informacje da się wydobyć o samej przestrzeni, którą mózg reprezentuje jedynie na podstawie zarejestrowanych potencjałów czynnościowych grupy komórek miejsca. Model zaproponowany przez Curto i Itscov [7] sugeruje, że możliwa jest rekonstrukcja topologii.



Rysunek 24 Uproszczony przykład działania modelu neuronalnej reprezentacji otoczenia. (A) przedstawia zapis potencjałów 5 komórek oraz okna czasowe w których zachodzi koaktywacja, (B) jest kompleksem zbudowanym na danych, gdzie n-1 sympleks reprezentuje zgrupowanie, (C) obrazuje rodzinę zbiorów, którego kompleks (B) jest nerwem. Źródło: [7]

Stojąca za modelem idea konstrukcji kompleksu symplicjalnego jest bardzo prosta. W pierwszym kroku utożsamia się każdą z rejestrowanych komórek z wierzchołkiem (0-sympleksem). Następnie analizując wzorce aktywności, wyszukiwane są *grupowania komórek*, czyli zbiór neuronów, który wykazuje wyraźną współaktywność w przyjętym oknie czasowym (~250ms). Zgrupowanie n+1 komórek odpowiadać będzie n-sympleksowi rozpiętemu na odpowiadających komórkom wierzchołkach w konstruowanym kompleksie (rys. 5A i B). Jeśli dwie lub więcej komórki są współaktywne, oznacza to, że kodowane przez nie obszary częściowo na siebie nachodzą (rys. 5C). Zakładając, że każdy punkt przestrzeni znajduje się w obrębie obszaru przynajmniej jednej komórki miejsca i dodając techniczne założenie, że zbiory te są otwarte oraz że ich przecięcia są ściągalne (tzn. wszystkie grupy homologii są trywialne, do czego wystarcza założenie, że obszary te są np. elipsami) otrzymujemy pokrycie przestrzeni. Skonstruowany kompleks stanowi dokładnie tzw. nerw pokrycia otwartego. Jako bezpośrednie następstwo wniosku 4G.3 z [21] wiadomo, że homologie nerwu, w szczególności również liczby Bettiego, będą identyczne z homologiami pokrywanej przestrzeni. Zatem ten nieskomplikowany tok rozumowania doprowadził do modelu rozpoznającego np. liczbę dziur w labiryncie po którym porusza się szczur laboratoryjny jedynie na podstawie rejestracji aktywności grupy neuronów.

Autorzy [7] proponują pójść krok dalej dodając założenie o równomiernym rozłożeniu i wielkości obszarów kodowanych przez komórki miejsca. Wtedy oprócz cech topologicznych, model pozwala na przybliżoną rekonstrukcję całej geometrii.

W innym badaniu Dabagian i wsp. [8] zbadali zależność metody od wyboru podstawowych parametrów, takich jak: częstotliwość potencjałów czynnościowych generowanych przez komórki miejsca, rozmiar obszarów kodowanych przez pojedynczy neuron oraz liczbę komórek użytych do reprezentacji przestrzeni. W rezultacie zaobserwowali, że poprawne dekodowanie cech topologicznych zachodzi jedynie dla pewnego zakresu wartości tych parametrów. Przedział ten maleje wraz ze wzrostem złożoności geometrycznej rozważanej przestrzeni. Co więcej, komórki miejsca nie kodują globalnie jedynego obszaru. Gdy osobnik przemieszcza się w inne otoczenie, wtedy aktywna staje się mapa neuronalna. Zatem komórka może kodować inne położenie oraz odpowiadający jej obszar może mieć inny rozmiar. Powoduje to, że podczas potencjalnego, empirycznego testowania modelu, strategia doboru poprawnej kombinacji parametrów może napotkać kolejne trudności.

Sam model przetestowano później wprost na danych empirycznych przez Dabagiana i wsp. [9] oraz poprzez analizę aktywności komórek hipokampa w trakcie eksploracji i w fazie snu REM przez Giusti i wsp. [10]. Co prawda wyniki sugerują potrzebę przeformułowania modelu i być może nowego spojrzenia na rzeczywistą rolę komórek miejsce, lecz jest to jednocześnie sukces TDA, pokazując swoją użyteczność w weryfikacji hipotez przyrodniczych.

4. Homologie persystentne

4.1. Filtracja

Motywacją do analizy wielu wartości pewnego parametru może być również np. sensowność danych jedynie dla pewnego zakresu wartości parametru, nie znanego explicite (jak w poprzednim podrozdziale), lub istnienia wielowarstwowej albo hierarchicznej struktury, której konkretna parametryzacja nie jest w stanie uchwycić. Metoda homologii persystentnych służy dokładnie przezwyciężeniu tego typu problemów.

Idea stojąca za znajdywaniem cech topologicznych, zależnych od parametru, może wydawać się naiwna, bo polega po prostu na zbadaniu homologii dla całego spektrum wartości parametru. Istnieją jednak algorytmy, gdzie wyniki dla jednej wartości wykorzystywane są przy wyliczaniu homologii dla parametru kolejnego [23]. Z tak przeanalizowanych danych wychwytuje się trwałe cechy topologiczne, tzn. te które są obecne dla długiego przedziału wartości parametru.

Niech konstrukcja kompleksu K zależy od wyboru parametru ε. Z racji, rozważanych skończonych zbiorów danych, istnieć będzie ciąg $\varepsilon_0 < \varepsilon_1 <...< \varepsilon_m$, dla pewnego $m > 0$, taki, że pomiędzy dwoma kolejnymi następuje zmiana w topologii kompleksu. Wymagamy, aby zachodziło:

$$K(\varepsilon_0) \subseteq K(\varepsilon_1) \subseteq \cdots \subseteq K(\varepsilon_m), \qquad (15)$$

gdzie $K(\varepsilon_i)$ odpowiada kompleksowi skonstruowanemu dla parametru ε_i. Oznacza to, że jeśli sympleks pojawił się w kompleksie dla ε_i, to już będzie on obecny dla dowolnego innego ε_j, takiego, że $\varepsilon_j > \varepsilon_i$. Powyższy ciąg inkluzji nazywa się filtracją, a ε *parametrem filtracji*.

W wielu kontekstach parametr filtracji stanowi wartość względem której progujemy dane na podstawie pewnej metryki lub funkcji odległości. Przykładowo, jeśli nasze obserwacje możemy utożsamić z punktami w przestrzeni $\mathbb{R}^n$ to miara euklidesowa jest taką funkcją i twierdzimy, że między dwoma punktami istnieje połączenie jeżeli odległość między nimi jest mniejsza niż ε. W ogólności, parametrem może być dowolna wielkość prowadząca do filtracji. U Dabagiana i wsp. [8] parametrem był czas który symulowany szczur spędził na eksploracji labiryntu. Wraz z odwiedzeniem nowych obszarów do kompleksu dodaje się sympleks jeśli zaobserwowana została nowa współaktywność. Sympleksów nie ubywało w czasie, więc wymagania filtracji są spełnione.

W przykładach z kolejnego podrozdziału za funkcję odległości będziemy przyjmować stopień korelacji aktywności między obszarami kory, tj. im korelacja jest silniejsza, tym dwa regiony mózgu są funkcjonalnie sobie bliższe.

4.2. Kompleks Vietorisa-Ripsa

W zależności od posiadanych danych, konieczny jest wybór strategii konstrukcji kompleksu, celem dalszego obliczenia ich charakteryzacji topologicznej. Ze względu na prostotę i obliczeniową wydajność, częstą decyzją jest budowa tzw. *kompleksu Vietorisa-Ripsa* [24].

Konstrukcja to oparta jest na strukturze grafu. Niech $V \subseteq \boldsymbol{X}$ będzie zbiorem wierzchołków w pewnej przestrzeni $\boldsymbol{X}$ (np. $\mathbb{R}^n$) oraz niech $\varepsilon > 0$ będzie ustalony. $\mathbf{G}_\varepsilon=(V, E_\varepsilon)$ będzie grafem ε-sąsiedztwa. Przy czym E_ε jest zbiorem krawędzi wyznaczonym następująco:

$$E_\varepsilon := \{\{u, v\} \mid d(u, v) \leq \varepsilon, u \neq v \in V\}, \qquad (16)$$

gdzie $d(u, v)$ oznacza funkcję odległości.

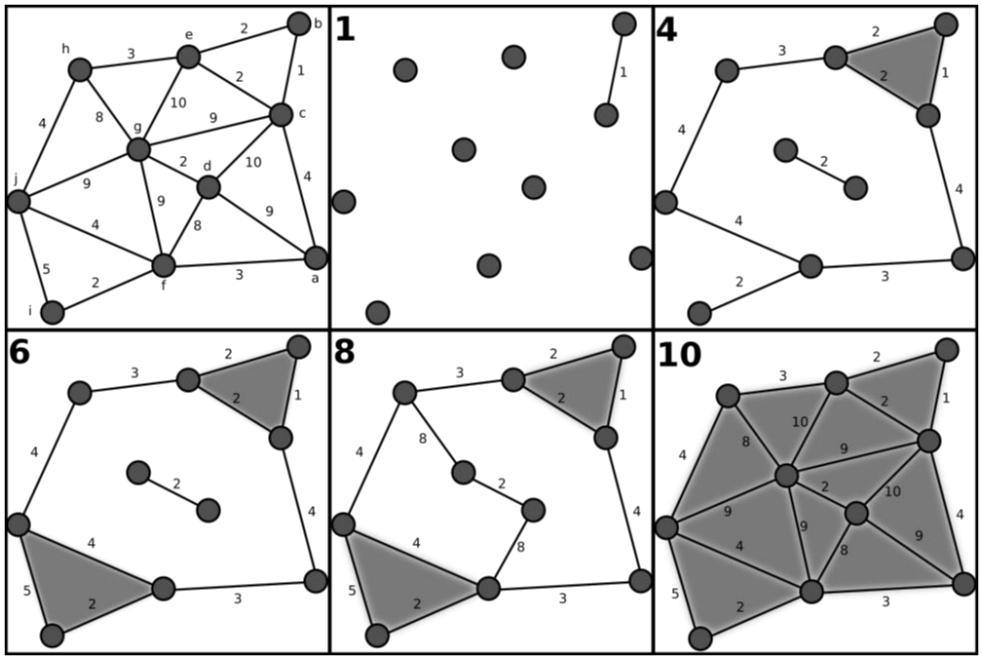
Klika oznacza pełny podgraf, tzn. taki podzbiór wierzchołków grafu $\mathbf{G}_\varepsilon$, że w E_ε zawarte są już wszystkie możliwe krawędzie między tymi wierzchołkami.

Kompleks Vietorisa-Ripsa polega na „wypełnieniu" każdej z klik, czyli jeśli w grafie istnieją wszystkie krawędzie łączące wzajemnie zbiór n+1 wierzchołków $\{v_0, v_1,...,v_n\}$, to dodajemy do kompleksu sympleks $[v_0v_1...v_n]$. Oznaczmy tak powstały kompleks przez $\mathbf{R}_\varepsilon(V)$.

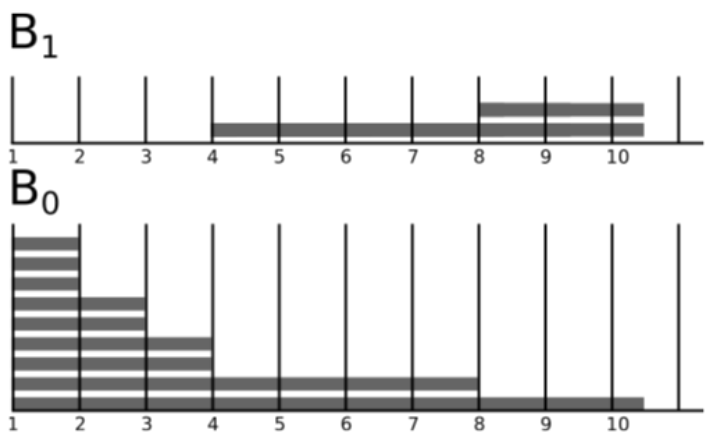
Warto zauważyć, że sam graf również jest kompleksem o wymiarze 1, w szczególności $\mathbf{G}_\varepsilon$ jest 1-szkieletem kompleksu $\mathbf{R}_\varepsilon(V)$. W pewnych sytuacjach uzasadnione jest myślenie o kompleksie jako o pewnym uogólnieniu pojęcia grafu.

4.3. Homologie persystentne

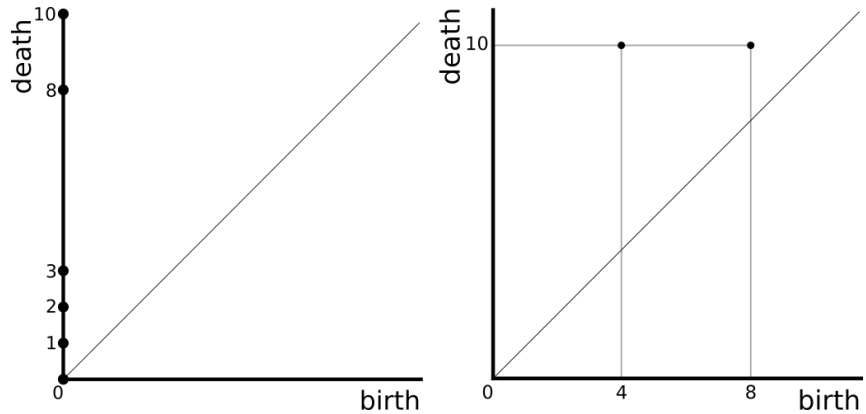
Mając zdefiniowany ciąg zawierających się kompleksów symplicjalnych, można przystąpić do badania homologii persystnentnych. Kluczowa część procedury została przedstawiona na rysunku 6. Pierwszy z nich przedstawia początkowy graf z wagami, które będą definiować funkcję odległości i będzie stanowić podstawę do filtracji. Filtracja przebiega od wartości najmniejszych. Przyjmujemy, że siła relacji między dowolnymi dwoma węzłami między którymi nie zaznaczono krawędzi waga wynosi ∞ (czyli są zupełnie niezależne). W ogólności kierunek filtracji nie ma znaczenia, dopóki spełnione są założenia (zawsze można skonstruować funkcję odwracającą wagi). W kolejnych krokach filtracji dodawane są krawędzie – 1-sympleksy o nie większych niż aktualna wartość parametru filtracji ε, a w momentach powstania wzajemnie połączonych trójek węzłów, czyli klik, również trójkąty – 2-sympleksy. Tym sposobem dostajemy ciąg kompleksów Vietorisa-Ripsa.



Rysunek 25 Pierwszy obraz przedstawia oryginalny graf na bazie którego konstruowane są kolejne etapy filtracji, wagi na krawędziach stanowią jednocześnie odległość między wierzchołkami, litery stanowią oznaczenie dla węzłów. Brak krawędzi między dwoma wierzchołkami oznacza odległość równą ∞. Na pozostałych ilustracjach znajdują się kompleksy Vietorisa-Ripsa odpowiednio dla wartości parametru filtracji $\varepsilon \in \{1, 4, 6, 8, 10\}$. [opracowanie własne]



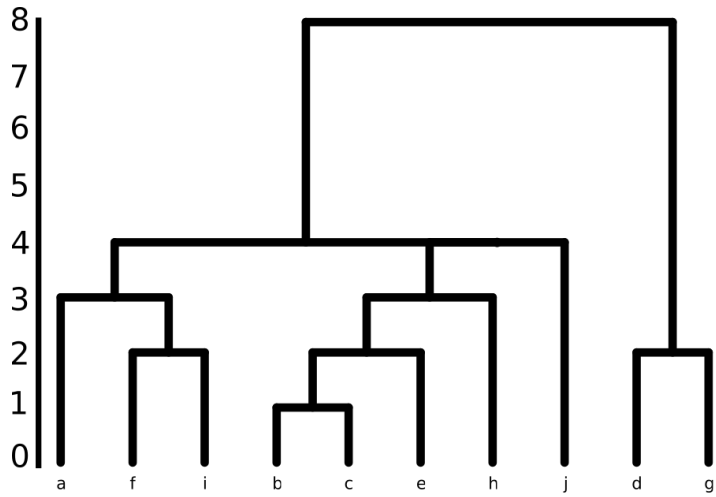
Rysunek 26 Barkody dla generatorów 1-homologii i 0-homologii dla filtracji przedstawionej na rysunku 6. Każdy pasek odpowiada istnieniu cechy topologicznej odpowiednio 1- lub 0- wymiaru. Pasek ma swój początek i koniec, odpowiednio dla wartości parametru, gdzie cecha powstała i zanikła. [opracowanie własne]



Rysunek 27 Diagramy persystencji 0-wych i 1-szych homologii – reprezentacja równoważna do barkodów. Każdy punkt na wykresach odpowiada paskowi z rys. 7. Współrzędna x odpowiada momentowi narodzin cechy topologicznej, współrzędna y – momentowi jej śmierci. [opracowanie własne]

Zmiany zachodzące w topologii przestrzeni/grafu najczęściej przedstawia się za pomocą tzw. *barkodów*. Barkody z rysunku 7 odpowiadają filtracji grafu z przykładu (rys. 6). Każdy z pasków reprezentuje pewien generator grupy homologii. Pasek ma swój początek przy wartości parametru dla której odpowiadająca cecha topologiczna pojawiła się, i kończy przy parametrze dla której zanikła. Ponieważ w przykładzie przyjęliśmy, że wszystkie wierzchołki obecne są od początku filtracji, więc wszystkie barkody B_0 (opisujące spójne składowe) rozpoczynają się już dla $\varepsilon = 0$ i po kolei zanikają wraz z uspójnianiem się grafu, aż do momentu pozostania jednej składowej, tzn. gdy w grafie znajdzie się ścieżka między dowolną parą węzłów. Na barkodach B_1 (rys. 7), odpowiadających „dziurom" 1-wymiarowym, można np. zaobserwować, zgodnie z filtracją (rys 6), powstanie pętli dla parametru $\varepsilon = 4$. Generator ten zanika, wspólnie z drugą pętlą powstałą w międzyczasie, przy ostatnim kroku filtracji, gdy wszystkie trójkąty zostają wypełnione.

Diagramy persystencji (rys. 8) stanowią reprezentację równoważną do barkodów i bywają bardziej czytelną wizualizacją w przypadku analizy dużej liczby danych. Każdy pasek – barkod – odpowiada dokładnie jednemu punktowi na diagramie, gdzie współrzędne *x* oznaczają pojawienie się generatora, a współrzędne *y*, jego zniknięciu. Ze względu na to, że zanik zawsze następuje później niż narodziny wszystkie punkty na diagramie znajdują się powyżej diagonali. Zauważmy również, że im barkod jest dłuższy tym znajduje się dalej do diagonali. Przedstawione na rys. 8 diagramy odpowiadają dokładnie barkodom z rys. 7.



Rysunek 28 Dendrogram, wizualizacja scalania się spójnych składowych dla filtracji z przykładu na rys. 6. Litery odpowiadają wierzchołkom grafu. Scalenie między dwoma węzłami następuje, gdy w grafie pojawia się ścieżka między nimi. [opracowanie własne]

Podczas analizy barkodów i diagramów persystencji interesują nas przede wszystkim te generatory, które są trwałe (persystentne). Zatem im dłuższy barkod lub odpowiadający mu punkt znajduje się dalej od diagonali, tym z większym prawdopodobieństwem zaobserwowana własność topologiczna uchwytuje istotną cechę abstrakcyjnej przestrzeni, z której pochodzą dane. Z kolei krótkie paski lub przydiagonalne punkty z dużą szansą stanowią efekt szumu oraz dyskretnego próbkowania i reprezentacji danych. Z rozważanego przykładu można wyczytać trwałość dwóch generatorów 0-wymiarowych oraz jednego 1-wymiarowego. Wnioskować można zatem o istnieniu dwóch relatywnie niezależnych podsieci oraz tworzenie przez jedną z nich struktury pierścienia.

Kolejnym sposobem wizualizacji, w tym przypadku zmian w homologiach 0-wymiarowych jest tzw. *dendrogram* (rys. 9). Obrazuje on jak, i w jakiej kolejności następuje scalanie się grafu. Pozwala także na szybką identyfikację grup wierzchołków odseparowanych od pozostałych. W podanym przykładzie natychmiastowo widać separację węzłów *d* i *g* od pozostałej części grafu. Dystans między dwoma wierzchołkami w tym przypadku zostaje wyznaczony przez tzw. *odległość pierwszego złączenia* (ang. single-linkage distance) [11]:

$$d(x_i, x_j) := min\{max_{l=0,\dots,k-1}\, c(w_l, w_{l-1}) \mid x_i = w_0, \dots, w_k = x_j\} \qquad (17)$$

gdzie c jest funkcją zwracającą wagę krawędzi między (w_l, w_{l-1}), a ciąg $w_l,\dots, w_k$ oznacza ścieżkę w grafie między wierzchołkami x_i i x_j.

5. Analiza sieci funkcjonalnych

5.1. Badania

Na przykładzie wybranych badań zostaną w tym podrozdziale przedstawione standardowe strategie aplikacji TDA w neuroobrazowaniu funkcjonalnym oraz możliwe interpretacje uzyskiwanych tą drogą rezultatów.

Lee i wsp. [11] wykorzystali TDA w celu zbadania danych pochodzących z obrazowania za pomocą FDG-PET. W badaniu uczestniczyły trzy grupy dzieci: u których zdiagnozowano zespół nadpobudliwości z deficytem uwagi (ADHD), autyzm (ASD) oraz grupę kontrolną (łącznie 61 osób). Wyodrębniono 103 obszary mózgu, które w dalszym etapie były reprezentowane przez wierzchołki w konstruowanym grafie. Między każdą parą węzłów tworzono krawędź o wadze, wyrażającej się wzorem 1-corr (x_i, x_j), gdzie corr oznacza korelację między aktywnością w obrębie dwóch regionów mózgu. Korelacja została odjęta od jedności, aby wagi przyjmowały dodatnie wartości z przedziału [0, 2].

W badaniu skupiono się wyłącznie na 0-wych liczbach Bettiego, czyli persystencji spójnych składowych. Dodatkowo, informacje przedstawione na barkodach rozszerzają poprzez konstrukcję dendrogramów, które oprócz liczby generatorów w danym etapie filtracji kodują również to, które spójne składowe i w jakiej kolejności się scalają. Do porównania dendrogramów służy im metryka Gromova-Hausforffa.

Tempo z jakim spójne składowe scalały się istotnie różniły się między badanymi grupami. Grupa ASD wyróżniała się szybkim spadkiem liczby generatorów dla małych wartości filtracji oraz potrzebna była większa wartość ε do utworzenia grafu jednospójnego, innymi słowy, w tej grupie graf funkcjonalny charakteryzuje się intensywnie współaktywnymi, lokalnymi podsieciami przy jednoczesnej słabszej synchronizacji globalnej. Grupa ADHD wykazywała opóźnioną redukcję generatorów, czyli słabszą korelację między obszarami współaktywnych w grupie kontrolnej. Dodatkowa analiza wzajemnej odległości pierwszego złączenia wykazała w grupie ASD słabszą siłę połączeń całego mózgu z obszarem Broki, natomiast wśród badanych z ADHD słabszą korelacje regionów sensomotorycznych z m.in. przednią częścią obręczy.

Oprócz jakościowej analizy, badacze porównali skuteczność klastrowania danych na podstawie różnic w strukturze dendrogramów a miarami grafowymi, będących już standardowym narzędziem w badaniu sieci funkcjonalnych [18]. Miary grafowe, takie jak charakterystyczna długość ścieżki, współczynnik małego świata, czy globalna homogeniczność sieci, wykazały się skutecznością na poziomie ponad 90%, podczas, gdy wspomniana już miara Gromova-Hausdorffa zapewniła 100% skuteczność.

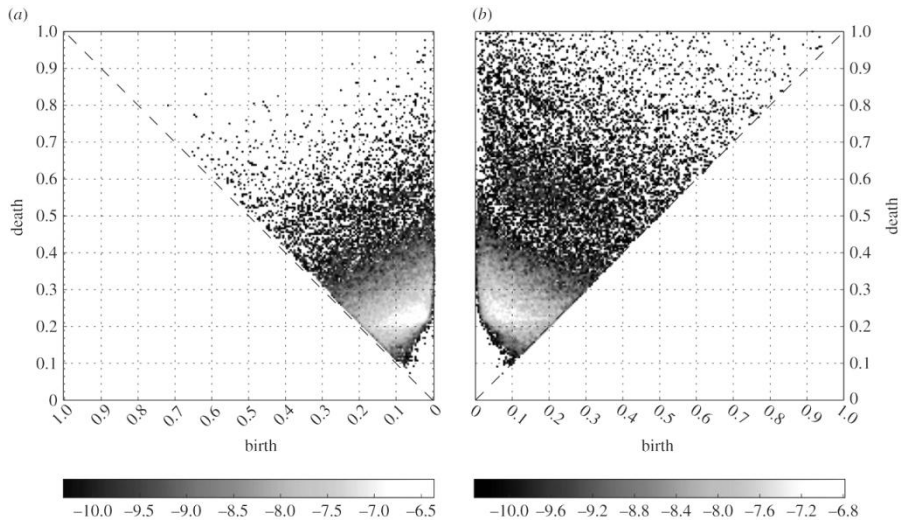
Khalid i wsp. [12] przeprowadzili analizę danych podobną do [11]. Ich badanie dotyczyło jednak różnic w sieci funkcjonalnych konstruowanych na podstawie zarejestrowanego sygnału EEG. Badaniu poddano dwie grupy szczurów: z indukowaną depresją (model CORT) oraz grupę kontrolną. Graf funkcjonalny był drastycznie mniejszy niż w poprzednim przytoczonym przykładzie, jako, że sygnał zbierany był jedynie z 8 elektrod. Wierzchołek grafu odpowiadał elektrodzie. Ponownie ograni-

czono się do badania tylko generatorów zerowej homologii, tj. spójnych składowych. Osobno rozpatrywano pięć pasm fal mózgowych: alfa, beta, gamma, delta i theta.

Po przeprowadzeniu obliczeń okazało się, że zanikanie spójnych składowych zaczyna się później (dla wyższych ε) oraz jego tempo było niższe w grupie CORT, niż grupie kontrolnej, dla każdego z pasm. W języku aktywności mózgowej, możliwa jest interpretacja mówiąca o obniżeniu globalnej synchronizacji kory i/lub zmianę trybu pracy mózgu na lokalne przetwarzanie informacji. Odseparowane węzły, stanowiące długo żyjące, samodzielne spójne składowe, mogą być tłumaczone na występowanie obszarów szczególnie odizolowanych z perspektywy funkcjonowania mózgu. Obserwacja odległości pierwszego złączenia sugeruje zmiany w grupie CORT przejawiające się np. poprzez obniżenie korelacji lewej kory czołowej z obszarami ciemieniowymi, wzmożonej aktywności, bilateralnie, między obszarami czołowymi i obniżenie, również bilateralnie, między obszarami potylicznymi. Dysfunkcje czołowo-ciemieniowej sieci funkcjonalnej w stanach depresji, uważa się za jedną z przyczyn zaburzeń w znajdujących się w jej obrębie sieci uwagowych i decyzyjnych [12]. Należy jednak pamiętać o ograniczonej możliwości uogólniania tych wyników ze względu na różnice w anatomii.

Przedmiotem badań Petri i wsp. [13] były sieci funkcjonalne mózgu utworzone na bazie skanów fMRI. Każdy z 14 badanych uczestniczył w dwóch sesjach, w jednej z nich otrzymywał dożylną iniekcję psylocybiny (substancji psychodelicznej), a w drugiej placebo. Aby porównać oba warunki metodami homologii persystentnych skonstruowano najpierw graf korelacji złożony z 169 wierzchołków reprezentujących wyodrębnione obszary kory mózgu. Podobnie do poprzednich przykładów dla każdej wartości parametru filtracji ε pozostawiano wyłącznie te krawędzie grafu, które posiadały wagę nie mniejszą niż ε. Każda klika składająca się z n węzłów w grafie G_ε traktowana jest jako sympleks n-1 wymiarowy, budowano zatem kompleks Vietorisa-Ripsa. Wyniki analizy homologii persystentnych dla tak generowanych kompleksów symplicjalnych porównywano zbiorczo dla obu warunków badawczych.

Badacze skupili się na analizie 1-homologii, czyli istnienia pętli w sieci funkcjonalnej. Na dwóch uzyskanych diagramach persystencji (rys. 10) przedstawiono sumarycznie pierwsze grupy homologii dla odpowiednio warunku placebo oraz psylocybiny. Okazuje się, że liczba cykli w obu przypadkach jest porównywalna, lecz ich dystrybucja jest istotnie różna. W przypadku placebo rozkład długości życia generatora H_1 jest bardziej jednostajny. Natomiast w warunku psylocybiny średnia długość życia jest niższa (większość cykli zanika, niedługo po pojawieniu się), ale jednocześnie powstają cykle wyjątkowo trwałe o długości życia, jak niemal nie występuje w warunku kontrolnym.



Rysunek 29 Sumaryczne diagramy persystencji dla wszystkich badanych osób dla (a) warunku kontrolnego i (b) warunku pod wpływem psylocybiny. Natężenie koloru koduje gęstość rozkładu generatorów 1-homologii zgodnie z legendą pod diagramami. Źródło: [13]

Kolejnym etapem była konstrukcja grafów pochodnych, tj. *szkielet homologicznej persystencji* (ang. persistence homological scaffold) oraz *szkielet homologicznej częstotliwości* (ang. frequency homological scaffold). Pierwszy powstał poprzez nadanie każdej krawędzi wagi równej sumie wszystkich długości życia generatorów cykli do których krawędź ta należała. Wagą krawędzi w drugim grafie jest liczba różnych cykli do których krawędź należy. Obie miary, pozwalają stwierdzić o funkcjonalnej istotności krawędzi w sieci i tym samym zidentyfikować krawędziowy odpowiednik huba [18]. Analiza obu statystyk wskazuje, że pod wpływem substancji psychodelicznej odpowiadające krawędzie należały sumarycznie do podobnej liczby cyklów, przy jednoczesnej większej trwałości tych cykli.

Nasuwa się parę wniosków płynących z tej analizy. Analiza szkieletu homologicznej persystencji oraz homologicznej częstotliwości pokazały zwiększoną globalną integrację obszarów kory mózgowych w warunku eksperymentalnym. Inna dystrybucja generatorów na diagramach persystencji (rys. 10) pokazuje możliwość zaobserwowania różnic również dla wyższych grup homologii. Jednakże nie istnieje dobra psychologiczna lub neuronalna interpretacja, która wyjaśniałaby znaczenie generatorów 1-homologii lub wyższych. Autorzy proponują hipotezę, mówiącą, że być może struktura sieci funkcjonalnej pod wpływem substancji psychodelicznej nie koniecznie staje się bardziej przypadkowa jak sugerują miary grafowe, ale następują w niej zmiany w sposób bardziej subtelny, w których wykryciu metody topologiczne mogą okazać się pomocne.

Yoo i wsp. w pracy [14] zastosowali rozwinięcie metody homologii persystentnych pozwalające na rozszerzenie analizy grafu korelacji również o zmienność następującą

w czasie. Metoda ta nosi nazwę persistence vineyard [25]. Autorzy zbadali jedynie 0-homologie, choć metoda da się uogólniać na wyższe wymiary. Analizowany w pracy sygnał EEG pochodził od grupy 26 zdrowych osób, przy czym skupiono się na uśrednionych danych. Celem badawczym było porównanie sieci funkcjonalnych mózgu podczas gry w grę video oraz stanu odpoczynku. Cała procedura składała się z 3 sekwencji gra + odpoczynek (5 min. + 5 min.) oraz dodatkowych 5 minut odpoczynku przed pierwszym uruchomieniem gry. Przetwarzany sygnał powstał z konkatenacji wszystkich segmentów.

Graf do którego zaaplikowano metodę homologii persystentnych, został zbudowany na jedynie 8 węzłach odpowiadających wybranym elektrodom EEG, natomiast wagi krawędzi określone zostały na podstawie korelacji sygnału w każdym z nich. Każde z pasm sygnału, (alfa, beta, delta i gamma) analizowano osobno.

Wyniki pracy są zgodne z rezultatami otrzymywanymi wcześniej, tj. zaobserwowano wyłonienie się odpowiednich podsieci szczególnie współaktywnych obszarów mózgu. Wyodrębniono mianowicie klaster złożony z 4 węzłów odpowiadających tzw. sieci bazowej aktywności mózgu dla częstotliwości alfa podczas fazy odpoczynku oraz wzmocnionej synchronizacji czołowo-skroniowej i ciemieniowo-czołowej dla fal beta, wiązanych z siecią uwagi i pamięcią roboczą, w trakcie gry. Wyodrębnienie wspomnianych podsieci uzyskano poprzez obserwację dynamiki persystencji 0-homologii w czasie. Różnice w krokach filtracji, podczas których następuje zanik kolejnych spójnych składowych między etapami gry oraz odpoczynku wskazują na jakościową różnicę następującą w sieci funkcjonalnej mózgu. Wraz z rozpoczęciem fazy gry, węzły grafu dla fal beta łączą się dopiero w późniejszych krokach filtracji, co próbuje się interpretować, jako zmianę trybu pracy z globalnej na skoncentrowaną w obrębie wyspecjalizowanych podsieci. Co więcej, liczba i położenie tzw. *switchy*, czyli momentów na osi czasu w których nastąpiła istotna zmiana w topologii przestrzeni może posłużyć do oceny trwałości i stabilności aktualnego stanu sieci.

Opisane badania oczywiście nie wyczerpują wszystkich przykładów. Przytoczymy zatem parę kolejnych. Choi i wsp. [15] przeprowadzili analizę sieci funkcjonalnych FDG-PET w poszukiwaniu różnic między grupą kontrolną szczurów a grupą z indukowaną epilepsją. Dłotko i wsp. [16] badali strukturalno-funkcjonalną sieć na poziomie pojedynczych neuronów pochodzącą z cyfrowej rekonstrukcji fragmentów mózgu szczura w ramach projektu The Blue Brain Project. Natomiast Ellis i Klein [18] badali różnice skanów fMRI między grupą kontrolną a osobami ze zdiagnozowanym ADHD. Posłużyli się metodą *zbieżności topologicznej* (ang. concurrence topology), którą można potraktować jako uogólnienie podejścia stosowanego przez Curto i Itscov [7] w modelu komórek miejsca.

5.2. Wnioski metodologiczne

Najczęściej wykorzystywanymi statystykami w analizie sieci funkcjonalnych są wspominane już miary grafowe. W badaniach [11,14] autorzy porównywali metody topologiczne oraz grafowe i ostatecznie raportowali podobną skuteczność obu z nich. Można jednak wymienić parę sytuacji lub aspektów, które mogą przemawiać na korzyść TDA.

Po pierwsze, jak wspominają autorzy [11] podejście grafowe analizuje jedynie grafy jednospójne, niejako pomijając „drugoplanową dynamikę" zachodzącą dla krawędzi o mniejszych wagach. W przypadku stosowania statystyk dla grafów ważonych problem ten jest częściowo rozwiązany. Jednak gdy sieci osiągają większe rozmiary, standardową procedurą bywa binaryzacja grafu w celu redukcji badanej struktury. Krok odrzucania nieistotnych krawędzi wiąże się z koniecznością wyboru arbitralnej wartości progowania.

Powyższą sytuację można zobrazować następująco. Rozważmy przypadek 3 grup, jednej kontrolnej, jednej w której pojawia się hiper-, a w kolejnej hypo-aktywacja pewnych lokalnych podsieci funkcjonalnych. Może zdarzyć się sytuacja, w której, w przypadku binaryzacji grafu dla niskiej wartości progowej wagi (czyli łatwo dopuszczamy istnienie krawędzi), grupa kontrolna i z hiperaktywnością będą tworzyły grafy o podobnej topologii. Natomiast w przypadku restrykcyjnym (wysokiej wartości progowania), struktura grafu będzie się pokrywać między grupą kontrolną i hypoaktywności. Konieczność wyboru poziomu progowania, które właściwie odda różnice w populacji, może powodować pominięcie informacji faktycznie zawartych w danych i ignorować np. istnienie hierarchicznej struktury sieci.

Badacz zatem będzie prawdopodobnie skłonny do przetestowania paru różnych wartości progujących, tak aby rozpoznać ich wpływ na uzyskiwane wyniki. Jednak powyższemu dokładnie służy homologia persystentna, pozwalająca wydobyć te poziomy parametrów ε dla których kolejne grafy G_ε przejawiają interesujące zmiany w topologii.

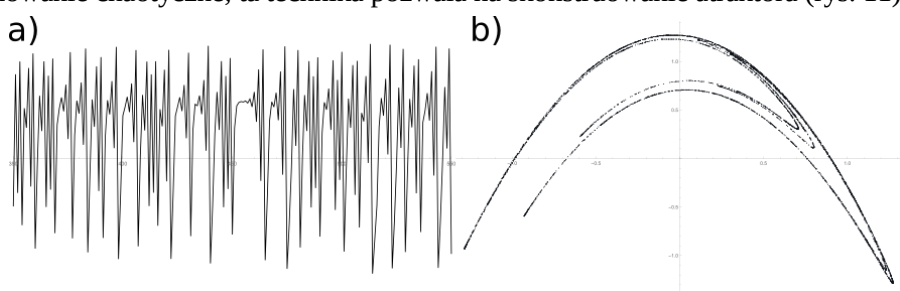
Kolejną istotną kwestią jest tzw. „zależność n-k" opisaną dokładniej w [19], powodująca, że wartości charakterystyk grafu są silnie powiązane z liczbą wierzchołków grafu (n) oraz krawędzi (k). W rezultacie dwa grafy o zbliżonej topologii, ale przeskalowane pod względem liczby węzłów będą przejawiały inne własności z punktu widzenia miar grafowych, jednocześnie charakterystyka topologiczna pozostanie niemal niezmieniona [14].

Istnieje szeroki wybór miar grafowych wraz z gotową interpretacją i opisujących konkretne aspekty sieci [18]. Można wśród nich wyróżnić charakterystyki lokalne (np. stopień wierzchołka – liczba krawędzi przystających) i globalne (np. charakterystyczna długość ścieżki – średnia odległość między parą węzłów). Istotnym podziałem tych miar jest również rozróżnienie na statystyki integracji (np. globalna efektywność) i segregacji grafu (współczynnik klasteryzacji). Propozycje odpowiedników niektórych miar grafowych w języku kompleksów symplicjalnych dopiero się pojawiają [26], a nowe jeszcze ewoluują. Rozsądne wydaje się jednak potraktowanie obu podejść jako komplementarne, dostarczające innego typu informacji o badanej sieci funkcjonalnej.

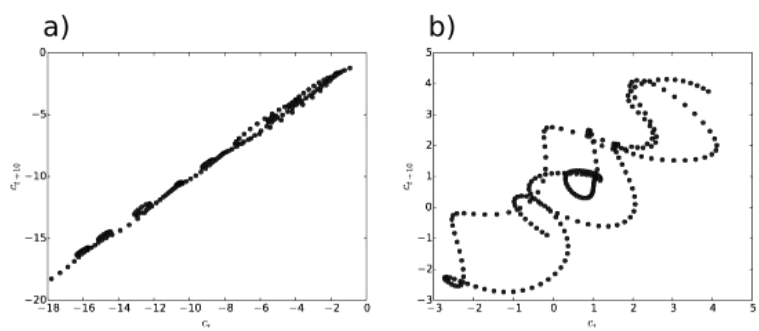
Yoo i wsp. [14] porównali swoje podejście również z metodą eigenconnectivity zaproponowana w [27]. Ta druga także wygenerowała m.in. podsieci charakterystyczne dla każdego etapu, tj. gry i spoczynku, podobne do tych powstałych przy analizie homologii, lecz ponadto powstały artefakty w postaci podsieci nie mających odzwierciedlenia w grafach, co prawdopodobnie wynika ze sztucznego założenia o istnieniu niezależnych podgrafów poszukiwanych przez tę metodę.

6. Sygnał EEG jako układ dynamiczny

Interesującym podejściem do danych EEG wykazali się autorzy pracy [28] mającej przede wszystkim charakter eksploracyjny. Badacze zaproponowali ogólny schemat analizy danych mających charakter szeregu czasowego. Sygnał EEG nie był jedynym zestawem testowanych danych, ale z racji przeglądowego charakteru tej pracy, warto wspomnieć o tym oryginalnym spojrzeniu na dane. Pomysł opiera się na twierdzeniu Takensa [29], w którym zawarta jest idea, że z szeregu czasowego $X = \{x_t\}$, składającego się z częściowych (skalarnych) pomiarów dyskretnego i deterministycznego układu dynamicznego, możliwe jest odtworzenie cech dynamiki oryginalnego układu poprzez analizę nowego układu $Y = \{y_t\}$ skonstruowanego na bazie X, gdzie $y_t = [x_t, x_{t+\tau}, ..., x_{t+n\,\tau}]$, dla pewnego skończonego n i τ, gdzie n jest wymiarem przestrzeni w której osadzamy rekonstruowany układ, a τ jest parametrem opóźnienia. Punkty z nowego ciągu możemy umieścić w przestrzeni $\mathbb{R}^n$ analizując następnie wygenerowaną trajektorię. Przykładowo, dla odwzorowania Henona, wykazującego zachowanie chaotyczne, ta technika pozwala na skonstruowanie atraktora (rys. 11).



Rysunek 30 (a) Jednowymiarowy, chaotyczny układ dynamiczny Henona oraz (b) atraktor powstały na jego podstawie w oparciu o konstrukcję zgodnie z tw. Takensa. [opracowanie własne]



Rysunek 31 Trajektorie powstałe na podstawie sygnału EEG w oparciu o ideę pochodzącą z tw. Takensa, na (a) sygnałem źródłowym był zapis EEG osoby z warunku kontrolnego, natomiast na (b) od pacjenta uzależnionego od alkoholu. Źródło: [28]

Niech X będzie ciągiem wartości odpowiadającym sygnałowi EEG. Przyjmując naiwnie założenie, że sygnał ten zachowuje się jak ciąg generowany poprzez pewien

chaotyczny układ dynamiczny oraz posiada własności wymagane przez twierdzenie Takensa, możemy wygenerować z niego nowy ciąg *Y*, metodą wspomnianą w poprzednim akapicie. Punkty z *Y* możemy się umieścić w $\mathbb{R}^n$, a następnie zbadać własności topologiczne tak wygenerowanej chmury punktów.

Pereira i Mello [28] zbadali sygnał z elektrody CPZ, czyli przedniej części kory ciemieniowej. Dane pochodziły z ogólnodstępnej bazy danych i dotyczyły grupy badawczej składającej się z osób uzależnionych od alkoholu oraz grupy kontrolnej osób zdrowych. Przyjęli parametr n=2. Następnie porównali skuteczność klasyfikacji obu grup badanych na podstawie cech topologicznych trajektorii oraz całego sygnału jako wektora cech wejściowych dla algorytmu K-means (metody powszechnie wykorzystywanego do klastrowania danych). Rozróżnienie grup na podstawie cech topologicznych okazało się być bardziej skuteczne.

7. Podsumowanie

W niniejszej pracy przedstawiono podstawowe idee związane z topologiczną analizą danych, możliwości, czasem innego niż klasycznie, spojrzenia na dane oraz przykłady jej zastosowań TDA w analizie aktywności mózgu. Największy nacisk został położony na metodę homologii persystentnych, która stanowi aktualnie trzon TDA. Opisywane podejście wykazuje się stabilnością na zniekształcenie danych, pozwala na przeprowadzenie analizy dla całego spektrum wartości parametru badanej, sparametryzowanej przestrzeni, a także umożliwia wgląd w nowy typ regularności, które dane mogą posiadać. Wydaje się również, że metody analizy sieci funkcjonalnych, takich jak te przytoczone w tej pracy, mogą ewoluować w kierunku idei dostarczonych przez TDA.

Literatura

1. Mrozek M., Żelawski M., Gryglewski A., Han S. i Krajniak A. *Homological methods for extraction and analysis of linear features in multidimensional images*, Pattern recognition, 45 (2012): s. 285-298.
2. Camara P.G., Rosenbloom D.I., Emmet K.J., Levine A.J. i Rabadan R. *Topological data analysis generates high-resolution, genome-wide maps of human recombination*, Cell Systems, 3 (2016): s. 83-94.
3. Sato M. *Can TDA be a new risk measure? An application to finance of persistent homology*, (2016) Dostępne na: http://dx.doi.org/10.2139/ssrn.2710275.
4. Kim E., Kang H., Lee H., Lee H., Suh M., Song J., Oh S. i Lee D. *Morphological brain network assessed using graph theory and network filtration in deaf adults*, Hearing research, 215 (2014): s. 88-98.
5. Chung M., Bubenik P. i Kim P. *Persistence diagrams of cortical surface dat*, Lecture notes in computer science, 5636 (2009): s. 386-397.
6. Sizemore A., Giusti Ch., Kahn A., Betzel R. i Bassett D. *Cliques and cavities in the human connectome*, (2016) arXiv: 1601.01580v1 [q-bio.NC].
7. Curto C. i Itskov V. *Cell groups reveal structure of stimulus space.* PLoS Computational biology, 4 (2008): s. 1-13.

8. Dabagian Y., Memoli F., Frank L. i Carlsson G. *A topological paradigm for hippocampal spatial map formation using persistent homology*, PLoS Computational biology, 8 (2012): s. 1-14.
9. Dabagian Y., Brandt V. i Frank L. *Reconceiving the hippocampal map as a topological template*, Elife, 3 (2014): s. 1-17.
10. Giusti Ch., Pastalkova E., Curto C. i Itskov V. *Clique topology reveals intrinsic geometric structure in neural correlations*, PNAS, 112 (2015): s. 1-29.
11. Lee H., Kang H., Chung M. K., Kim B. i Lee D. S. *Pesrsistent brain network homology from the perspective of dendrogram*, IEEE Transactions on Medical Imaging, 31 (2012): s. 2267-2277.
12. Khalid A., Kim B. S., Chung M. K., Ye J. C. i Jeon D. *Tracing the evolution of multi-scale functional networks in a mouse model of depresion using persistent brain network homology*, NeuroImage, 101 (2014): s. 351-363.
13. Petri G., Expert P., Turkheumer F. Carhart-Harris R., Nutt D., Hellyer P. J. i Vaccarino F., *Homological scaffolds of brain functional networks*. Journal of the royal society Interface, 11 (2014): s. 1-10.
14. Yoo J., Kim E.Y., Ahn Y. M. i Ye J. Ch. *Topological persistance vineyard for dynamic functional brain connectivity during resting and gaming stages*, Journal of Neuroscience Methods, 267 (2016): s. 1-13.
15. Choi H., Kim Y., Kang H., Lee H., Im H., Hwang D., Kim E., Chung J. i Lee D. *Abnormal metabolic connectivity in the pilocarpine-induced epilepsy rat model: A multiscale network analysis based on persistent homology*, Neuroimage, 99 (2014): s. 226-236.
16. Dłotko P., Hess K., Levi R., Nolte M., Reimann M. Scolamiero M., Turner K., Muller E. i Markram H., *Topological analysis of the connectome of digital reconstructions of neural microcircuits*, (2016) arXiv:160101580 [q-bioNC].
17. Ellis S. i Klein A. *Describing high-order statistical dependence using "concurrence topology", with application to functional MRI brain data*, (2013) arXiv: 1212.1642v3 [stat.ME].
18. Rubinov M. i Sporns O. *Complex network measures of brain connectivity: Uses and interpretations*, NeuroImage, 52 (2010): s. 1059-1069.
19. Van Vijk B. C. M., Stam C. J. i Daffertshofer A. *Comparing brain networks of different size and connectivity density using graph theory*, PLoS ONE, 5 (2010): s. 1-13.
20. Cohen-Steiner D., Edelsbrunner H. I Harer J. *Stability of persistence diagram*, Discrete and Computational Geometry, 37 (2007): s. 103-120.
21. Hatcher A. (2001). *Algebraic topology*, Cambridge: Cambridge University Press.
22. Burgess N. *Spatial cognition and the brain*, Annals of the New York academy of sciences, 1124 (2008): s. 77-97.
23. Edelsbrunner H. i Harer J. (2010). *Computational topology: An introduction.* Providence, R.I: American Mathematical Society.
24. Zomorodian A. *Topological data analysis*, Proceedings of symposia in applied mathematics: Advances in applied and computational topology, 70 (2011): s. 1-39.
25. Cohen-Steiner, D., Edelsbrunner, H. i Morozov, D. *Vines and vineyards by updating persistence in linear time*, W: Twenty-Second Annual Symposium on Computational Geometry, ACM, (2006): s. 119–126.
26. Giusti Ch., Ghrist R. i Bassett D. *Two's company, three (or more) is simplex*, Journal of computational Neuroscience, 41 (2016): s. 1-14.

27. Leonardi N. i Van De Ville D. *Eigenconnectivities of dynamic functional networks: consistency across subjects*, w: Conference on Signals, Systems and Computers, Asilomar (2014).
28. Pereira C. M. i de Mello R. *Persistent homology for time series and spatial data clustering*, Expert systems with applications, 42 (2015): s. 6026-6038.
29. Takens F. *Detecting strange attractors in turbulence*, W: Rand D. A. i Young L. S. (red.), Dynamical systems and turbulance, lecture notes in mathematics, Springer-Verlag, 898 (1981): s. 366-381.

Topologiczna analiza danych w neuroobrazowaniu funkcjonalnym

Streszczenie
Relatywnie młode narzędzia topologii algebraicznej coraz częściej wykorzystuje się do wykrywania nowego typu wzorców również w neuroobrazowaniu funkcjonalnym. Aktualnie najpopularniejsza metoda spośród nich – analiza homologii persystentnych, pozwala na zbadanie np. sieci funkcjonalnych o wielowarstwowej architekturze oraz na uniknięcie arbitralnego wyboru parametru (np. progowania szumu). W pracy zaprezentowane zostały podstawowe intuicje i pojęcia związane z topologiczną analizą danych oraz przykłady zastosowań tych narzędzi w analizie danych pochodzących z eksperymentów neuropsychologicznych. Uwzględnione zostały m.in. model rekonstrukcji otoczenia jedynie na podstawie współaktywności komórek miejsca w hipokampie, analiza sygnału EEG potraktowanego jako układ dynamiczny oraz analiza grafów korelacji powstałych na podstawie obrazowania funkcjonalnego. Szczególnie obiecujący jest ostatni przypadek (tj. sieci funkcjonalnych), gdzie narzędzia topologiczne mogą stanowić mocne wsparcie dla bardziej standardowych miar grafowych lub w przyszłości nawet je uogólnić. Jednak pomimo swej skuteczności np. przy klasyfikacji grup w populacji, częstym problemem pozostaje bezpośrednia interpretacja uzyskanych tą drogą wyników.
Słowa kluczowe: topologia, tda, mózg, sieci funkcjonalne

Topological data analysis in functional neuroimaging

Abstract
Usage of relatively young tools of algebraic topology in functional neuroscience in order to find new types of patterns are more and more common. The most popular method – the persistence homology analysis introduces a scheme to investigate multilayered networks. It also provide an approach in which number of arbitrary choices of parameters (i.e. data tresholding) can be reduced. The aim of this paper is to present both the basic definitions and ideas of topological data analysis, as well as examples of practical applications of these tools to neuropsychological data. Among listed research one can find model of environment coding by hippocamal place cells, analysis of EEG signal based on Taken's theorem and studies of correlation graphs built on functional imaging data. The last example seems to be particularly interesting due to opportunity arised from combining graphs measures with topological tools or even possibility of partialy replace them. This new, presented tool shows its effectivness, but sometimes the direct interpretation of the results can be problematic.
Keywords: topology, tda, brain, functional networks

Róża Dzierżak[1]

Przetwarzanie i analiza obrazów medycznych uzyskanych metodą tomografii komputerowej

1. Wprowadzenie

Obrazowanie medyczne to dynamicznie rozwijający się obszar inżynierii biomedycznej zajmujący się metodami pozyskiwania, przedstawiania oraz analizy obrazów otrzymanych podczas badań medycznych. Uzyskane obrazy wykorzystywane są do celów badawczych oraz co ważniejsze, wspomagania technicznych procesów diagnostycznych. Obecnie stosowane urządzenia pozwalają na dokładną obserwację funkcjonowania narządów wewnętrznych, stanu ich struktur oraz pracy poszczególnych układów [1]. Pozwala to na wysunięcie bardziej precyzyjnych wniosków diagnostycznych. Poza własnymi zmysłami lekarze dysponują także analizą informacji wykonaną przez coraz to nowsze urządzenia medyczne. Jedną z najczęściej stosowanych technik jest tomografia komputerowa. Dzięki swojej dokładności oraz szerokim możliwościom przetwarzania otrzymanych obrazów stała się ona bezcennym narzędziem diagnostycznym [2]. Główne wskazania do jej wykonania to obrazowanie mózgowia, odcinka lędźwiowo-krzyżowego kręgosłupa, tętniaków aorty, guzów mózgu, nowotworów pęcherza moczowego, prostaty, jajników, macicy, udarów mózgu, zmian w centralnym układzie nerwowym, tętniaków, dokładnego badania jamy brzusznej oraz urazów czaszki [1, 3, 4].

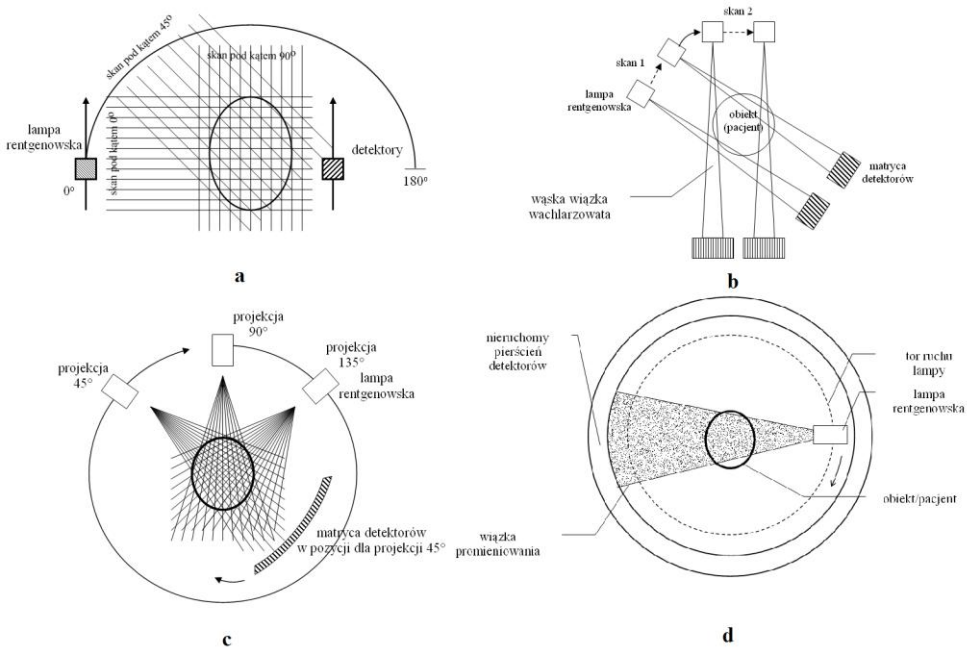
2. Tomografia komputerowa

Tomografia komputerowa (Computer Tomography) jest metodą wykorzystującą promieniowanie X oraz jego pochłanianie w strukturach wewnętrznych ciała człowieka. W tomografie komputerowym wiązka promieni X jest wysyłana przez lampę ze specjalnym kolimatorem. Stopień osłabienia wiązki przenikającej ciało pacjenta mierzony jest przez zestaw detektorów obracający się razem z lampą wokół pacjenta. Pozwala to na obejrzenie narządów wewnętrznych z różnych stron i uniknięcie zasłonięcia niektórych fragmentów [1, 2, 5].

Jednym z uczonych, którzy przyczynili się do stworzenia tomografii komputerowej był Johann Radon. W 1917 roku przedstawił on transformację całkową, która jest matematyczną podstawą tomografii. Stworzył także transformację odwrotną, która teoretycznie umożliwia odwrócenie procesu – rekonstrukcję obrazu [1]. W 1967 roku Godfrey Newbold Hounsfield opracował pierwsze urządzenie, uznane za tomograf komputerowy, natomiast w 1970 roku odkrył on, że dzięki pomiarowi adsorpcji promieniowania rentgenowskiego można ocenić gęstość tkanki. Wydarzenie to

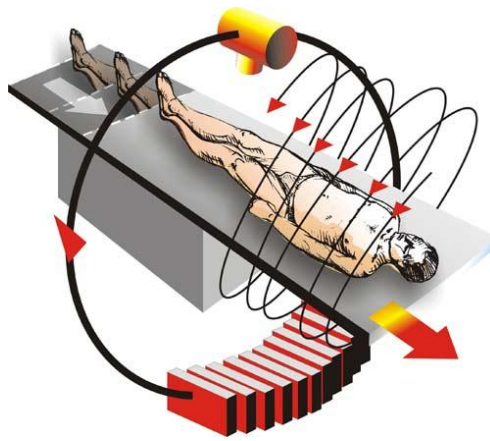
[1] r.dzierzak@pollub.pl, Instytut Elektroniki i Technik Informacyjnych, Wydział Elektrotechniki i Informatyki, Politechnika Lubelska, pollub.pl

zapoczątkowało dynamiczny rozwój w dziedzinie tomografii. W stosunkowo krótkim czasie pojawiły się urządzenia z wachlarzową wiązką promieniowania. Po kilkunastu latach stworzono tomograf spiralny. Dalsze udoskonalenia polegały na zwiększaniu liczby warstw prześwietlanych podczas jednego skoku [1-3, 5].



Rys. 1. Konstrukcje tomografów komputerowych: a) I generacji, b) II generacji, c) III generacji, d) IV generacji [5].

W tomografach I generacji (Rys. 1a) znajdowała się lampa RTG oraz jeden detektor, przez co wykorzystano jedną wiązkę skierowaną równolegle do osi detektorów, a skanowanie pacjenta uzyskiwano przez ruch translacyjno-obrotowy układu lampa-detektor. Lampa i detektory po każdym obrocie wykonywały ruch translacyjny i dokonywały serii naświetlań wzdłuż badanego obiektu. W II generacji tomografów (Rys. 1b) wzrosła jedynie liczba detektorów, a kształt wiązki zmieniono na wachlarzowaty. Natomiast III generacja była przełomem dla techniki tomografii komputerowej. Wachlarzową wiązkę promieniowania skierowano na zestaw detektorów nieruchomych względem lampy rentgenowskiej. Czas skalowania skrócił się do 10 sekund dzięki wyeliminowaniu ruchu translacyjnego. Lampa z detektorami wykonywała jedynie ruch obrotowy - raz w lewo, raz w prawo. W tomografach IV generacji zastosowano cały ich pierścień zamiast rzędu detektorów. Lampa RTG pozostała jedynym elementem ruchomym. Zmiany pozwoliły na uzyskanie podobnych czasów naświetlania jak w urządzeniach poprzedniej generacji [1, 5].



Rys. 2. Konstrukcja tomografu komputerowego V generacji [5].

Za V generację tomografów komputerowych uważa się dzisiaj stosowaną powszechnie tomografię spiralną (Rys. 2.), w której pacjent przesuwany jest wzdłuż osi gantry z jednoczesnym ruchem obrotowym lampy [5]. Wyniki badań zapisywane są w standardzie DICOM (Digital Imaging and Communications in Medicine), stworzonym w celu ujednolicenia wymiany i interpretacji danych medycznych. Dane w tym formacie mają dużą objętość przez co wymagają specjalnego oprogramowania i sprzętu komputerowego jak również łączy o wysokiej przepustowości. Pozwalają jednak na zachowanie bardzo wysokiej jakości obrazów [1, 2]. Ważną informacją zawartą w plikach DICOM jest informacja kalibracyjna. Dzięki kalibracji wszystkich obrazów przy użyciu odpowiedniej przeglądarki DICOM możliwe jest dokonywanie pomiarów i obliczeń analogicznie do sposobu oferowanego przez urządzenia diagnostyczne [6].

3. Rekonstrukcja i charakterystyka obrazów CT

Obraz pochodzący z tomografii komputerowej uzyskiwany jest na drodze obliczeń matematycznych. Uzyskanie rekonstrukcji obrazu wymaga zastosowania metod przetwarzania. Wszystkie z nich bazują na próbie odtworzenia osłabień rozkładu promieniowania w obiekcie na podstawie serii pomiarów. Istotną rolę odgrywają tu najmniejsze elementy przestrzeni w obrazach trójwymiarowych zdefiniowane jako woksele. Możemy odnieść je do roli pikseli w grafice dwuwymiarowej [5].

3.1. Metody rekonstrukcji

Odtworzenie promieniowania musi zostać wykonane w odniesieniu do każdego woksela [5]. Najczęściej stosowane metody rekonstrukcji to:

- metody algebraiczne,
- metody iteracyjne,
- metody Fouriera,
- filtrowany rzut wsteczny [7, 8].

Algebraiczna metoda rekonstrukcji obrazu polega na utworzeniu odpowiedniej ilości niezależnych równań dla wszystkich pomiarów w każdej projekcji. Przedmiot pomiaru dzielony jest na regularną siatkę wyznaczającą liczbę równań i liczbę niewiadomych w macierzy dwuwymiarowej funkcji osłabienia promieniowania. Wartość znana dla jednej projekcji i danego przesunięcia jest sumą tylu niewiadomych przez ile elementów siatki przechodzi promień skanujący. Rekonstrukcja obrazu jest w tym przypadku rozwiązaniem układu niezależnych równań [7].

Metoda iteracyjna umożliwia rekonstrukcję obrazu na podstawie pomierzonych wartości osłabień dla danych projekcji, przy przyjęciu warunków początkowych. Te same operacje matematyczne wykonywane są do momentu osiągnięcia warunku końca co oznacza rekonstrukcje obrazu. Kolejnym etapem jest dokonanie filtracji poprzez przemnożenie lub dodanie odpowiednio skonstruowanej zmiennej [7].

Metoda filtrowanego rzutu wstecznego bazuje na połączeniu odwrotnej transformaty Rodona z twierdzeniem przekroju Fouriera. Wymaga to skomplikowanych przekształceń matematycznych. Wykorzystanie tego sposobu rekonstrukcji pozwala na dokładniejszą redukcję szumów co daje nam obraz o lepszej jakości. W zależności od tego jakie struktury wewnętrze chcemy zobrazować musimy dobrać odpowiedni algorytm. Dla narządów o dużym kontraście struktur wewnętrznych jak np. kości lub płuca zaleca się stosowanie tzw. algorytmu „twardego" o wysokiej rozdzielczości, który powoduje wzmocnienie krawędzi. Obrazowanie narządów o mniejszym kontraście struktur wewnętrznych wymaga zastosowanie tzw. algorytmu „miękkiego", ponieważ algorytm „twardy" zwiększyłby poziom szumu [7, 8].

3.2. Artefakty

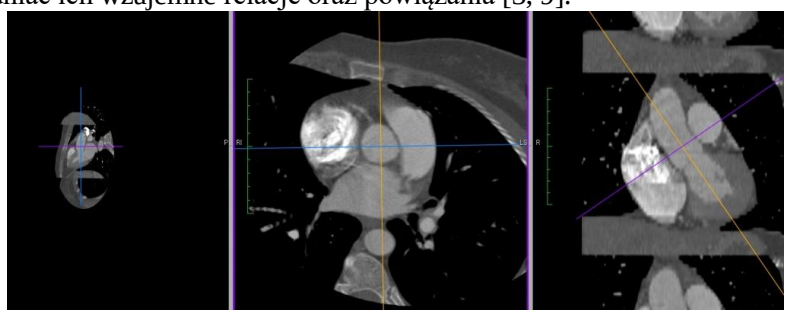
Na obrazach otrzymanych metodą tomografii komputerowej występują niekiedy elementy niewchodzące w skład badanego obiektu, tak zwane artefakty. Powoduje to utrudnienia w właściwej interpretacji otrzymanego materiału. Pojawienie się artefaktów może być związane z utwardzeniem wiązki promieniowania, całkowitym pochłanianiem fotonów, zniekształceniem obrazu badanych przedmiotów w wyniku rekonstrukcją obrazu otrzymywanego w technice spiralnej oraz efektu uśredniania.

Artefakty związane z właściwościami fizycznymi promieniowania X wynikają z niejednorodnego energetycznie charakteru wiązki promieniowania rentgenowskiego [9]. W przypadku utwardzenia wiązki promieniowania na otrzymanym obrazie widoczne są smugi i ciemne pasy pomiędzy obiektami o dużej gęstości. Kolejnym z problemów występujących podczas badania CT jest ruch pacjenta powodujący rozmycie obrazu bądź podwójne kontury. Czynnikiem pozwalającym wyeliminować tego typu artefakty jest odpowiednia prędkość pomiaru. Artefakty zależne od urządzenia tomograficznego zwykle występują w postaci pierścieni i są wynikiem niewłaściwej kalibracji detektora, bądź jej brakiem [9, 10].

4. Cechy i jakość obrazów tomografii komputerowej

Podstawowa zasada działania tomografii komputerowej opiera się na założeniu, że wewnętrzną strukturę ciała człowieka można zrekonstruować na podstawie pewnej

liczby pomiarów zewnętrznych. Wynikiem badania tomograficznego jest seria przekrojów badanej części ciała. Każdy przekrój uzyskany podczas pomiarów, charakteryzuje się zadaną grubością warstwy. Decyzje diagnostyczne podejmowane są zawsze w oparciu o równoczesną analizę kilku przekrojów. Stwarza to wyzwanie dla systemów automatycznej interpretacji tego typu danych, ponieważ muszą one posiadać funkcję równoczesnej analizy serii obrazów we wszystkich przekrojach obiektu i uwzględniać ich wzajemne relacje oraz powiązania [3, 9].



Rys. 3. Różne płaszczyzny obrazu tomografii komputerowej serca

Wartość pikseli i wokseli jest określana w jednostkach Housfileda. Jednostka Hounsfielda jest przekształceniem pomiaru liniowego współczynnika osłabienia, gdzie gęstość radiologiczna wody destylowanej w standardowej temperaturze i ciśnieniu przyjmuje wartość zera jednostek Hounsfielda (HU). Gęstość powietrza w tych samych warunkach wynosi -1000 HU. Zależność tą dla danego materiału x możemy wyrazić zależnością:

$$\frac{\mu_x - \mu_{H_20}}{\mu_{H_20} - \mu_{powiertza}} \times 1000$$

gdzie:

μ_{H_20} - liniowy współczynniki osłabienia wody

$\mu_{powiertza}$ - liniowy współczynniki osłabienia powietrza [3,6,9].

Tab. 1. Średni współczynnik pochłaniania promieni X przez różne tkanki [3]

Rodzaj tkanki	Współczynnik pochłaniania (jH.)
Kości	od 300 do 1000
Tarczyca	70 ± 10
Wątroba	65 ± 5,0
Śledziona	50 ± 5,0
Nerka	30 ± 10
Trzustka	40 ± 10
Tkanka tłuszczowa	–65 ± 10
Płuca	od –600 do –800
Płyny ustrojowe: Krew wynaczyniona Krew żylna	 80 ± 100 55 ± 5,0

Na jakość uzyskanych obrazów wpływa:

- rozdzielczość przestrzenna określająca zdolność systemu do rozpoznawania na obrazie bardzo małych obiektów,
- rozdzielczość kontrastowa określające zdolność systemu do rozróżnienia na obrazie elementów o małej różnicy gęstości,
- rozdzielczość czasowa odnosząca się do szybkości zbierania danych pomiarowych [9].

Ważnym aspektem w technice obrazowania CT jest odpowiednia zdolność rozdzielcza. Musi być ona swego rodzaju kompromisem pomiędzy możliwością rozróżniania szczegółów oraz wartością dawki promieniowania na które narażony jest pacjent. Na rozdzielczość wpływa różnica osłabienia promieniowania między szczegółem a otoczeniem. Obrazy CT ze współczesnych tomografów, w porównaniu z innymi technikami obrazowania medycznego, charakteryzują się dużą dokładnością, wysoką rozdzielczością i kontrastem. Wadą są jednak dość duże dawki promieniowania rentgenowskiego. Najlepsze efekty możemy uzyskać w przypadku obrazowania tkanek twardych. Pewne trudności mogą wystąpić przy rozróżnianiu tkanek miękkich, które charakteryzują się zbliżonym współczynnikiem absorpcji promieniowania [3, 9].

5. Etapy przetwarzania i analizy obrazów medycznych

Obraz pochodzący z urządzenia medycznego bezpośrednio po uzyskaniu jest rejestrowany w systemie określanym jako PACS (ang. Picture Achiving and Comunication System). Obraz medyczny przechowywany w pamięci systemu komputerowego w postaci reprezentacji cyfrowej staje się przedmiotem analizy pozwalającej na rozpoznawanie konkretnych struktur i właściwą ich interpretacje [6].

Komputerowe przetwarzanie obrazu medycznego może być ukierunkowane na wydobycie z ogólnego obrazu jego konkretnych elementów. W typowych przypadkach przetwarzanie obrazów jest procesem wieloetapowym, w którym kolejne przekształcenia układają się w pewien łańcuch określony przez nadrzędny cel, który należy osiągnąć [11].

Rys. 4. Etapy przetwarzania i analizy obrazów medycznych [1].

W zależności od informacji jakie chcemy uzyskać dobieramy odpowiedni sposób przetwarzania obrazów. Możliwe do przeprowadzenia operacje są w stanie pomóc nam uwypuklić cechy, które nie zawsze są widoczne podczas obserwacji „surowych" obrazów. Pierwszym etapem jest wstępne przetwarzanie, które ma na celu redukcje szumów oraz ogólnie rozumianą poprawę jakości interesujących nas obszarów. Kolejnym etapem jest segmentacja obiektów, które chcemy dokładniej przeanalizować. Proces ten pozwala zwizualizować struktury, eliminując jednocześnie elementy niepożądane, które często zaciemniają obraz danych struktur. Analiza jest procesem,

w którym na obrazie znajdowane są określone cechy pozwalające na scharakteryzowanie zasadniczych właściwości tego obrazu. Po określeniu wszystkich istotnych cech możemy wyznaczyć zbiory wartości pozwalające na klasyfikację badanych obiektów oraz interpretację ich wartości [1, 2, 11].

5.1. Przetwarzanie wstępne

Proces wstępnego przetwarzania obrazów polega na przygotowaniu go do dalszej analizy. Głównymi operacjami wykonywanymi w tym etapie są redukcja występujących szumów i stosowanie filtrów poprawiających jakość otrzymanych obrazów. Dzięki temu jesteśmy w stanie wyeliminować niepotrzebne zakłócenia oraz uwydatnić cechy obrazu istotne w dalszych etapach przetwarzania. W celu odpowiedniej interpretacji obraz poddawany jest odpowiedniej filtracji. Ten rodzaj przetwarzania nie powoduje zwiększenia ilości informacji zawartej w obrazie jednak przetworzona forma z subiektywnego punku widzenia osoby diagnozującej jest znacznie korzystniejsza [11].

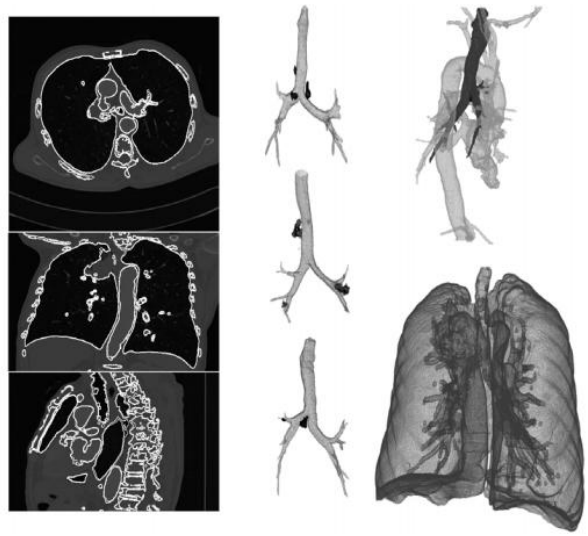
Cechą charakterystyczną obrazów jest obecność szumu. Najbardziej niekorzystny wpływ szumów występuje w obiektach o niskim kontraście, które znajdują się na pograniczu widoczności. Źródła szumów występują w każdym z etapów tworzenia obrazów medycznych. Wynikają one głównie z różnicy energii kwantów promieniowania X, losowo zmieniających się właściwości urządzeń oraz metod rekonstrukcji. Główna minimalizacja poziomów szumu prowadzona jest w trakcie rejestracji obrazu. W przypadku, gdy jest ona niewystarczająca stosuje się metody filtracji uwydatniające informacje użyteczne. Szum zazwyczaj ma widmo ciągłe przez co staje się bardziej widoczny kiedy zwiększa się średni poziom przenoszenia kontrastu w obrazie [1, 2, 11].

Wpływ szumu najprościej zredukować poprzez filtrację dolnoprzepustową obrazu, uwydatnienie znanych cech sygnału lub lokalny dobór kontrastu. Stosowane są również bardziej skomplikowane metody bazujące na transformatach falkowych jak np. transformata curvelet [12]. Podczas redukcji szumu w obrazach medycznych musimy uważać na ryzyko pogorszenie treści diagnostycznej [1, 11].

5.2. Segmentacja

Operacją umożliwiającą odpowiednie przestawienie struktur anatomicznych poddawanych dalszym procesom jest segmentacja wybranego obiektu. Pozwala zwizualizować struktury, eliminując jednocześnie elementy niepożądane, często zaciemniające obraz końcowy. Odpowiednie przeprowadzenie segmentacji jest szczególnie ważne ze względu na to, iż właśnie z tego fragmentu obrazu pobierany będzie materiał do dalszych analiz.

Należy podkreślić, że proces segmentacji jest niezwykle trudnym do precyzyjnego wykonania ze względu na zbliżoną do siebie wartości pikseli odpowiadających poszczególnym elementom sąsiadującym ze sobą [13].



Rys. 5. Przykładowe wyniki segmentacji; Pierwsza kolumna: przekroje przez dane CT z naniesionymi wynikami segmentacji w postaci białych konturów; Druga kolumna: Drzewa oskrzelowe oraz węzły chłonne (kolor czarny); Trzecia kolumna: od góry, drzewo oskrzelowe (kolor czarny) oraz okoliczne naczynia krwionośne, drzewo oskrzelowe oraz płuca [14].

Kolejnym z problemów podczas analizy obrazów CT jest nierównomierny rozkład jasności. Oznacza to, że jasność pikseli należących do tej samej tkanki jest zmienna w zależności od miejsca w obrazie. W przypadku zmian chorobowych (np. guzów, zwapnień) występujących w badanym obiekcie będą one przyjmowały wartości spoza zakresu przypisanego danej strukturze [13, 15].

Obecnie stosowanych jest kilka metody segmentacji. Najprostszym podejściem jest ręczne wyodrębnienie pożądanych struktur, które polega na obrysie konturów na poszczególnych przekrojach obrazu. Ze względu na dużą ilość czasu potrzebną na przeprowadzenie segmentacji ręcznej, powstały metody zautomatyzowane pozwalające na szybką izolację badanych struktur. Są to metody bazujące na:

- Progowaniu jasności,
- Segmentacji na podstawie brzegów,
- Segmentacji na podstawie konturów,
- Dopasowanie wzornika,
- Segmentacja tekstur [13-16].

5.3. Analiza cech obrazów na podstawie tekstury

Po przeprowadzeniu segmentacji możemy dostosować metody dalszej analizy do konkretnego obiektu którym się zajmujemy. W zależności od efektu jaki chcemy osiągnąć wyznaczamy najbardziej pożądane cechy badanego elementu pozwalające na scharakteryzowanie zasadniczych właściwości obrazów. Do najbardziej powszechnych

metod analizy obrazów należy wyznaczanie ich cech w dziedzinie przestrzennej oraz analiza tekstury.

Tekstura reprezentuje regularne cechy powierzchni obiektu. Rozpatrywaną właściwością może być na przykład ziarnistość, kierunek ułożenia wzoru, jednorodność, lokalny kontrast, czy też średni poziom jasności pikseli danego obszaru obrazu. W literaturze spotyka się cztery główne podejścia do ekstrakcji cech teksturalnych. Są to metody statystyczne (analiza histogramu poziomów szarości, macierzy gradientów, macierzy współwystępowania), modele matematyczne (głównie modele autoregresji i fraktalne), metody transformacyjne (transformata falkowa, filtry Gabora) oraz podejście strukturalne w którym poszukiwany jest podstawowy, powtarzający się element tekstury (tzw. teksel) i określane reguły jego rozmieszczenia [17-20].

W kolejnym etapie wyznaczone cechy podlegają redukcji w celu znalezienie ich podzbioru, który optymalny z punku widzenia charakterystyki tekstur występujących w obrazach [18].

Analiza tekstur ma na celu znalezienie zbioru parametrów, nazywanych cechami teksturalnymi, z których każdy jest liczbową miarą określonej właściwości tekstury. Parametry tekstury definiuje się w celu scharakteryzowania jednorodnych obszarów obrazu, określonych za pomocą różnorodnych rozkładów jasności [19]. Wyznaczone parametry tekstury muszą spełniać następujące wymagania:

- umożliwienie jednoznacznego rozróżniania analizowanych tekstur nawet dla niewielkich obszarów ROI (Region of interest),
- brak wzajemnej korelacji , wartość dyskryminacyjna skorelowanego zbioru cech jest zbliżona do jej wartości dla poszczególnych cech,
- możliwość skorelowania danego parametru z cechami struktury lub właściwościami fizycznymi obrazowanych narządów lub tkanek [17-20].

5.4. Klasyfikacja tekstury obrazu

Klasyfikacja tekstury obrazu związana jest z określeniem właściwości charakteryzujących obszary zainteresowania. W analizach diagnostycznych jest to ocena zmian patologicznych narządu uwidocznionego w danym obszarze obrazu. Wartości parametrów opisujących ten obszar mogą zostać przypisane do jednej z określonych wcześniej klas określających stopień zmian patologicznych w danym narządzie [21-23].

Metody klasyfikacji możemy podzielić na nadzorowane i nienadzorowane. W tych dwóch grupach rozróżniamy:

- metody wykorzystujące odległość wyznaczoną pomiędzy próbkami w przestrzeni cech jak np. klasyfikator k-NN oraz metoda k-średnich należąca do analizy skupień,
- metody przeprowadzające podział przestrzeni cech na obszary przypisane różnym klasom jak np. sieci perceptorowe, liniowa i nieliniowa analiza dyskryminacyjna, metoda SVM i drzewa decyzyjne,
- metody probabilistyczne, gdzie estymuje się prawdopodobieństwo przynależności próbki do poszczególnych klas jak w przypadku klasyfikatora Bayesa [18, 22, 23].

6. Podsumowanie

Obecnie stosowane tomografy komputerowe oraz metody rekonstrukcji pozwalają na uzyskanie wysokiej jakości obrazów ciała człowieka. Dzięki serii obrazów wykonanych przy niewielkiej grubości zadanej warstw oraz trzem przekrojom badanego elementu, możliwe jest dotarcie do prawie każdego zakamarka badanego narządu. Wachlarz metod przetwarzania otrzymanych obrazów pozwala usunąć z niego ewentualne niedoskonałości i uwypuklić najbardziej pożądane cechy. Metody analizy cech wyselekcjonowanych elementów obrazu pozwalają na stworzenie klasyfikatorów, dzięki którym jesteśmy w stanie rozróżniać obszary tkanek ze zmianami patologicznymi. Proces przetwarzania i analizy obrazów medycznych ma nieocenioną funkcję diagnostyczną. Możliwość dostosowania obrazu tak, aby wyeksponować cechy świadczące o obecności zmian, często decyduje o postawieniu przez lekarza właściwej diagnozy. Rozwój zagadnień dotyczących obrazowania medycznego zmierza w kierunku całkowitej automatyzacji procesu diagnostycznego, aby wyeliminować ewentualne pomyłki związane z ludzką percepcją. Metody te są bardziej precyzyjne i pozwalają uzyskać więcej informacji w znacznie krótszym czasie niż tradycyjne metody. Szybka i dokładna diagnostyka komputerowa wspomaga leczenie oraz profilaktykę wielu chorób.

Literatura

1. Tadeusiewicz R., Śmietański J. *Pozyskiwanie obrazów medycznych oraz ich przetwarzanie, analiza, automatyczne rozpoznawanie i diagnostyczna interpretacja*, Wydawnictwo Studenckiego Towarzystwa Naukowego, Kraków 2011.
2. Tadeusiewicz R. pod red. nauk. *Inżynieria biomedyczna : księga współczesnej wiedzy tajemnej w wersji przystępnej i przyjemnej*, AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Kraków 2008.
3. Budzik G., Dziubek T., Turek P. *Podstawowe czynniki wpływające na jakość obrazów tomograficznych*, Problemy Nauk Stosowanych, Tom 3, s. 77 – 84, 2015.
4. Kulawiak-Gałąska D., Gałąska R., Pieńkowska J., Fijałkowska J., Szurowska E. *Aplication of computed tomography in the diagnostic of cardiac diseases*, Ann. Acad. Med. Gedan, 43,s. 135-146, 2013.
5. Chmielewski L., Kulikowski J. L., Nowakowski A., *Obrazowanie Biomedyczne* , Tom 8, Biocybernetyka i inżynieria biomedyczna 2000 pod redakcją M. Nałęcza, Wyd. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2003.
6. Orgiela M., R., Tadeusiewicz R., *Modern computational intelligence methods for the interpretation of medical images*, Springer, 2008.
7. Cierniak R. *Tomografia komputerowa : budowa urządzeń CT : algorytmy rekonstrukcyjne*, Akademicka Oficyna Wydawnicza Exit, 13-199, Warszawa 2005.
8. Budzik G., Turek P., *Proces rekonstrukcji obrazów tomograficznych*, Problemy Nauk Stosowanych, Tom 4, s. 057 – 064, 2016.
9. Przelaskowski A., *Miary jakości* w „Multimedia - Algorytmy i Standardy kompresji" pod redakcją W. Skarbka, Akademicka Oficyna Wydawnicza PLJ, s. 111-142.
10. Świątkowski J., Jarkiewicz-Kochman E., Pacholec E., Benke G., Gołębiowski M., Błażewicz S., Wojciechowski A., Goździk J. *Zakłócenia obrazu w badaniach tomografii komputerowej (TK) i tomografii rezonansu magnetycznego (MR)*, Inżynieria Biomateriałów, 8, 65—66, 2005.
11. Materka A., Strumiłło P. *Wstęp do komputerowej analizy obrazów* Politechnika Łódzka, Łódź 2009.

12. Mala K., Sadasivam V. *Automatic Segmentation and Classification of Diffused Liver Diseases using Wavelet Based Texture Analysis and Neural Network*, Annual IEEE INDICON Conference, s. 216 – 219, 2005.
13. Ławicki, T. and Zhirnova, O. *Application of curvelet transform for denoising of CT images*, Proc. of SPIE 9662, 966226, 2015.
14. Skalski A., *Segmentacja 3D danych medycznych pochodzących z tomografii komputerowej oraz endoskopowych zapisów wideo*, rozprawa doktorska, AGH, Kraków 2009.
15. Strzelecki, M. *Segmentacja tekstury obrazów z wykorzystaniem neuronowych sieci oscylacyjnych i metod statystycznych*, Politechnika Łódzka, Z. 336, 3-177, 2004.
16. Suri J. S., Kamaledin Setarehdan S., Singh S. *Advanced Algorithmic Approaches to Medical Image Segmentation*, Springer 2002.
17. Duda D., Krętowski M., Bézy-Wendling J. *Ekstrakcja cech teksturalnych w klasyfikacji obrazów tomograficznych wątroby*, Zeszyty Naukowe Politechniki Białostockiej,Informatyka – Zeszyt 2, 2007.
18. Strzelecki M., Materka A. *Tekstura obrazów biomedycznych. Metody analizy komputerowej* , Wyd. PWN, Warszawa 2017.
19. Bruno A., Collorec R., Bézy-Wendling J., Reuzé P., Rolland Y. *Texture analysis in medical imaging* W: Roux C., Coatrieux J. L. (edytorzy): Contemporary Perspectives in Three-dimensional Biomedical Imaging, IOS Press, s. 133-164, 1997.
20. Duda D., Krętowski M., Bézy-Wendling J. *Texture characterization for Hepatic Tumor Recognition in Multiphase CT* , Biocybernetics and Biomedical Engineering, 26(4), s. 15-24, 2006.
21. Chen C., Daponte J. S., Fox M. D. *Fractal feature analysis and classification in medical imaging* , IEEE Transactions on Medical Imaging, 8, s. 133- 142, 1989.
22. Chen E. L., Chung P. C., Chen C. L., Tsai H. M., Chang C. I. "An automatic diagnostic system for CT liver image classification", IEEE Transactions on Biomedical Engineering, 45(6), s. 783-794, 1998.
23. Haralick R. M., Shanmugam K., Dinstein I. *Textural features for image classification*, IEEE Transactions on Systems, Man and Cybernetics, 3, s. 610-621.

Przetwarzanie i analiza obrazów medycznych uzyskanych metodą tomografii komputerowej

streszczenie

W pracy przedstawiony został mechanizm powstawania oraz rekonstrukcji obrazów uzyskiwanych metodą tomografii komputerowej. Opisane metody przetwarzania i analizy obrazów pozwalają na usunięcie ewentualnych zakłóceń obrazów oraz uwydatnienie badanych struktur. Zaprezentowane metody ekstrakcji cech pozwalają na utworzenie zbioru parametrów określających badane tkanki. Dzięki temu możemy zbudować klasyfikatory pozwalające na przypisanie badanego obszaru do określonej grupy zmian patologicznych. Jest to podstawą do tworzenia systemów automatycznej interpretacji diagnostycznej.

Słowa kluczowe: tomografia komputerowa, przetwarzanie obrazów, obrazowanie medyczne, analiza obrazów, diagnostyka medyczna

Processing and analysis of medical images obtained by computed tomography

Summary

The paper presents the mechanism of creation and reconstruction of images obtained by computer tomography. Described methods of image processing and analysis allow to remove any disturbance of images and enhance the examined structures. Presented methods of feature extraction allow to create a set of parameters determining the examined tissues. This makes it possible to build classifiers that allow the examined area to be assigned to a specific group of pathological changes. This is the basis for creating automated diagnostic interpretation systems.

Key words: computer tomography, image processing, medical imaging, image analysis, medical diagnostics

Sebastian Piłat[1], Joanna Szulżyk-Cieplak[2]

Charakterystyka nowoczesnych technologii multimedialnych w aspekcie efektywnego wykorzystania w procesie nauczania

1. Wstęp

Od wielu lat obserwuje się dynamiczne przemiany kulturowe, ekonomiczne i naukowo-techniczne, które powodują znaczący wzrost znaczenia technologii informacyjnej. Wobec zaistniałych zmian pojawiła się konieczność przygotowania młodzieży do prawidłowego funkcjonowania w społeczeństwie informacyjnym, co wymusza ścisłą korelację procesu współczesnego kształcenia z nowoczesnymi technologiami. Osiągnąć to można poprzez wprowadzenie do programów kształcenia ogólnego, zagadnień związanych z gromadzeniem, selekcją, opracowywaniem i prezentacją informacji oraz twórczym rozwiązywaniem zadań. [1]

Zastosowanie nowoczesnych technologii, w tym informacyjnych, intensyfikuje proces uczenia się, pomaga uczącym się w lepszym zrozumieniu przekazywanych treści, a nauczycielom z kolei dostarcza zaawansowanych, interaktywnych i angażujących uwagę ucznia metod nauczania. Istotną rolę w procesie dydaktycznym odgrywają wszelkiego rodzaju środki multimedialne. Wprowadzenie multimediów do procesu kształcenia przyczyniło się do powstania nowej jakości w edukacji. Oprócz większej efektywności, wynikającej z wykorzystania sfery emocji w procesie poznawczym, technologie te oferują szereg mechanizmów dzięki którym prezentacja skomplikowanych problemów staje się bardziej prosta i przystępna.

Multimedia umożliwiają swobodną oraz interaktywną wymianę informacji w postaci tekstu, grafiki, obrazu lub dźwięku. Są one także głównym źródłem informacji o otaczającym świecie. Ich ogromna rola w procesie kształcenia wynika przede wszystkim z umożliwienia uczniom szybkiego i łatwego dostępu do wiadomości oraz związana jest z dużą skutecznością w zapamiętywaniu przekazywanych treści. Potrafią one zaciekawić ucznia i dostosować treści do jego umiejętności i zainteresowań. Wymienione cechy sprawiają, że stanowią one w obecnych czasach istotny czynnik stymulujący przemiany w edukacji.

2. Technologie multimedialne w procesie kształcenia

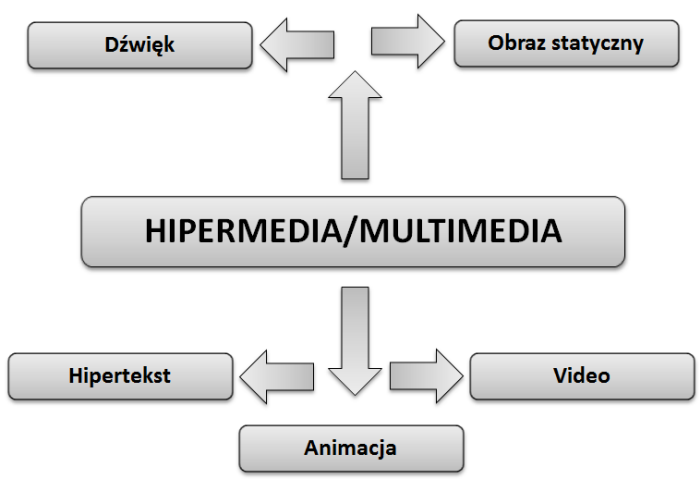
Proces kształcenia jest podstawową formą realizacji celów edukacyjnych i stanowi system powiązanych ze sobą działań nauczycieli i uczniów, w toku których uczący przekazuje uczniom wiedzę, bądź kieruje ich pracą, stwarzając warunki do samodzielnego zdybywania wiadomości, rozwijania osobowości i kształtowania postaw. W procesie tym coraz większą rolę odgrywają różnego rodzaju media. [2]

[1] sebastian.pilat@pollub.edu.pl; Wydział Podstaw Techniki, Politechnika Lubelska, www.pollub.pl
[2] j.szulzyk-cieplak@pollub.pl; Katedra Podstaw Techniki, Wydział Podstaw Techniki, Politechnika Lubelska, www.pollub.pl

Z punktu widzenia procesu dydaktyczno-wychowawczego terminów media oraz multimedia używa się zarówno w odniesieniu do środków masowego oddziaływania, jak i środków dydaktycznych. Pod pojęciem media rozumie się różnego rodzaju przedmioty i urządzenia przekazujące odbiorcom określone informacje poprzez słowa, obrazy i dźwięki, a także umożliwiające im wykonywanie określonych czynności intelektualnych i manualnych. Zaliczamy do nich m.in. telewizję, gazety, radio i Internet. Podają one informacje w formie komunikatu oraz umożliwiają zdybywanie informacji i porozumiewanie się. [3] Z kolei używając terminu multimedia mamy na myśli zastosowanie równocześnie kilku różnych form przekazu informacji, np. połączenie obrazu w formie animacji, tekstu oraz dźwięku jako podkładu muzycznego.

Zasadniczą rolą technologii multimedialnych jest pośrednictwo między człowiekiem i tym wszystkim, co jest poza zakresem jego bezpośredniego oglądu i doświadczenia. Dają one możliwość prezentacji zjawisk których nie jesteśmy w stanie bezpośrednio zaobserwować. Multimedia nie są ograniczone fragmentami rzeczywistości, na temat którym chcemy wiedzę przekazać, a jedynie naszymi wiadomościami o nich oraz umiejętnościami posługiwania się narzędziami multimedialnymi w celu zaprezentowania wybranych treści. Dla przykładu, aby obejrzeć pod mikroskopem próbkę materiału musimy fizycznie posiadać odpowiedni sprzęt oraz tą próbkę. Aby ją zademonstrować w formie multimedialnej wizualizacji wystarczy, że mamy wiedzę na temat jej budowy i właściwości, a także umiejętności wykonania jej modelu przy wykorzystaniu odpowiednich technologii multimedialnych. W efekcie możemy przekazać wiedzę tak samo jak przy użyciu mikroskopu, a dodatkowo odpowiednie wykonanie wizualizacji pozwoli uczniom lepiej zrozumieć przekazywane treści.

Niezwykle ważne znaczenie w procesie współczesnego kształcenia zajmuje system multimedialny, przez który rozumiemy dowolny system teleinformatyczny zdolny do przetwarzania, archiwizacji i dystrybucji danych w postaci dźwięku, ruchomych obrazów, fotografii, grafiki komputerowej i tekstów. [2] Przy konstruowaniu multimedialnego systemu kształcenia należy określić jego elementy składowe, takie jak: hipertekst, obraz statyczny, obraz ruchomy, animację oraz dźwięk (rys. 1).



Rysunek 1. Elementy systemu multimedialnego [4]

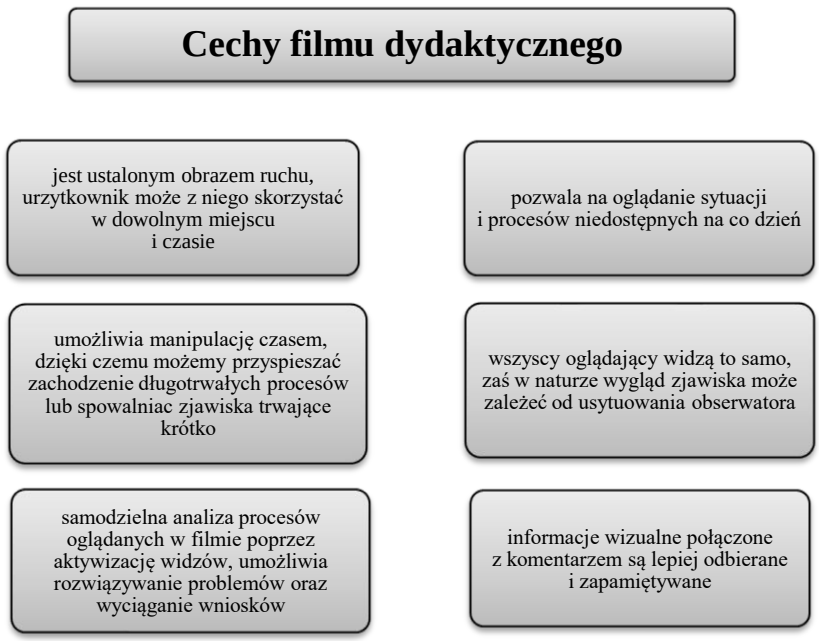
2.1. Elementy systemu informatycznego

Obraz (zarówno statyczny, jak i dynamiczny) stanowi istotny element sytemu multimedialnego. Ułatwia on osobom uczącym się zrozumienie omawianych treści oraz przyśpiesza proces ich zapamiętywania. Zastosowanie grafiki, np. w prezentacji multimedialnej pozwala uniknąć monotonii. Należy jednak pamiętać, że wszystkie elementy graficzne powinny tworzyć spójną całość, wyróżniając jednocześnie ważne informacje. [5]

Animacja komputerowa w ogólnym ujęciu polega na ożywieniu obiektów na ekranie, poprzez odpowiednio szybkie przesuwanie klatka po klatce serii obrazków, ukazujących kolejne fazy ruchu. Z animacją jest ściśle związane pojęcie FPS (ang. *frames per second*), czyli ilością klatek statycznego obrazu wyświetlanych w ciągu jednej sekundy. Oczywiście, im więcej takich klatek jest wyświetlanych, tym obraz który finalnie widzimy na ekranie sprawia wrażenie bardziej płynnego. W przypadku filmów nadawanych w telewizji ilość klatek wyświetlanych na sekundę jest stała i wynosi, w zależności od kodowania sygnału, od 25 do 30. Jest to w zupełności wystarczające do tego aby w odbiorze przez oko ludzkie imitować zjawisko płynnego ruchu. Oczywiście ilość FPS potrzebna do wrażenia płynności jest zależna od szczegółowości i stopnia złożoności wyświetlanego obrazu. W przypadku prostych obrazków dwuwymiarowych wywołanie złudzenia ruchu obiektów może nastąpić już przy 4-5 klatkach na sekundę. W przypadku grafiki renderowanej w czasie rzeczywistym wartość FPS jest zmienna i zależy w dużej mierze od wydajności sprzętu, na którym jest prezentowana. Oko człowieka reaguje na zmienne ilości FPS w nieco inny sposób niż na stałą jego wartość i przyjmuje się, że aby imitować pełną płynność w przypadku zmiennej ilości klatek wyświetlanych w każdej sekundzie musi ich być nie mniej niż 55-60. Animacje w systemach multimedialnych mogą przybierać rożne formy – począwszy od umieszczenia anonimowanych obrazków na wybranych slajdach, przez animację pojawiających się na slajdzie pól tekstowych, schematów, wykresów, diagramów i tabel, a kończąc na anonimowanych przejściach pomiędzy kolejnymi slajdami. Taka forma prezentacji pobudza wyobraźnię uczniów, przyciąga uwagę oraz motywuje do aktywnego odbioru prezentowanych treści, a także pobudza do dyskusji[6]

Dźwięk z uwagi na dużą możliwość aktywizowania emocji jest bardzo ważnym elementem systemu multimedialnego stosowanego w procesie kształcenia. Jest on wykorzystywany głównie w dwóch aspektach. Pierwszy, to rola lektora w przypadku np. filmów edukacyjnych lub samodzielne nagranie głosu wykorzystywane m.in. przy nauce języków obcych. Drugi aspekt dotyczy zabiegu angażowania sfery emocjonalnej ucznia w proces nauczania. Polega on na wykorzystaniu dodatkowego medium do przykucia uwagi ucznia do przekazywanych treści, a także stymulowania skupienia się. Działanie takie sprowadza się często do zastosowania podkładu dźwiękowego w filmach lub prezentacjach multimedialnych. Należy jednak pamiętać, aby z takich zabiegów korzystać z rozwagą. Odpowiednie wykorzystanie dźwięku może przykuć uwagę ucznia, jednak jeśli zechcemy wykorzystać tę technikę zbyt dosadnie, może ona w konsekwencji odwrócić uwagę ucznia od prezentowanych treści, powodując że skupi się on na dźwięku bardziej niż na informacji, która ma zostać przekazana.

Wideo łączy w sobie zarówno obraz, jak i dźwięk. Spośród wszystkich elementów systemu multimedialnego film edukacyjny stanowi jedną z najbardziej atrakcyjnych form przedstawiana treści dydaktycznych, a jego głównym celem jest przekazanie informacji w prosty i ciekawy sposób. Na schemacie (rys. 2) przedstawiono najważniejsze cechy filmu dydaktycznego, decydujące o jego dużej przydatności w procesie kształcenia.



Rysunek 2. Główne cechy filmu dydaktycznego [2]

Pliki wideo znalazły szerokie zastosowanie nie tylko wśród metod wykorzystywanych w procesie nauczania, ale również jako forma sprawdzania wiedzy. Najlepszym przykładem jest egzamin na prawo jazdy. W trakcie egzaminu osoba zdająca jest postawiona w określonej sytuacji która jest wizualizowana właśnie na podstawie krótkiego filmu. Ciężko jest sobie wyobrazić, aby w inny sposób sprawdzić nie tylko wiedzę, ale również czas reakcji osoby podchodzącej od egzaminu.

Ostatnim elementem sytemu multimedialnego jest hipertekst definiowany jako sposób pisania tekstu i łączenia go za pomocą odsyłaczy, które wyglądają jak podkreślony tekst (najczęściej pisany niebieską lub zieloną czcionką) – po kliknięciu na który czytający zostaje przeniesiony na inną, często związaną tematycznie stronę. Najbardziej typowym przykładem zastosowania hipertekstu są strony WWW. Hipertekst łączy informację w „pajęczynę", po której użytkownik może się dowolnie poruszać.

2.2. Charakterystyka wybranych narzędzi multimedialnych stosowanych w procesie kształcenia

W codziennej pracy nauczyciel ma przed sobą jeden podstawowy cel, a mianowicie przekazać informacje w sposób jasny i czytelny. Przedstawione treści powinny w jak największym stopniu zapadać w pamięć, a sposób ich prezentacji powinien zainteresować słuchaczy. Opierając się wyłącznie na środkach tradycyjnych, czyli użyciu kredy i tablicy oraz ustnej wypowiedzi, nauczyciel musi wykazać się niebywałymi umiejętnościami retorycznymi. Na uzyskanie większej efektywności w przekazywaniu wiadomości pozwala wykorzystanie w procesie dydaktycznym nowoczesnych multimedialnych technologii informatycznych.

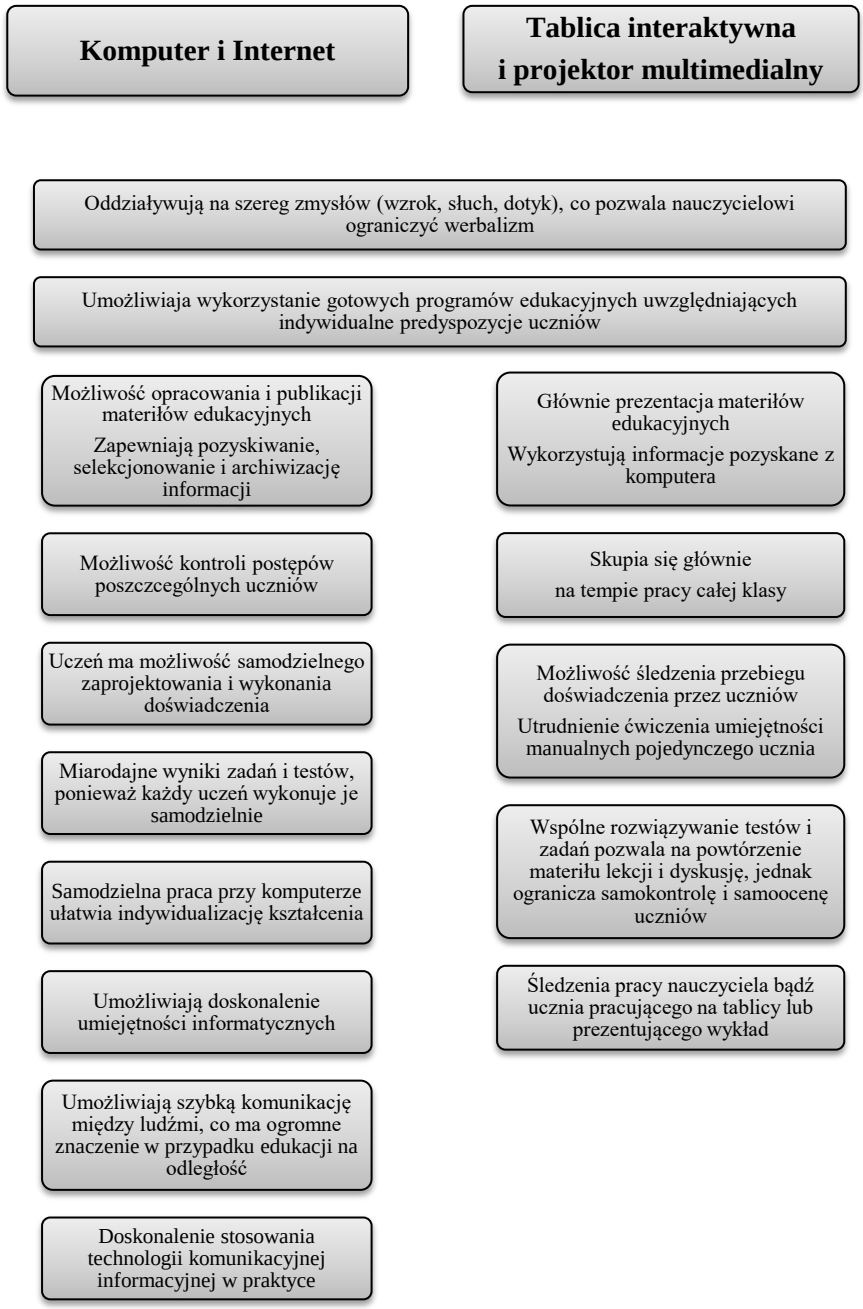
Nauczyciel będący częścią społeczeństwa informacyjnego jest niejako zobligowany do korzystania z dobrodziejstw nowoczesnych narzędzi dydaktycznych, wśród których wymienić należy: prezentacje multimedialne, tablice interaktywne oraz komputer i Internet. Stosowanie wymienionych narzędzi informatycznych podczas zajęć wzbudza u uczniów zainteresowanie i fascynację przedmiotem, co ma pozytywny wpływ na ich motywację do nauki. [1]

Jednym z najnowocześniejszych multimedialnych narzędzi dydaktycznych jest tablica interaktywna. Posiada ona szereg funkcji (m.in. wczytywanie notatek z tablicy do pamięci komputera, odtwarzanie filmów z możliwością wykonywania notatek bezpośrednio na pojedynczych kadrach, praca z aplikacjami komputerowymi na powierzchni tablicy), którymi nie dysponują pozostałe środki dydaktyczne, nawet komputer sprzężony z projektorem multimedialnym. Stosując tablice interaktywne podczas zajęć nauczyciel realizuje postulat wizualizacji w procesie nauczania.

Innymi, niezwykle potężnymi narzędziami multimedialnymi wspomagającymi pracę nauczyciela, są komputer i Internet. Komputer stanowi wszechstronny środek dydaktyczny. Umożliwia polisensoryczność przekazu, czyli poznanie wielozmysłowe. Następuje w ten sposób przekroczenie werbalizmu i emocjonalne zaangażowanie uczniów w kojarzeniu wartości poznawczych z estetycznymi (tekst + słowo + muzyka + grafika + film) [1]. Zastosowanie Internetu w edukacji jest przykładem nowego sposobu kształcenia i samokształcenia, o czym decydują następujące jego cechy:

- interaktywność – uczeń ma możliwość odbierania komunikatów i reagowania na nie oraz możliwość wymiany danych z innymi użytkownikami),
- multimedialność – Internet łączy w sobie wszystkie media tradycyjne, zawiera komunikaty w formie tekstu, obrazu, dźwięku, animacji, filmu wideo,
- hipertekstowość – Internet jest zbiorem informacji publikowanych w sposób nielinearny, ułatwiający sprawne dotarcie do tematów pokrewnych oraz haseł kluczowych,
- prostota wyszukiwania informacji – wyszukiwarki internetowe ułatwiają szybkie dotarcie do poszukiwanych informacji, a systemy znaczników wspomagają ich porządkowanie i archiwizację [7].

Nauczyciel wykorzystujący w swojej pracy nowoczesne środki dydaktyczne przed ich zastosowaniem musi przeanalizować zalety i wady poszczególnych rozwiązań, tak aby przekazywanie treści odbywało się w sposób jak najbardziej efektywny (rys.3).



Rysunek 3. Porównanie nowoczesnych multimedialnych narzędzi informatycznych stosowanych w procesie kształcenia [1]

3. Efektywność procesu kształcenia, a korzystanie z technologii multimedialnych

Jak przedstawiono na wstępie artykułu, multimedia umożliwiają uczniowi dotarcie do wartościowych pod względem naukowym i edukacyjnym wiadomości, jak również stymulują jego rozwój. Wśród zalet wykorzystania technologii multimedialnych w procesie edukacji wymienić należy:

- pobudzanie ciekawości oraz chęci i gotowości uczenia się,
- możliwość indywidualnej pracy i przyswajania nowej wiedzy w tempie dostosowanym do każdego ucznia,
- wzrost samodzielności uczniów i rozwój umiejętności twórczych,
- wzbogacenie zajęć lekcyjnych o nowe rozwiązania metodyczne,
- umożliwienie jednoczesnego uczenia się i zabawy, co powoduje podniesienie efektywności nauczania,
- wzrost aktywności uczniów na interaktywnych lekcjach,
- wzrost komunikatywności podczas zajęć z komputerem (uczniowie rozmawiają ze sobą, wzajemnie zadają sobie pytania i udzielają na nie odpowiedzi, głośno komentują swoje sukcesy),
- wyrabianie u uczniów nawyku myślenia twórczego [8].

Jednym z następstw szybkiego rozwoju technologii multimedialnych jest konieczność wykształcania u uczniów odpowiednich kompetencji ułatwiających, a niekiedy umożliwiających wykorzystanie pełnego potencjału kryjącego się w zasobach sieci Internet. Tradycyjne metody pamięciowe, należy zastąpić szeregiem metod i umiejętności związanych z wyszukiwaniem, selekcją i weryfikacją informacji. (Rys. 4)

Rysunek 4. Proces prowadzący do odnalezienia potrzebnej informacji oraz czynności które należy wykonać w każdym z kroków.

W celu jak najszerszego wykorzystania możliwości oferowanych przez Internet, w pierwszej kolejności należy opanować umiejętność wyszukiwania wartościowych informacji. Podstawową umiejętnością w tym zakresie jest zdolność do określenia problemu, a następnie nazwania go w taki sposób, aby zoptymalizować działanie wyszukiwarki. Następnie, z racji na dużą ilość treści w zasobach sieci, konieczna jest selekcja oraz proces weryfikacji. Sam proces definiowania problemu okazuje się nie być zbyt skomplikowany, kiedy wiemy konkretnie jakiej informacji szukamy. Jednak

zdarzają się sytuacje w których dysponujemy jedynie szczątkowymi informacjami dotyczącymi zagadnienia. W takim przypadku często nie jesteśmy pewni, czy szukamy w odpowiedni sposób, jednak to możemy sprawdzić na etapie weryfikacji.

W momencie kiedy mamy już określony i nazwany problem, należy zastanowić się w jaki sposób opisać go, aby został dobrze zinterpretowany przez wyszukiwarkę i aby wyświetlone wyniki były optymalne. Istnieje kilka technik usprawniania procesu wyszukiwania. Jedną z nich jest wyszukiwanie hasłowe. Polega ono na tym, że podajemy wyszukiwarce jedynie słowa kluczowe z poszukiwanego zagadnienia i w uzyskanych wynikach szukamy informacji które nas interesują. W tym przypadku należy wspomnieć również o operatorach wyszukiwania. Są to znaki które możemy wstawić w polu wyszukiwania pomiędzy wyszukiwanymi słowami, aby uzyskać pożądany efekt. Przykładem niech będzie znak „-". Wpisując w polu wyszukiwania frazę: „klucz –wiolinowy" zostaną wyszukane informacje dotyczące słowa „klucz", z pominięciem słowa „wiolinowy". Podobnych operatorów funkcjonuje wiele i są one bardzo przydatnym narzędziem przy wyszukiwaniu informacji i zawężaniu wyników wyszukiwania.

Po uzyskaniu wyników wyszukiwania należy poddać je procesowi selekcji, a więc wstępnie określić, czy wyszukana informacja jest tą o którą nam chodziło i czy nie jest inną interpretacją naszego zapytania. W wyniku selekcji, powinno się uzyskać kilka informacji, które odpowiadają naszym wymaganiom. Kiedy posiadamy już informacje które przeszły pozytywnie przez proces selekcji, należy poddać je jeszcze weryfikacji. W dzisiejszych czasach Internet pełen jest informacji niepotwierdzonych i nierzetelnych, dlatego nie można ufać wszystkiemu, co się w nim znajduje. Najważniejsze jest określenie źródła. Jeśli nasza informacja pochodzi z artykułu czy publikacji naukowej, możemy przyjąć, że informacja jest sprawdzona i poprawna merytorycznie. W innych jednak przypadkach każdą odnalezioną informację należy sprawdzić. Najprostszą metodą jest odnalezienie takiej samej, bądź znacząco zbliżonej informacji w innych źródłach. Jeśli informacje potwierdzają się, można przyjąć, że są poprawne.

Cały opisany proces wyszukiwania informacji jest jedną z podstawowych kompetencji, które powinien nabywać uczeń w procesie nowoczesnej edukacji. Przy tak szerokim i powszechnym dostępnie do informacji jaki oferuje dzisiaj Internet zagrożenie ze strony błędnych informacji jest bardzo duże, dlatego nadzwyczaj ważna jest umiejętność odnalezienia wartościowej informacji.

Jak pokazują badania [9, 10] pomimo rosnącego zainteresowania wykorzystaniem technologii multimedialnych nauczyciele przejawiają o wiele mniejsze zainteresowanie w tym zakresie aniżeli uczniowie. Może to oczywiście wynikać z faktu, że młodzi ludzie są przyzwyczajeni do stanu, w którym znaczną częścią ich codziennego życia stanowią nowoczesne technologie. Wyniki badań M. Abelite [9] wskazują, że ponad 70 % uczniów spędza na pracy z komputerem dwie, lub więcej godzin w ciągu dnia. Ten wynik ukazuje jak dużą rolę w życiu współczesnej młodzieży odgrywa technologgia komputerowa, przy czym należy pamiętać, że są to wyniki badań sprzed czterech lat. Przy dzisiejszym tempie rozwoju technologii oraz jej upowszechnieniu, można spekulować o wzroście tej wartości w na przestrzeni ostatnich lat. W przypadku nauczycieli częstotliwość korzystania z zasobów Internetu w celu przygotowania zajęć

dydaktycznych wygląda następująco: 53,9% korzysta codziennie lub prawie codziennie, 26,3% – co 2-3 dni, 13 % – raz w tygodniu, 6,5% - rzadziej niż raz w tygodniu, natomiast 0,3 % nie korzysta wcale. [10] Wykorzystanie technologii informacyjnych w celu pozyskania specjalistycznych informacji jest zdecydowanie najczęstszym działaniem podejmowanym przez nauczycieli (93%), nieco mniejszą popularnością cieczy się zastosowanie TI do przygotowania prezentacji multimedialnej, czy pliku tekstowego (78%), najrzadziej zaś komputer oraz Internet wykorzystywane są w celu logowania na platformy e-learningowe (1,3%). [10]

Ważnym aspektem odnoszącym się do wykorzystania nowoczesnych technologii multimedialnych są kompetencje nauczycieli w zakresie technologii informatycznych. Jak wskazują wyniki badań [10] są one bardzo mocno zróżnicowane. Umiejętności dotyczące edycji tekstu, tworzenia prezentacji multimedialnych oraz obsługi arkuszy kalkulacyjnych plasują się na zadowalającym poziomie, natomiast pozostałe pozostawiają pole do dalszego podkoszenia kwalifikacji i kompetencji.

Mając na uwadze efektywność edukacji z wykorzystaniem środków mul timedialnych w odniesieniu do tradycyjnych form kształcenia dane literaturowe wskazują zdecydowanie na lepsze wyniki w przypadku zastosowania multimediów. B. Steinbrik [11] wykazał wzrost/spadek efektywności na następujących poziomach:

- skuteczność nauczania – wzrost o 56%,
- zrozumienie tematu – wzrost o 50-60%,
- nieporozumienie przy przekazywaniu wiedzy – spadek o 20-40%,
- oszczędność czasu – wzrost o 38-70%,
- tempo uczenia – wzrost o 60%,
- zakres przyswojonej wiedzy – wzrost o 25-50% .

Badania przeprowadzane na gruncie polskim potwierdzają powyższe zależności. Z badań G. Gulińskiej, w których porównywano wyniki nauczania z wykorzystaniem multimediów oraz wyniki nauczania z wykorzystaniem innych środków kształcenia wynika, iż przyrost wiedzy ogólnej po zastosowaniu multimediów wśród nauczycieli był większy o 5,4%, zaś wśród studentów o 34%. [12]

O ważności kwestii wykorzystywania multimediów w procesie kształcenia świadczą międzynarodowe raporty i wytyczne, zawierające założenia programowe edukacji multimedialnej w epoce globalizacji i społeczeństwa informacyjnego. UNESCO w przyjętej w 2007 roku agendzie wskazuje 12 zaleceń dotyczących rozwoju edukacji medialnej, wśród których znajdują się również postulaty dotyczące wprowadzenia edukacji medialnej do szkół, stosowania metod aktywizujących w nauczaniu, objęcia nauczycieli programami kształcenia w zakresie mediów, czy też zachęty do rozwoju akademickiej pedagogiki mediów. [13]

4. Podsumowanie

Nowoczesna edukacja powinna odbywać się z wykorzystaniem interaktywnego przekazu multimedialnego. Wyniki analizowanych badań wskazują na wysoką efektywność nowoczesnych technologii multimedialnych w odniesieniu do metod tradycyjnych. Obserwuje się rosnące zainteresowanie takimi technologiami, zarówno ze strony nauczycieli, jak i uczniów. Pozostawanie w próżni w tej kwestii powoduje spadek efektywności nauczania i brak zainteresowania zajęciami ze strony uczniów.

Do realizacji zadań szkoły w zakresie edukacji multimedialnej powinni być przygotowani, poprzez umiejętność odpowiedniego posługiwania się technologiami informatycznymi w pracy własnej oraz w pracy z uczniami, wszyscy nauczyciele. Komputer podłączony do Internetu, projektor multimedialny oraz tablica interaktywna to narzędzia, które w sposób niewyobrażalny zmieniły pracę nauczyciela. Od nauczyciela, jego znajomości, czasami podstawowych zagadnień informatycznych, zależy jakość i profesjonalizm tworzonych materiałów dydaktycznych. Jednak aby proces kształcenia odznaczał się wysoką efektywnością, nauczyciel wykorzystujący w swojej pracy nowoczesne środki dydaktyczne przed ich zastosowaniem musi przeanalizować zalety i wady poszczególnych rozwiązań i wybrać rozwiązanie najbardziej korzystne z punktu widzenia przekazywanych treści.

Literatura

1. Szulżyk-Cieplak J., Lenik K., Pietroń D. *Kształcenie na kierunku edukacja techniczno-informatyczna w dobie społeczeństwa informacyjnego*, W: Aspekty wizualne w edukacji szkolnej i akademickiej, pod red. H. Rarot, M. Śniadkowski, Politechnika Lubelska, Lublin, 2016.
2. Bednarek J. *Multimedia w kształceniu*, Wydawnictwo PWN, Warszawa, 2006.
3. Strykowski W. *Media i edukacja*, Edukacja medialna, nr 1, 1996.
4. Siemieniecki B. *Technologia informacyjna w edukacji*, W: Edukacja multimedialna, pod red. J. Gajdy, S. Juszczyka, B. Siemienieckiego, K. Wenty. Wydawnictwo Adam Marszałek, Toruń, 2005.
5. Wenta K. *Metodyka wykorzystania technologii informacyjnej w edukacji medialnej*, W: Edukacja multimedialna, pod red. J. Gajdy, S. Juszczyka, B. Siemienieckiego, K. Wenty. Wydawnictwo Adam Marszałek, Toruń, 2005.
6. Lenik K., Gawrylak B. *Dobór form prezentacji multimedialnych w dydaktyce szkoły wyższej na przykładzie problematyki dystrybucji wyrobów*, W: Informatyka w kształceniu, pod red. K. Lenika, G. Borowskiego, t. I. Lubelskie Towarzysko Naukowe, Lublin ,2006.
7. Walter N. *Obszary edukacyjnych zastosowań Internetu,* Studia edukacyjne, nr 23, Poznań, 2012.
8. Lipiński M. *Multimedia strumieniowe w kształceniu informatycznym w świetle badań sondażowych*, praca magisterska, Politechnika Lubelska, Lublin, 2011, [maszynopis niepublikowany].
9. Abelite, M. *Pobudzanie aktywności twórczej uczniów szkoły podstawowej poprzez multimedia*, Szczecin, Uniwersytet Szczeciński, 2013.
10. Mikołajczyk, K., Pietraszek K. *Czy nauczyciele wykorzystują nowoczesne technologie informacyjno-komunikacyjne w kształceniu?*, Raport z badań, Centrum Rozwoju Edukacji Niestacjonarnej SGH, 2013.

11. Steinbrink B. *Multimedia u progu technologii XXI wieku,* Wydawnictwo Robomatic, Wrocław, 1993.
12. Gulińska H. *Strategie multimedialne kształcenia chemicznego,* Wydawnictwo Naukowe UAM, Poznań, 1997.
13. *Paris Agenda or 12 Recommendations for Media Education,* Paris, UNESCO, 2007. Dostęp: http://www.diplomatie.gouv.fr (17 marca 2017 r.)

Charakterystyka nowoczesnych technologii multimedialnych w aspekcie efektywnego wykorzystania w procesie nauczania

Streszczenie
W niniejszym artykule scharakteryzowano nowoczesne technologie multimedialne w zakresie wybranych form ich prezentacji oraz wybranych narzędzi ułatwiających ich wykorzystanie w procesie nauczania. Przedstawiono również powody wykorzystania technologii multimedialnych w edukacji, a także opisano metody efektywnego ich zastosowania. Dokonano również analizy badań poziomu zainteresowania technologiami multimedialnymi wśród nauczycieli oraz uczniów, a także analizy poziomu umiejętności związanych z technologiami informacyjnymi wśród nauczycieli.
Słowa kluczowe: technologie multimedialne, efektywne nauczanie, proces nauczania

Characteristic of modern multimedia technologies in terms of effective use in the teaching process

Summary
This article describes modern multimedia technologies in selected forms of their presentation and selected tools to facilitate their use in the teaching process. The reasons for the use of multimedia technologies in education are also presented, as well as the methods of their effective application. The study also investigated the level of interest in multimedia technologies among teachers and students, as well as an analysis of the level of information technology skills among teachers.
Keywords: multimedia technologies, effective teaching, teaching process

Agata Płecha[1], Joanna Szulżyk-Cieplak[2]

Aktywizujące metody nauczania i ich wpływ na efektywność procesu kształcenia

1. Wprowadzenie

Metody aktywizujące wpływają na wzrost efektywności kształcenia, co potwierdza skuteczność dydaktyczną angażowania w ten proces osób uczących się. Już w latach 70-tych, Wincenty Okoń w swojej książce pt. „Nauczanie problemowe we współczesnej szkole" [1], promował hasło odejścia od biernej postawy uczniów biorących udział w procesie nauczania opartego na stosowaniu wyłącznie metod tradycyjnych, gdzie sposób myślenia zostawał narzucany, a gotowe wiadomości były podawane przez osobę prowadzącą zajęcia. Nowak i Sobieszczyk [2] w oparciu o przeprowadzone badania na grupie osób w wieku od 7 do 10 lat wykazały, że aktywizowanie ucznia poprzez zachęcanie go do konstruowania, interpretowania oraz indywidualnego przetwarzania rzeczywistości na podstawie wcześniej prowadzonych przez niego obserwacji i doświadczeń, skutkuje podniesieniem efektywności w zakresie poziomu wiedzy, stopnia jej zrozumienia oraz umiejętności odpowiedniego wykorzystania w konkretnej sytuacji.

Aktywizowanie osób uczących się może odbywać się m.in. z wykorzystaniem technologii informacyjnych, których szybki rozwój na przestrzeni ostatnich lat sprawił, że są one w coraz szerszym zakresie wykorzystywane w placówkach oświatowo-wychowawczych. Użycie komputera nie sprowadza się obecnie jedynie do prowadzenia lekcji informatyki, czy zajęć komputerowych. Nauczyciele coraz chętniej sięgają po niego w celu usprawnienia przyswajania wiedzy przez uczniów z różnych przedmiotów, zarówno humanistycznych jak i ścisłych. Zastosowanie wspomagania komputerowego umożliwia m.in. wizualizację danych oraz aktywny udział uczniów w prowadzonych zajęciach poprzez praktyczne działanie [3].

Mając na uwadze wspomniane wyżej przesłania oraz korzyści płynące ze stosowania aktywizujących metod nauczania, w pracy podjęto próbę zbadania jak interaktywne pomoce dydaktyczne wpływają na efektywność procesu kształcenia. W tym celu wykorzystano autorską aplikację internetową wspomagającą nauczenie wybranych zagadnień matematyki w szkole ponadgimnazjalnej.

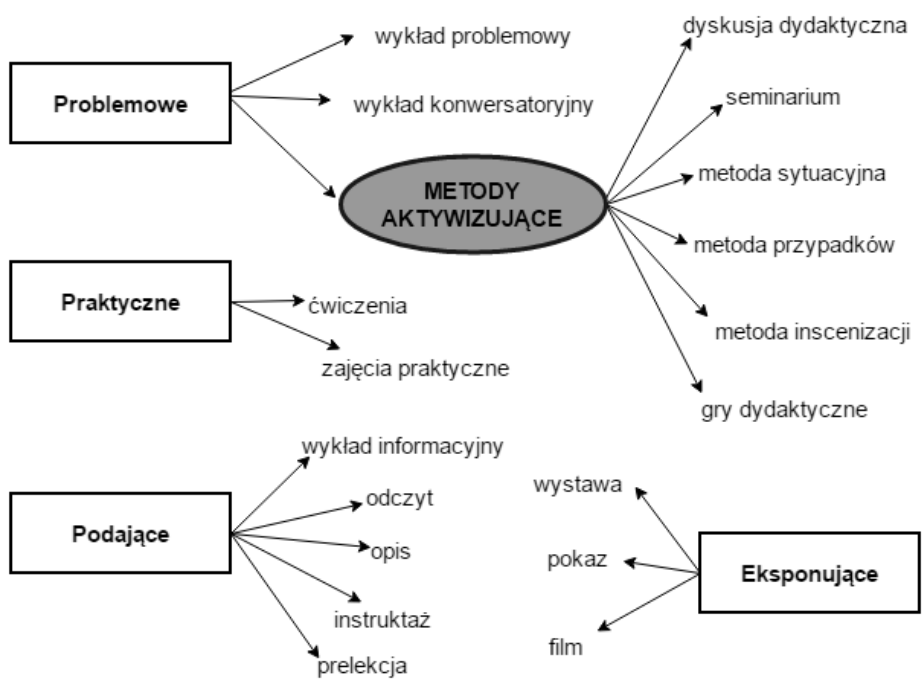
2. Metody aktywizujące w nauczaniu

Metody aktywizujące, występujące obok wykładu problemowego i konwersatoryjnego (rys. 1), wywodzą się z grupy metod problemowych, korzystnie wpływających na wszechstronny rozwój osobowości uczniów. Charakteryzują się one tym, że

[1] agata.plecha@wp.pl; Wydział Podstaw Techniki, Politechnika Lubelska, www.pollub.pl
[2] j.szulzyk-cieplak@pollub.pl; Katedra Podstaw Techniki, Wydział Podstaw Techniki, Politechnika Lubelska, www.pollub.pl

pozwalają na doskonalenie umiejętności poznawczych, kształtują zdolności obserwacyjne oraz sprzyjają procesom samodzielnego myślenia. Głównym celem stosowania tych metod jest zwrócenie szczególnej uwagi na rozbudzenie aktywności i zainteresowań ucznia, podniesienie poziomu atrakcyjności przyswajania omawianych treści, co pozwala na uniknięcie wkradania się monotonii w trakcie trwania procesu nauczania-uczenia się. Skłaniają one do samodzielnego rozwiązywania problemów stawianych przez nauczyciela podczas prowadzonych zajęć dydaktycznych, które nawiązują do znanych już wiadomości, doświadczeń, czy przeżyć [4].

Zaangażowanie w naukę z zastosowaniem metod aktywizujących odbywa się poprzez kierowanie się celem bezpośrednio związanym z wykonaniem zadania, wyzwalającym pozytywną motywację do działania oraz zachęcającym do podejmowania czynności prowadzących do jego rozwiązania. Wykonanie zadania stanowi nagrodę wzmagającą potrzebę podejmowania nowych wyzwań. Motywacją do osiągnięcia wspomnianego celu może być chęć zaspokojenia własnej ciekawości, zdobywania wiadomości przydatnych w życiu codziennym bądź nauka dla osobistej satysfakcji. O tym jakie cele stawia sobie uczeń w procesie nauczania oraz w jakim stopniu stara się je realizować jest uzależnione od różnorodnych czynników zarówno zewnętrznych jak i wewnętrznych wśród których szczególnie liczą się cechy osobowości traktowane jako umiejętność świadomego kierowania własnym postępowaniem [1].



Rysunek 32. Podział metod dydaktycznych – klasyfikacja według Wincentego Okonia, opracowanie własne na podstawie [5]

2.1. Rodzaje metod aktywizujących

W skład metod aktywizujących wchodzą dyskusje dydaktyczne, seminaria, a także metody takie jak: sytuacyjna, przypadków, inscenizacji oraz gier dydaktycznych (rys. 1).Dyskusja dydaktyczna polega na wymianie informacji oraz konfrontacji wysnutych na ich podstawie opinii z nauczycielem, aby w ten sposób opanować podstawowe wiadomości związane z poruszanym tematem. Charakteryzuje się ona tym, że uczestników należy wcześniej odpowiednio do niej przygotować poprzez wydanie poleceń polegających między innymi na zaznajomieniu się z materiałami, aby mogli się oni wykazać bogatą wiedzą dotyczącą omawianego zagadnienia. Dyskusja stanowi ważną umiejętność społeczną, która jest wykorzystywana podczas rozwiązywania problemów intelektualnych. Służy do poszukiwania rozwiązania oraz drogi postępowania podczas występowania różnorodnych poglądów [5].

Seminarium to metoda pozwalająca na zgłębianie wiedzy uczestniczących w nim osób z określonej dziedziny, którzy sami przygotowują jego przebieg dzięki zaznajomieniu się z literaturą związaną z omawianym tematem. Podstawowe problemy pojawiające się podczas trwania seminarium zostają rozwiązane przez prowadzących w oparciu o analizę własnego życia zawodowego lub społecznego. Metoda ta wykorzystuje dyskusję dydaktyczną, którą kieruje nauczyciel czuwający nad jej przebiegiem. Skuteczność seminarium zależy od tego w jakim stopniu jego tematyka jest związana z doświadczeniami uczestników [4].

Metoda sytuacyjna umożliwia doskonalenie umiejętności związanych z analizowaniem podanych zagadnień oraz podejmowaniem właściwych decyzji w konkretnych sytuacjach przedstawionych jako opis sytuacyjny. Tworząc daną sytuację na potrzeby tej metody powinno się mieć na uwadze to, aby była jak najbardziej zbliżona do rzeczywistości. Ma ona w obiektywny sposób prezentować nieoczywisty i prawdziwy problem, którego ocena, analiza oraz wybór optymalnego rozwiązania należy do uczestników, indywidualnie zgłaszających swoje propozycje. Nauczyciel jako osoba nadzorująca proces uczenia się z wykorzystaniem metody sytuacyjnej ma za zadanie przygotować plan oraz cel zajęć [4].

Metoda przypadków zaliczana jest do teoretycznych sposobów nauczania, z wyłączeniem przypadków, gdzie należy rozstrzygnąć faktyczną sytuację, która miała miejsce w konkretnej instytucji, dlatego też stanowi ona idealną sposobność do wykorzystania zdobytej wiedzy teoretycznej w działaniu praktycznym. Jej istotą jest przedstawienie zwięzłego opisu zdarzenia z trudno wykrywalnym rozwiązaniem, aby uczeń mógł zadając pytania nauczycielowi wyjaśniać wątpliwe sobie kwestie. Dzięki tej metodzie uczeń rozwija myślenie analityczne, syntetyczne oraz ekonomiczne, a także kształtuje zdolności podejmowania decyzji oraz gotowości stawiania czoła trudnym problemom [5].

Metoda inscenizacji służy kształtowaniu odpowiednich zachowań i wyrabianiu prawidłowych postaw w razie wystąpienia sytuacji problemowej poprzez odgrywanie ról przydzielonych przez prowadzącego w oparciu o przygotowany scenariusz. Aktywność ucznia przejawia się tutaj bezpośrednim zaangażowaniem go w dane zdarzenie, a podejmowane przez niego decyzje wynikają z posiadanej wiedzy jak

również z pobudek emocjonalnych. Po zakończeniu inscenizacji ma miejsce dyskusja zapoczątkowana przez widownię, oceniającą zaprezentowany problem. Do zadań nauczyciela należy tutaj jedynie sporządzenie scenariusza, rozdzielenie ról i przygotowanie uczniów do inscenizacji, minimalne kierowanie zaistniałą dyskusją, a także jej podsumowanie [4].

Gry dydaktyczne stanowią rodzaj metod nauczania, w których pewne treści przyswajane są przez uczniów drogą zabawy. Gra umożliwiająca efektywne uczenie się powinna posiadać przemyślaną i celowo zorganizowaną sytuację dydaktyczną. Aktywizowanie uczniów odbywa się tutaj poprzez wymuszanie interakcji oraz rywalizacji między pozostałymi uczestnikami rozgrywki. Różne gry spełniają różnorodne funkcje, najczęściej stosowane rodzaje to: gry problemowe oparte na analizie sytuacji problemowych, gry sprawdzające służące do weryfikowania podejmowanych decyzji, gry problemowo-dyskusyjne pozwalające na wymianę poglądów na temat przedstawionego problemu, gry dyskusyjne konfrontujące własne strategie w werbalny sposób, gry symulacyjne przedstawiające świat modelowy oraz gry komputerowe [5].

2.2. Stosowanie metod aktywizujących z wykorzystaniem technologii informacyjnych

Tworzenie interaktywnych materiałów dydaktycznych z wykorzystaniem różnorodnych technologii informacyjnych generuje szereg korzyści związanych z szansą oddziaływania na różne zmysły odbiorcy, a tym samym z pobudzeniem jego motywacji, utrzymaniem dłuższego okresu koncentracji oraz skróceniem czasu nauki wywołanego lepszym zrozumieniem materiału o wzbogaconym przekazie treści [6]. Taki sposób zdobywania wiedzy wymaga posiadania odpowiedniego sprzętu oraz dłuższych przygotowań do zajęć, jednak warto go stosować, ponieważ wzbogacenie procesu kształcenia takimi formami, które będą aktywizować ucznia bywa dużo ciekawsze i skuteczniejsze od stosowania tylko i wyłącznie metod tradycyjnych.

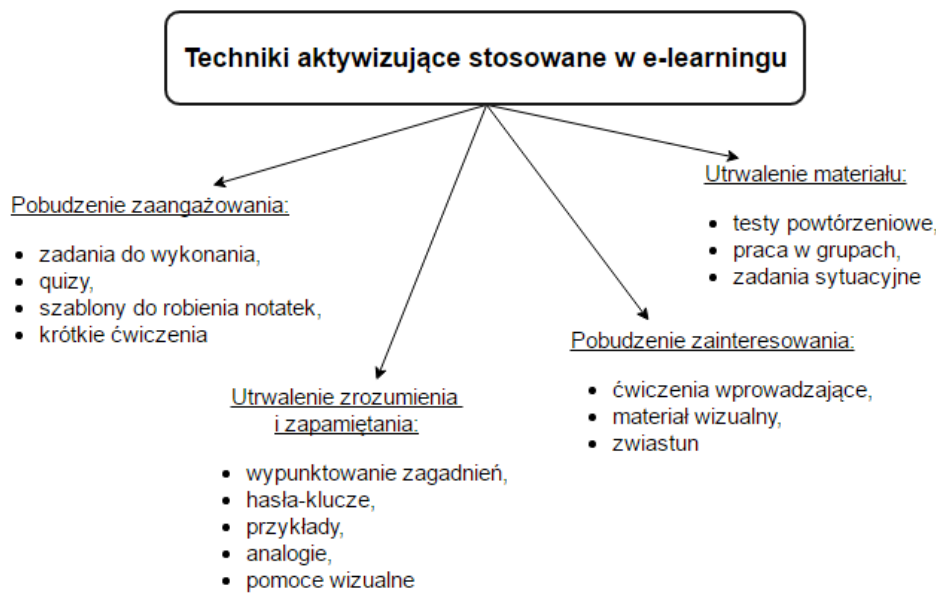
Popularna w obecnych czasach e-edukacja, realizowana z wykorzystaniem technologii informacyjnych, jako interaktywny sposób kształcenia także charakteryzuje się czynnym udziałem oraz zaangażowaniem jej odbiorców. W takim podejściu metody aktywizujące stosuje się w odniesieniu do omawianych treści, sposobu ich prezentowania oraz szybkości przyswajania. Te trzy obszary wykorzystywane w e-learningu pozwalają na:

- prezentowanie jedynie takiej ilości informacji, jaka będzie potrzebna do rozwiązania określonego problemu, z możliwością zakończenia nauki przez uczącego się w dowolnej chwili;
- dostosowanie treści do indywidualnych potrzeb ucznia, który posiada możliwość wyboru interesujących go w danym czasie zagadnień;
- dostarczenie wiedzy i umiejętności w odpowiednim momencie, czyli podczas wykonywania danego zadania, przy czym materiały dydaktyczne są dostępne w chwili odpowiadającej użytkownikowi.

Zgodnie z powyższymi informacjami e-szkolenia wyróżniają się ograniczoną ilością materiału nauczania, który przyswajany jest zróżnicowanymi sposobami.

Kształtują one światopogląd oraz motywacje poprzez rozwijanie praktycznych umiejętności oraz zrozumienie poszczególnych zagadnień i koncepcji. Stwarzają możliwości do pracy grupowej dzięki czemu istnieje szansa na wykorzystanie wiedzy osób uczących się. Zaistniałe problemy przeznaczone do rozwiązywania nawiązują do znanych już teorii. Takie podejście sprzyja planowaniu czynności, które będą realizowane po szkoleniu [7].

Istnieje wiele technologii oraz narzędzi wspomagających metody aktywizujące, które są stosowane w e-edukacji. Są to różnego rodzaju serwisy oraz platformy umożliwiające tworzenie interaktywnych pomocy dydaktycznych, wspomagających pobudzanie zaangażowania, zainteresowania, zrozumienia, zapamiętania oraz utrwalenia materiału przez uczniów. Przykłady takich pomocy pobudzających zaangażowanie uczestników w naukę zostały przedstawione na rysunku 2.



Rysunek 33. Techniki aktywizujące stosowane w e-learningu, opracowanie własne na podstawie [6]

2.3. Charakterystyka wybranych narzędzi do tworzenia interaktywnych materiałów dydaktycznych

Zanim uczniowie będą mogli korzystać z dóbr jakie niesie ze sobą wykorzystywanie technologii informacyjnych w edukacji, ktoś musi zająć się odpowiednim ich zastosowaniem w procesie kształcenia. W Internecie dostępnych jest mnóstwo narzędzi ułatwiających nauczycielom tworzenie interaktywnych materiałów dydaktycznych wspomagających proces nauczania. Ten podrozdział został poświęcony omówieniu wybranych sposobów pozwalających na przygotowanie takich materiałów z wykorzystaniem zasobów udostępnionych w sieci Internet.

2.3.1. Hot Potatoes

Hot Potatoes jest darmowym programem udostępnionym dla środowiska edukacyjnego, umożliwiającym szybkie i łatwe tworzenie interaktywnych materiałów, które można między innymi eksportować do druku bądź zamieszczać na wybranych portalach lub stronach internetowych. W skład niniejszego programu wchodzi sześć narzędzi pozwalających na przygotowywanie testów, zadań polegających na wypełnianiu luk umieszczonych w tekście lub ułożeniu rozsypanych liter, ćwiczeń związanych z przyporządkowywaniem oraz krzyżówek o różnej wielkości siatki. Ostatnie narzędzie umożliwia łączenie wszystkich utworzonych dotychczas zadań. Program Hot Potatoes umożliwia dokonywanie obliczeń procentu poprawnie udzielonych odpowiedzi w zadaniach przez niego utworzonych. Takie rozwiązanie sprzyja łatwemu wystawianiu ocen użytkownikom, którzy je rozwiązują [8].

2.3.2. LearningApps.org

LearningApps.org to wspomagająca edukację aplikacja zawierająca małe interaktywne moduły, pozwalające na opracowanie zadań aktywizujących osoby je rozwiązujące. Charakteryzuje się ona tym, że większość zawartych w niej treści została przygotowana przez osoby z niej korzystające, które są ograniczone jedynie wcześniej wspomnianymi modułami. Aplikacja umożliwia korzystanie z gotowych rozwiązań wygenerowanych przez innych użytkowników jak również opracowanie własnych materiałów dydaktycznych. Gotowe wzory zamieszczone na stronie zostały posortowane według poziomów kształcenia. Oferowane są tutaj: edukacja wczesnoszkolna, kształcenie podstawowe, gimnazjalne, ponadgimnazjalne oraz zawodowe i ustawiczne. Ponadto dostępny jest podział według kategorii tematycznych zamieszczonych w ramach wybranego etapu kształcenia [9].

2.3.3. mInstructor

mInstructor to bezpłatne narzędzie umożliwiające przygotowanie oraz udostępnianie samodzielnie wygenerowanych interaktywnych materiałów dydaktycznych. Portal ten posiada bogaty zasób lekcji opracowanych przez rzeszę nauczycieli, z których można korzystać wyłącznie na prywatny, niekomercyjny użytek. Możliwość przygotowywania własnych materiałów jak i przeglądanie już istniejących wymaga tutaj od użytkownika wcześniejszej rejestracji. Po zalogowaniu do portalu mIntructor w zakładce „Repozytorium" dostępne są wszystkie dotychczas wygenerowane lekcje. Zostały one pogrupowane według następujących etapów kształcenia: szkoła podstawowa klasy 1-3 i 4-6, gimnazjum, szkoła ponadgimnazjalna, a także szkoła językowa. Wybranie interesującej kategorii skutkuje wyświetleniem kolejnego podziału, tym razem na przedmioty. Tak posortowane materiały można przeglądać oraz zapisywać do swoich zasobów w celu ich dalszego wykorzystania [10].Zaprezentowane wyżej narzędzia służące do opracowywania interaktywnych materiałów dydaktycznych zostały wybrane ze względu na to, że są całkowicie darmowe, a ponadto powszechnie występujące w zasobach sieciowych. Omówione w podrozdziale narzędzia są dostępne w języku polskim, co znacznie ułatwia obsługę, a ich stosowanie nie wymaga od nauczycieli dużego doświadczenia z zakresu informatyki.

3. Wpływ interaktywnych pomocy dydaktycznych na poszerzenie poziomu wiedzy – badania własne

3.1. Cel, przedmiot i metoda badań

Przeprowadzone badania miały charakter diagnostyczny. Ich celem było określenie wpływu interaktywnych pomocy dydaktycznych na proces kształcenia. Do badań wykorzystano autorską aplikację webową wspomagającą nauczanie wybranych zagadnień z matematyki. W pracy podjęto próbę udzielenia odpowiedzi na pytanie: jak treści prezentowane z wykorzystaniem aktywizujących materiałów edukacyjnych wpływają na poszerzenie poziomu wiedzy uczniów szkół ponadgimnazjalnych z zakresu funkcji i ich własności?

W trakcie prowadzenia badań podstawowym założeniem było stwierdzenie, że uczestnicy badania korzystający z interaktywnej aplikacji internetowej posiadają podstawową wiedzę z zakresu prezentowanych treści programowych. Wiadomości te uczniowie mieli okazję zdobyć na zajęciach matematyki, prowadzonych w ramach realizacji zagadnień zawartych w podstawie programowej obowiązującej w szkole średniej. Proponowana aplikacja miała jedynie wzbogacić wiadomości przyswojone metodami tradycyjnymi.

W badaniach wykorzystano metodę ilościową, monitorującą wpływ czynnika niezależnego na efekt pomiaru zmiennej zależnej. Konieczne było zbadanie poziomu wiedzy uczniów przed i po zapoznaniu się z treściami prezentowanymi przez interaktywną aplikację. Takie podejście wymagało przeprowadzenia co najmniej dwukrotnego pomiaru dlatego też przygotowano dwa testy:

- pre-test sprawdzający stan wiedzy uczniów przed rozpoczęciem korzystania z interaktywnej pomocy dydaktycznej,
- post-test weryfikujący wiedzę osób, które zapoznały się z treściami zawartymi w przygotowanej aplikacji.

3.2. Organizacja badań

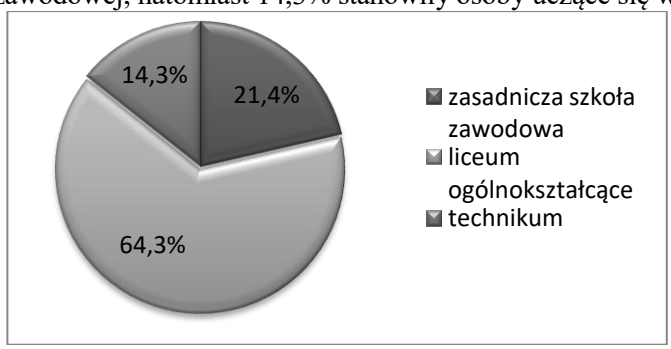
Opracowana aplikacja stanowiąca interaktywną pomoc dydaktyczną została udostępniona na stronie internetowej dzięki skorzystaniu z darmowych usług hostingowych. Testy jednokrotnego wyboru sprawdzające poziom wiedzy uczniów przygotowano z wykorzystaniem formularzy Google. W każdym z nich zawarto 10 pytań dotyczących treści prezentowanych przez aplikację. Za każdą poprawną odpowiedź testowani otrzymywali jeden punkt (maksymalnie można było uzyskać 10 punktów z każdego testu).

Linki odsyłające do pre-testu, aplikacji oraz post-testu zostały rozpowszechnione na portalu społecznościowym – Facebook. Weryfikacja uczniów biorących udział w badaniu miała miejsce w pierwszym etapie rozwiązywania testu, gdzie testowany musiał udzielić odpowiedzi na pytania związane z płcią, typem szkoły ponadgimnazjalnej oraz klasą, do której aktualnie uczęszcza.

3.3. Analiza badań

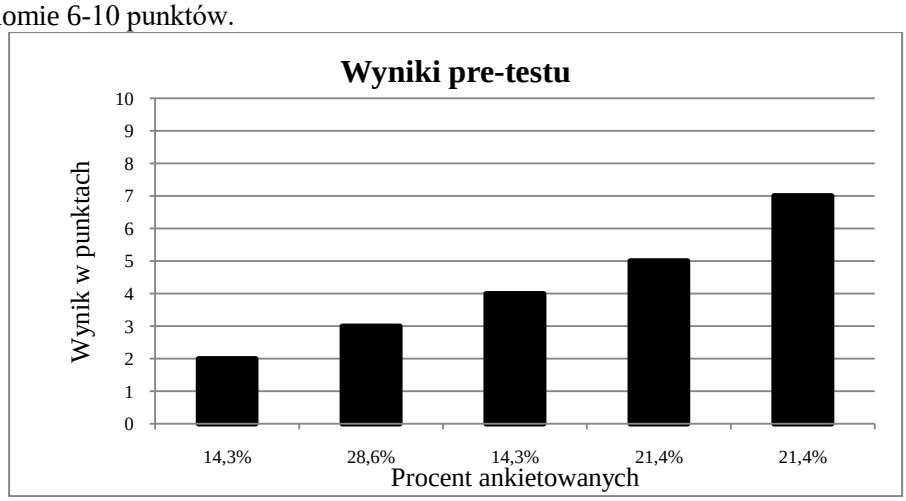
W badaniu wzięła udział grupa uczniów uczęszczających do różnych typów szkół ponadgimnazjalnych. Uczestnicy badania rozwiązali pre-test, zapoznali się z treściami prezentowanymi przez aplikację oraz udzielili odpowiedzi na pytania zawarte w post-teście.

Na rysunku 3 przedstawiono procentowy udział uczniów uczących się w poszczególnych typach szkół ponadgimnazjalnych. Najwięcej osób, bo aż 64,3% ankietowanych uczęszcza do liceum ogólnokształcącego, 21,4% to uczniowie zasadniczej szkoły zawodowej, natomiast 14,3% stanowiły osoby uczące się w technikum.



Rysunek 34. Procentowy udział badanych uczniów według typu szkoły ponadgimnazjalnej, do której uczęszczają

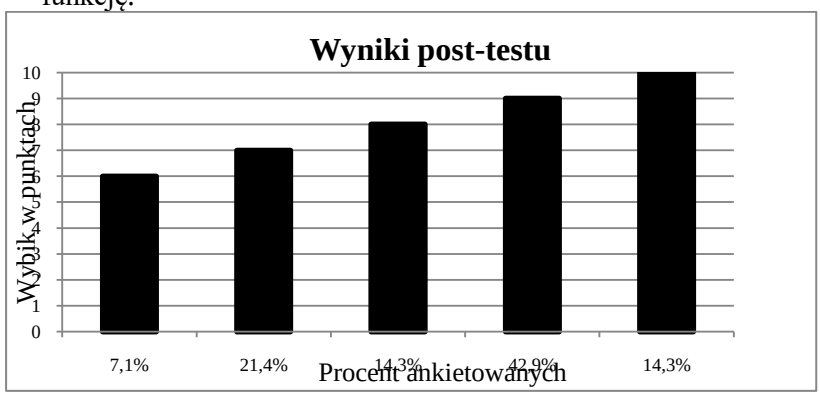
Na wykresach (rys. 4 i 5) przedstawiono rozkład uzyskanych przez uczniów punktów przed i po zapoznaniu się z treściami prezentowanymi przez aplikację. Uzyskane wyniki wskazują, że zastosowane narzędzie dydaktyczne w postaci interaktywnej aplikacji webowej przyczyniło się do średniej poprawy stanu wiedzy testowanych osób z 42,9% do 83,6%. Zakres punktów uzyskiwanych po rozwiązaniu pre-testu wahał się między 2-7, natomiast post-test uczniowie kończyli z wynikiem na poziomie 6-10 punktów.



Rysunek 35. Rozkład wszystkich punktów uzyskanych przez uczniów po rozwiązaniu pre-testu

W trakcie rozwiązywania pre-testu uczniowie największy problem mieli podczas udzielania odpowiedzi na pytania związane z:

- definicją dziedziny funkcji f;
- określeniem, czy każdy zbiór punktów w układzie współrzędnych jest wykresem funkcji;
- odnalezieniem miejsc zerowych funkcji;
- odczytaniem wartości największej i najmniejszej osiąganej przez daną funkcję.



Rysunek 36. Rozkład wszystkich punktów uzyskanych przez uczniów po rozwiązaniu post-testu

W trakcie rozwiązywania post-testu wciąż utrzymywał się problem z odczytaniem największej oraz najmniejszej osiąganej przez funkcję wartości, do czego mógł przyczynić się podchwytliwy sposób sformułowania dostępnych odpowiedzi.

Zgodnie z danymi zaprezentowanymi na rysunkach 4 i 5 żaden z ankietowanych nie otrzymał mniejszej ilości punktów z post-testu w porównaniu do wyników uzyskanych z pre-testu. Żaden z post-testów nie został ukończony z wynikiem poniżej 60%. Rozbieżność osiąganych rezultatów może wynikać z ilości czasu poświęconego na zapoznanie się z treściami prezentowanymi przez zaprojektowaną aplikację internetową.

3.4. Wskaźnik przyrostu wiedzy (WPW)

Zakładając, że treści prezentowane przez aplikacje są sposobem usuwania różnic między aktualnym, a oczekiwanym stanem wiedzy z zakresu funkcji i ich własności, możliwe jest wyznaczenie wskaźnika przyrostu wiedzy (WPW), który obliczono w następujący sposób [11]:

$$WPW = \frac{(wynik\ posttestu - wynik\ pretestu)}{(wyniki\ możliwe\ do\ osiągnięcia - wynik\ pretestu)} \cdot 100\%$$

Z wykonanych obliczeń wynika, że średni wskaźnik przyrostu wiedzy wyniósł w przybliżeniu 66%.

$$WPW_{śr} \approx 66{,}1\%$$

4. Podsumowanie

Zarówno stosowanie metod aktywizujących w sposób tradycyjny, jak i z wykorzystaniem technologii informacyjnych umożliwia efektywny rozwój osób uczestniczących w procesie kształcenia poprzez angażowanie uczniów oraz rozbudzenie ich ciekawości poznawczej, co skutkuje chęcią jej zaspokojenia. Nauczanie metodami tradycyjnymi, które zostają wzbogacone o aspekty aktywizujące jednostki uczącej się ograniczają ryzyko wkradania się monotonii, urozmaicając proces kształcenia o efekty wizualne, czy interaktywne.

Przeprowadzone badania wykazały, że interaktywne pomoce dydaktyczne efektywne wpływają na poszerzenie stanu wiedzy osób z nich korzystających. Średni wskaźnik przyrostu wiedzy ankietowanych uczniów wyniósł 66%, a żaden z post-testów nie został rozwiązany poniżej 60%, natomiast około 14% testowanych uzyskało maksymalną ilość punktów.

Stosowanie metod aktywizujących wiąże się jednak z pewnymi problemami takimi jak m.in. potrzeba większej ilości czasu w trakcie ich wdrażania, czy wymóg odpowiedniego sprzętu, jeżeli chodzi o ich stosowanie z wykorzystaniem technologii informacyjnych. Fakt ten powoduje, że omówione metody w niektórych placówkach oświatowych wykorzystywane są bardzo rzadko więc warto opracowywać interaktywne materiały, które mogłyby być wykorzystywane przez uczniów podczas zajęć pozalekcyjnych.

Literatura

1. Okoń W.*Nauczanie problemowe we współczesnej szkole*, Wydawnictwo Szkolne i Pedagogiczne, Warszawa 1975.
2. Nowak J., Sobieszczyk M., *Metody aktywizujące a efektywność kształcenia na etapie wczesnoszkolnym*, [w:] Siwek H. (red.): Efektywność kształcenia zintegrowanego. Implikacje dla teorii i praktyki, Katowice 2007, s. 234-241, [online] [dostęp 15.02.2017], dostępne w Internecie:
3. http://www.ukw.edu.pl/pracownicy/plik/jolanta_nowak/2161.
4. Szulżyk-Cieplak J., Lenik K., Pietroń D., *Kształcenie na kierunku edukacja techniczno-informatyczna w dobie społeczeństwa informacyjnego*, Politechnika Lubelska 2016.
5. Półturzycki J.,*Dydaktyka dla nauczycieli*, Wydawnictwo Adam Marszałek, Toruń 1999.
6. Plewka C. *Metodyka nauczania teoretycznych przedmiotów zawodowych*, Instytut Technologii i Eksploatacji, Radom 1999.
7. Adamkiewicz J. *Nowe technologie informacyjne w edukacji*, Wydawnictwo Adam Marszałek, Toruń 2015, s. 132-147.
8. Abramek E.,Piesiur T., Słaboń M., Żytniewski M., *Metody aktywizujące stosowane w dydaktyce na przykładzie wybranych narzędzi wykorzystywanych w wirtualnej edukacji*, Prace Naukowe/Akademia Ekonomiczna w Katowicach 2005, [online] [dostęp 15.02.2017], dostępne w Internecie:
9. http://www.swo.ae.katowice.pl/_pdf/185.pdf.

10. *Program Hot Potatoes*, [online] [dostęp 17.02.2017], dostępne w Internecie: https://hotpot.uvic.ca.
11. *Aplikacja LearningApps*, [online] [dostęp 17.02.2017], dostępne w Internecie: http://learningapps.org.
12. *Portal mInstructor*, [online] [dostęp 17.02.2017], dostępne w Internecie: https://www.minstructor.pl.
13. Skulmowski M. *Ocena efektywności kształcenia na przykładzie studiów podyplomowych z rachunkowości realizowanych na Wydziale Ekonomicznym Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie*, Folia Pomeranae Universitatis Technologiae Stetinensis 2013, Oeconomica 303(72), s. 198–199.

Aktywizujące metody nauczania i ich wpływ na efektywność procesu kształcenia

Streszczenie

Tematykę artykułu stanowi omówienie aktywizujących metod nauczania stosowanych w procesie kształcenia. Szczególną uwagę zwrócono tutaj na aktywizowanie uczniów z wykorzystaniem technologii informacyjnych. Przeprowadzone badania dotyczące wpływu metod aktywizujących na proces nauczania-uczenia się potwierdziły, że wykorzystywanie interaktywnych pomocy dydaktycznych skutkuje podniesieniem efektywności w zakresie poziomu wiedzy, stopnia jej zrozumienia oraz umiejętności odpowiedniego wykorzystania w konkretnej sytuacji.

Słowa kluczowe: aktywizujące metody nauczania, aktywny uczeń, edukacja

Activating teaching methods and their influence on efficiency of the education process

Summary

The subject of this article consists of discussion of activating teaching methods used in the education process. Particular attention has been paid to activating learner with using information technology. Conducted research proved that the interactive teaching aids raise efficiency of the level of knowledge, degree of understanding it and the ability to use it appropriately in a particular situation.

Keywords: activating teaching methods, activating learner, education

Monika Bogdanowska[1]

Interdyscyplinarny Słownik Wielojęzyczny on-line – możliwości i ograniczenia

1. Wstęp

Poszukiwanie koherencji terminologicznej w najróżniejszych obszarach ludzkiej aktywności jest zadaniem nienowym i towarzyszy nauce, rzec można, od zarania dziejów. Jednak dopiero dziś uzyskujemy dostęp do niemal nieograniczonych zasobów informacji, stąd też narastająca potrzeba z jednej strony objaśniana pojęć funkcjonujących w różnych językach, z drugiej ich porządkowania. W przypadku nauki ma to szczególne znaczenie – przekaz musi być spójny i klarowny.

Technologie informatyczne pozwalają na tworzenie narzędzi służących gromadzeniu, porządkowaniu i udostępnianiu zasobów terminologicznych. Jest to jednak zadanie bardzo złożone, wymagające przyjęcia określonej metodologii, która dodatkowo uzależniona jest od konkretnej dyscypliny.

W artykule omówiono wyniki i wnioski ukończonego projektu badawczego, który poświęcony był między innymi wskazanym wyżej zagadnieniom. Punktem wyjścia do badań była terminologia stosowana w szeroko rozumianym obszarze ochrony dziedzictwa kulturowego. W tym miejscu należy podkreślić, że zakres nauk wchodzących w pole zainteresowań tej dyscypliny jest niezwykle szeroki: od nauk humanistycznych, jak filozofia, prawo, czy historia sztuki, przez tradycyjne techniki malarstwa, czy budownictwa, po architekturę i nauki przyrodnicze: chemię, fizykę, biologię, czy petrografię. Do tego dochodzi ogromny zasób terminów związanych z historycznymi rzemiosłami. Jako dziedzina interdyscyplinarna, ochrona dziedzictwa kulturowego okazała się szczególnie cennym i inspirującym polem badawczym dla badań nad ekwiwalencją terminologii fachowej, a w konsekwencji stworzenia nowego typu słownika elektronicznego.

2. Stan badań

Konieczność tworzenia zasobów terminologii fachowej w różnych językach jest od dawna podnoszona, tak podczas konferencji, seminariów, jak i w raportach, czy podsumowaniach prac rozmaitych zespołów roboczych [1]. Punktem wyjścia dla podejmowanych prac są leksykony, glosariusze, terminologiczne słowniki jednotęzyczne lub też wspólne, wielojęzyczne opracowania wykonane przez autorów różnych narodowości. Poniżej przedstawię krótką charakterystykę ważniejszych dokonań w tym zakresie.

[1] mimero@interia.pl; Zakład Rysunku, Malarstwa i Rzeźby, Wydział Architektury, Politechnika Krakowska im. Tadeusza Kościuszki.

W pracach nad terminologią fachową dużą rolę odegrała możliwość budowania elektronicznych baz danych. W przypadku ochrony dziedzictwa kulturowego pierwsze bazy danych opierały się na systemach inwentaryzacji stosowanych w muzeach[2], ich uzupełnienie stanowiły zasoby pozyskiwane z tradycyjnych publikacji książkowych[3].

Zebranie listy terminów danego zagadnienia specjalistycznego, to pierwszy etap pracy. Należy wspomnieć, że często, wobec braku słowników, konieczne jest budowanie zasobu „od podstaw". Jest to zadanie niezwykle trudne i niejednokrotnie wymagające opracowania zupełnie odrębnej metodologii pracy. Systematyzacja terminów umożliwia w dalszej kolejności poszukiwanie ekwiwalentów w innych językach, a w konsekwencji tworzenie zasobów wielojęzycznych. Warto wspomnieć, że poza popularnymi językami jak angielski, niemiecki, czy hiszpański podejmowane są próby tworzenia leksykonów w językach mniej popularnych[4].

W przypadku hierarchicznego szeregowania terminów, zadanie okazało się bardzo złożone i wielowątkowe, niemniej jednak, jak pokazują istniejące bazy danych (np. AAT, HEREIN, CAMEO[5]), uwieńczone zostało sukcesem.

Na gruncie polskim zagadnieniami porządkowania terminologii fachowej zajmują się między innymi naukowcy związani z Uniwersytetem Wrocławskim, tworzący tezaurus terminów z historii sztuki i nauk pokrewnych [2]. W najnowszej publikacji poświęconej wykorzystaniu narzędzi informatycznych m.in. w digitalizacji zbiorów znajdziemy podsumowanie wielu zagadnień związanych z tym opracowaniem [3].

Inną strategię przyjęli twórcy projektu ARGOS, którego założeniem było stworzenie internetowej bazy danych terminologii konserwatorskiej podzielonej na rozdziały związane z zabiegami, technikami i technologią. W bazie tej każde hasło miało posiadać odpowiednią definicję, przełożoną na inne języki[6]. Szybko jednak

[2] Np. MIDAS Heritage - the UK Historic Environment Data Standard, [on-line] http://www.english-heritage.org.uk/publications/midas-heritage/, dostęp 2 03 2017

[3] By wymienić kilka ważniejszych źródeł: *Lexikon der Kunst. Architektur, Bildende Kunst, Angewandte Kunst, Industrieformgestaltung, Kunsttheorie*, wyd. Ludger A., Leipzig, Seeman 1968-1978; Koch W., Baustilkunde – Europäische Baukunst von der Antike bis zur Gegenwart, Orbis Verlag 1991; Glossarium Artis, wyd. Hubner R., Rieth R., Saur K.G., 1994; Dictionary of Building Preservation, wyd. Bucher W., A.I.A., C. Madrid, John Wiley & Sons Inc. 1996; Xarrié M., Glossary of Art Conservation, Vol I – III, Balaam, Barcelona 2006; Martínez L.R., Martínez Cabetas C., Diccionario Técnico Akal de Conservación y restauración de bienes culturales, Akal, Madrid 2003; z polskojęzycznych np. Kurzątkowski M., Mały słownik ochrony zabytków, Ministerstwo Kultury i Sztuki, Ośrodek Dokumentacji Zabytków, Warszawa 1989; *Słownik terminologiczny sztuk pięknych*, red. S. Kozakiewicz, PWN 1969.

[4] Na przykład prace nad tezaurusem chorwacko-włoskim podjęto na Uniwersytecie w Dubrowniku, a ICCROM opublikował glosariusz terminów arabskich odnoszących się do ochrony dziedzictwa kulturowego. Zob. Lucijana L., *Multilingual Multimedia Thesaurus for Conservation and Restoration – Collaborative Networked Model of Construction*, pdf; Mahdy H., *Glossary of Terms for the Conservation of Cultural Heritage in Arabic Alphabetical Order*, ICCROM, pdf.

[5] Euromed Heritage, HEREIN-European Heritage Network multilingual thesaurus, [on-line] http://www.euromedheritage.net/intern.cfm?menuID=8&submenuID=27&subsubmenuID=11; dostęp 10 02 2017; The Getty Research Institute, The Art and Architecture Thesaurus (AAT), [on-line] http://www.getty.edu/research/conducting_research/vocabularies/ dostęp: 10 02 2017; CAMEO: Conservation & Art Materials Encyclopedia Online, [on-line] http://cameo.mfa.org/wiki/Main_Page, dostęp: 5 03 2017.

[6] ARGOS, Art and Restoration Glossary Operating System, brak dostępu.

ujawnił się słaby punkt tego podejścia – otóż trudno jest o konsensus znaczeniowy w różnych obszarach kulturowych. Sytuacja ta dotyczy zresztą nie tylko definicji, ale też i samych terminów. Fakt ten ma niebagatelne znaczenie w pracach nad przekładem. W takiej sytuacji rozwiązania są dwa: albo uzgodnić definicje i terminy w oparciu o międzynarodowe gremia eksperckie, albo oprzeć się na jednojęzycznych, wyspecjalizowanych glosariuszach, jako formułach źródłowych, które mogą być następnie przekładane na inne języki. W przypadku pierwszego rozwiązania istnieją już bardzo dobre przykłady jego zastosowania, jak np. wielojęzyczny słownik zniszczeń kamienia, opracowany początkowo w wersji angielsko-francuskiej i stopniowo przekładany na inne języki[7]. Z kolei w 2015 roku ukończono prace nad pierwszym siedmiojęzycznym (!) słownikiem konserwacji malowideł ściennych EwaGloss[8], w którym międzynarodowa grupa ekspertów zebrała i opisała 250 terminów zgrupowanych w kilku rozdziałach. Ten model działania stosuje Komisja Europejska (CEN) przy ustalaniu norm (EN); w zakresie ochrony dziedzictwa przyjęto dotychczas dwadzieścia standardów[9].

Drugie rozwiązanie, a w więc budowanie „podręcznych" baz danych terminologicznych, wykorzystują twórcy specjalistycznych, internetowych glosariuszy różnych dziedzin powiązanych z ochroną dziedzictwa, np. Draper Conservation, czy Punchmarks[10] i wiele innych.

Ciekawym obszarem prac są jednojęzyczne opracowania, przygotowane przez międzynarodowe gremia w taki sposób, że umożliwiają przekład np. w oparciu o ilustracje, tabele czy wykresy. Tu warto wspomnieć o wieloautorskim dziele poświęconym cementom romańskim z glosariuszem w języku angielskim, zawierającym interesujący materiał ilustracyjny ułatwiającym zrozumienie i ewentualny dalszy przekład terminów [4].

Istnieje także wiele stron internetowych z których można uzyskać dostęp do zasobów tekstowych, takich jak prowadzona przez Amerykański Instytut Konserwacji strona AIC Wiki[11]. Nie powiodła się natomiast inicjatywa Presevapedia, gdzie teksty redagowane miały być podobnie jak w Wikipedii. Można przypuszczać, że taka

[7] ICOMOS-ISCS**:** *Illustrated glossary on stone deterioration patterns. Glossaire illustré sur les formes d'altération de la pierre*, pdf. Warsztaty podczas których uczestnicy z różnych krajów dyskutują i uzgadniają stosowaną terminologię organizuje także ICCROM np. *Terminology Used in Discussing Conservation Decisions, A contribution from the participants of SCD08 Course, Sharing Conservation Decisions ICCROM - International Course - Rome*, 3-28 November 2008, pdf.

[8] EwaGloss, [on-line] http://www.ewaglos.eu/; dostęp 3 03 2017

[9] English Heritage, [on-line] https://www.english-heritage.org.uk/professional/research/heritage-science/collections-conservation/centc346/ dostęp 10 02 2017; CEN, European Commeetee for Standarization, [on-line]: http://standards.cen.eu/dyn/www/f?p=204:32:0::::FSP_ORG_ID:411453&cs=11079A55D70F8377E3942E1C6704C7664; dostęp 10 02 2017

[10] Painting Conservation and Restoration Service, [on-line] http://www.draperconservation.com/glossary.htm , dostęp 10 02 2017); Punchmarks.net, [on-line] http://punchmarks.net/index.html dostęp 10 02 2017.

[11] AIC Wiki [on-line] http://www.conservation-wiki.com/wiki/Main_Page, dostęp 9 0 3 2017.

formuła pracy nie wzbudza większego zainteresowania specjalistów[12]. Większy zasięg, i zdecydowanie więcej materiałów, znaleźć można na stronie CoOL, gdzie zbierane są nadsyłane przez autorów opracowania poszczególnych tematów[13]. Nie są to jednak bazy danych, które możemy określić mianem baz „słownikowych" *per se*, jakkolwiek znajdziemy w nich liczne glosariusze różnych specjalistycznych zagadnień.

Dla polskiego odbiorcy szczególnie dotkliwy jest niedobór opracowań terminologii fachowej w języku polskim, nie mówiąc już o opracowaniach dwu- i więcej- językcznych. Na marginesie warto wspomnieć, że w powszechnie dostępnych w sieci ogólnych słownikach angielsko-polskich znajdziemy dwie dyscypliny z wyjątkowo obszernie reprezentowaną terminologią, są to prawo oraz medycyna.

3. Podsumowanie stanu badań i wytyczne nowego projektu

Dostępne w Internecie terminologiczne bazy danych są formalnie zróżnicowane. Mogą to być strony z listami haseł i ich krótkimi objaśnieniami, czasem są to wyszukiwarki. W tych ostatnich zazwyczaj znajdujemy pożądany termin wraz z objaśnieniem w języku wyszukiwania. Jeszcze inną propozycją są słowniki elektroniczne obsługiwane przez programy do edycji tekstu (np. w plikach pdf) – takimi leksykonami są na przykład są wspomniane „EwaGloss" i „Illustrated glossary on stone deterioration patterns".

Pod względem struktury można te zasoby podzielić następująco:

- struktura hierarchiczna: pozwala na wyświetlanie pojedynczych haseł,
- odpowiednio ulokowanych na gałęzi wraz z ekwiwalentami w innych językach i ewentualnie prostą definicją (np. AAT),
- baza danych podzielona tematycznie: pozwala na przeszukiwanie rozdziału i znajdowanie konkretnego hasła (czasem z ekwiwalentem w innym języku, definicją czy ilustracją (np. Plaster Architecture[14]),
- glosariusz w układzie alfabetycznym (np. Punchmarks),
- dokument w układzie tematycznym, z ekwiwalentami w innych językach, nieraz powiązany z ilustracją (np. EwaGloss).

Zakres tematyki poruszanej w opracowaniach jest bardzo zróżnicowany. W interesującym nas obszarze ochrony dziedzictwa kulturowego stosunkowo wiele baz danych zawiera listy różnych substancji stosowanych w wytwarzaniu dzieł sztuki (pigmenty, spoiwa), niektóre wzmiankują najważniejsze rodzaje zniszczeń, kluczowe zabiegi, podstawowe terminy teoretyczne. Pewną słabością jest powtarzalność haseł przy równoczesnym, niedoborze terminów wysokospecjalistycznych, niezwykle trudnych do znalezienia, zarówno w języku rodzimym, jak i w językach przekładu. Z drugiej jednak strony, czasem otrzymujemy nadmiar wyników, a zdarza się, że znajdowane hasła różnią się objaśnieniami i nie wiadomo, które źródło podaje je poprawnie.

[12] Presevapedia, [on-line] http://preservapedia.org/, dostęp 10 02 2017.
[13] CoOL Lexical and Classification Resource, [on-line]: http://cool.conservation-us.org/lex/, dostęp 8 03 2017.
[14] Plaster Architecture [on-line] http://www.palazzospinelli.org/plaster/, dostęp 5 03 2017

Słabością wyszukiwarki jest to, że użytkownik musi znać brzmienie szukanego terminu, tymczasem często wiemy, jak dana rzecz wygląda czy jak pojęcie jest definiowane, ale nie znamy konkretnej nazwy. Tu pomocne mogą być słowniki obrazkowe oraz tematyczne. Inne częste niedogodności to brak szukanego hasła, brak ekwiwalentu hasła w rodzimym języku, niemożność powiązania hasła z innymi terminami wynikającymi z kontekstu jego stosowania (jeśli słownik ma układ alfabetyczny, a nie tematyczny), trudność ze znalezieniem lub ustaleniem przydatnej kolokacji szukanego terminu w językach obcych (np. *tack time*). Dlatego tak ważne jest, by spojrzeć na słownik terminologiczny od strony potrzeb odbiorcy. Wydaje się, że ten aspekt bywa nieco marginalizowany.

Poważną niedogodnością w korzystaniu ze słowników internetowych jest to, że nie można konwertować wyszukanego hasła w celu zapisania czy wydrukowania go, co nie dotyczy oczywiście dokumentów pdf. W nich jednak znajdowanie słów, przy większej objętości dokumentu, jest wyjątkowo uciążliwe.

Przeglądając dostępne w sieci słowniki specjalistyczne łatwo dostrzec dominację niektórych języków (głównie angielskiego) i niedobory innych. Tymczasem różne obszary kulturowe wykształciły nie tylko własną tradycję budowalną, rzemieślniczą czy artystyczną wraz z oryginalną terminologią ale także różne sposoby postępowania z zasobami historycznych obiektów. Jest oczywiste, że podstawowym celem twórców wielojęzycznych słowników jest pewna standaryzacja pojęć. Jest to podejście słuszne w aspekcie międzynarodowych konwencji, publikacji wyników prac, działań prawnych, ale niepożądane z perspektywy zachowania różnic kulturowych. Jednakże nieprzekładalność pewnych pojęć czy ich niejednoznaczność nie powinna być przeszkodą w międzykulturowym porozumieniu, przeciwnie – warto włączać je w krwiobieg współczesnej cywilizacji. Tutaj pojawia się jedno z największych wyzwań w tworzeniu nowoczesnego słownika: powinien on z jednej strony zawierać koherentny zasób terminów w różnych językach, z drugiej zaś uwzględniać kulturową odmienność znaczeń. Co więcej, jego konstrukcja powinna umożliwiać promowanie różnorodności indywidualnie kształtowanych pojęć.

Przedstawiona powyżej analiza pozwala na sformułowanie wniosków i wytycznych niezbędnych do opracowania nowego typu słownika terminologii fachowej. Priorytetem powinna być jak największa przydatność takiego leksykonu dla użytkowników. Bazę terminologii należy budować w taki sposób, aby sprzyjać wydobywaniu i zachowywaniu kulturowych różnic i odmienności terminologicznej. Gromadzenie zasobu terminologicznego trzeba oprzeć co najmniej w równej mierze na istniejących już słownikach, jak i na badaniach korpusowych – analiza tekstów naukowych pozwala wychwytywać terminy pomijane w istniejących już zbiorach. Od strony translatorskiej głównym zadaniem jest poszukiwanie językowej ekwiwalencji wysokospecjalistycznych terminów. W większości przypadków zadanie takie może wykonać jedynie specjalista lub naukowiec związany z daną dziedziną ale także tłumacz posiadający najwyższe kompetencje językowe i znajomość terminów z danego zagadnienia. Wreszcie, niezwykle istotne założenie, odróżniające nowoczesną „encyklopedię terminów" od wydawnictw tradycyjnych. Otóż współczesna baza

danych terminologii fachowej powinna mieć charakter otwarty, a zatem poddawać się stałej rozbudowie, modyfikacjom i korektom. Zadanie to ułatwia wymiana uwag między użytkownikami słownika on-line, naukowcami i specjalistami, możliwa poprzez ogólnie dostępne forum powiązane z bazą danych słownika.

Należy jednak zastrzec, że jakkolwiek narzędzia cyfrowe i programy komputerowe dają możliwość tworzenia baz danych bez większych ograniczeń objętościowych, to pojawia się ryzyko zbyt dużej liczby danych, których przeszukiwanie stanie się żmudne i zniechęcające. Dlatego trzeba założyć pewną docelową wielkość zasobu. Sprzyja temu hierarchiczny układ słownika i możliwość dzielenia zasobów na gałęzie poszczególnych nauk.

Próby realizacji powyższych założeń zostały przeprowadzone w ramach dwóch projektów naukowych. Pierwszy, zakończony w 2010 roku, pozwolił na stworzenie prototypu elektronicznego słownika konserwacji, na tym etapie poświęconego konserwacji malowideł sztalugowych. Metodologia budowy bazy danych od podstaw została szczegółowo opisana w artykule [5]. Drugi projekt, rozpoczęty w ramach grantu NCN w roku 2012 i prowadzony na Politechnice Krakowskiej, został ukończony w roku 2015. Jego wyniki – wielojęzyczna baza danych – zostały udostępnione na stronie: www.imd.pk.edu.pl[15]. W słowniku, zwanym IMD, zebrano – jak wspomniano na wstępie – terminologię fachową konserwacji malarstwa w pięciu językach. Poniżej omówione zostaną najważniejsze rozwiązania, których celem była odpowiedź na postawione wcześniej zadania badawcze oraz podstawowe problemy.

4. Opis projektu

Podstawą budowy zasobu terminologii były badania korpusowe. Rozwiązanie to wynikało z przesłanek samego zadania, którym w przypadku opisywanego projektu było poszukiwanie ekwiwalencji terminologicznej, a nie tworzenie definicji. Z drugiej jednak strony w bazie danych, miała się także znaleźć terminologia tzw. „nauk pomocniczych" konserwacji, do których zaliczyć można m.in. historię sztuki, architekturę i budownictwo, technologię i technikę, biologię, chemię, fizykę czy petrografię (oczywiście w określonych obszarach). Nauki te wymagały odrębnego podejścia i początkowo wydawało się, że ich opracowanie – z uwagi i na liczne słowniki wielojęzyczne, i na istnienie międzyjęzyka (czy to terminologii binominalnej, czy wzorów chemicznych, czy wreszcie swoistej *interlingwy*, jaką są ilustracje) – nie będzie budzić większych problemów. Tak się jednak nie stało.

Przyjęto założenie, że poszczególne rozdziały będą opracowywane indywidualnie. Generalną zasadą było poszukiwanie ekwiwalentu, który powinien w danym rozdziale (a więc w danym kontekście) być tylko jeden. W przypadku jego braku pole hasła głównego pozostaje puste, a w polu opisu hasła pojawia się termin najbliższy znaczeniowo, lub objaśnienie odautorskie. Te „puste pola" słownika IMD są

[15] Projekt zrealizowany został w ramach grantu finansowanego ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji nr 2011/03/B/HS2/05355, „Wielojęzyczny słownik konserwacji. Część 1. Malarstwo sztalugowe, Część 2. Malarstwo ścienne."

szczególnie cenne: wskazują terminy i obszary, które albo nie zostały dotychczas wykryte w istniejących wielojęzycznych słownikach, albo też w znaczący sposób różnią się między sobą w poszczególnych krajach.

Kolejnym założeniem było budowanie bazy danych w formie struktury hierarchicznej. Rozwiązanie to sprawdziło się znakomicie. Tak więc terminy łączone są w rozdziałach tematycznych. Interesującym zjawiskiem jest „samoistne" dzielnie się rozdziałów na mniejsze jednostki, po przekroczeniu pewnej liczby wprowadzonych terminów. Przykładowo rozpatrzmy termin *dłuto*. Może on funkcjonować jako samodzielne hasło przełożone na kolejne języki, nie jest to jednak rozwiązanie odpowiednie dla słownika specjalistycznego. Dlatego ważne jest wyróżnienie elementów składowych dłuta (rączka, ostrze), dalej rodzajów dłut (ciesielskie, kamieniarskie, tokarskie i in.), ich nazw związanych z funkcją i kształtem (fazowane, trójkątne i in.), czy szlifami ostrzy (zerowy, płaski, wklęsły itd.). Każde z tych zagadnień wymaga opracowania osobnego podrozdziału. W następnej kolejności, wychodząc od przyjętego języka źródłowego (w którym pozyskano pierwszy zestaw terminów) poszukuje się ekwiwalentów w innych językach. W przypadku polskiego *dłuta* w języku angielskim mamy obok *chisel* także termin *gouge*, a w niemieckim *Meißel* i *Beitel*. Do tego dochodzą terminy stosowane przez rzemieślników poszczególnych specjalności w różny sposób określających konkretne rodzaje dłut: np. *żłobak, Beitel für die Herstellung von Aussparungen, blockmaker's chisel*. Jak łatwo się domyślić, w wielu przypadkach poszczególne nazwy nie mają odpowiedników w innych językach albo znaczenia pokrywają się częściowo, tylko w odniesieniu do konkretnych zastosowań.

Przykład ten daje wyobrażenie o złożoności zagadnień przed którymi stają autorzy specjalistycznego słownika wielojęzycznego. W IMD zebrano ponad 60 terminów określających rodzaje dłut do drewna, najwięcej w języku niemieckim, nieco mniej w angielskim, a najmniej w polskim. Stało się tak dlatego – i warto tu o tym wspomnieć – że to właśnie niemiecka literatura specjalistyczna w zakresie rzemiosła (choć nie tylko) jest wyjątkowo bogata i szczegółowo opracowana.

W przypadku dłut (ale też np. łączy stolarskich czy terminów architektonicznych) zadanie ułatwiają ilustracje, które w słowniku wielojęzycznym można traktować jako swoistą *interlingwę*. Co jednak z innymi naukami?

Jak wspomniano, nasze przewidywania dotyczące nauk biologicznych były raczej optymistyczne. Okazało się jednak, że właśnie terminologia tej dziedziny przysporzyła najwięcej problemów. Zjawisko to ma dwie przyczyny. Pierwsza z nich tkwi w bogactwie odmian gatunków występujących w poszczególnych obszarach geograficznych, które zgodnie z taksonomią należy prawidłowo rozróżniać i nazywać. Drugą przyczyną są kolokwialne, powszechnie stosowane nazwy, które nie opisują precyzyjnie gatunku, stąd trudno je sklasyfikować. Dobrze obrazuje tę kwestię terminologia nazw grzybów niszczących drewno. Na przykład nazwa *grzyb piwniczny*, to angielski *cellar (rot) fungus* ale też możemy się spotkać z określeniem *wet rot fungus*, co kieruje nas z powrotem do polskiej nazwy *grzyb zgnilizny mokrej*. W niemieckim to *Braune Kellerschwamm*, oraz inne: *Brauner Warzenschwamm*,

Dickhäutiger Braunsporrindenpilz. Binominalna nazwa gatunku *Coniophora puteana* ma 50 (!) ekwiwalentów[16], polski to *gnilica mózgowata*. Poprawne zestawienie tych terminów wymagało ogromnego nakładu pracy, pogłębionej wiedzy, oraz licznych konsultacji ze specjalistami[17].

W przypadku nazw związków chemicznych warto zwrócić uwagę na bogactwo synonimów. W słowniku IMD zawarto np. nazwy soli, które spotyka się w obiektach budowalnych. I tak np. *uwodniony siarczan sodu*,, to też: *10 hydrat siarczanu sodu, dziesięciowodny siarczan sodu, dekahydrat siarczanu sodu* oraz *mirabilit*"[18].

Wydawałoby się zatem, że rozdziały o charakterze teoretycznym, zawierające pojęcia wielokrotnie opisywane w różnych aktach prawnych, normach, ujmowane w glosariuszach konferencyjnych nie powinny nastręczać szczególnych trudności. I tu jednak zakres ekwiwalencji jest bardzo zróżnicowany, co znów wynika przede wszystkim z narodowych tradycji czy lokalnych uwarunkowań. Nawet podstawowe terminy, choć ich definicje są powszechnie znane i łatwe do znalezienia, mogą powodować zamieszanie i trudności z dobraniem odpowiedniego ekwiwalentu. Co więcej, okazuje się, że trudności z uzgodnieniem narodowych stanowisk co do definiowania pewnych pojęć stanowią poważną przeszkodę we współpracy zespołów przygotowujących słowniki konserwatorskie. Bardzo dobrym przykładem jest rozumienie i stosowanie tak podstawowych terminów, jak *zabytek*, *konserwacja* czy *restauracja*. Równocześnie pojawia się dużo terminów (np. *sustainable conservation*), które trzeba naprędce przekładać, czy akomodować w innych językach, ponieważ cieszą się one popularnością w języku angielskim.

Rozdział poświęcony teorii konserwacji-restauracji w IMD zawiera 63 hasła w pięciu językach, wszystkie uzupełniono o definicje. Problem różnic w definiowaniu pojęć rozwiązano w nowatorski sposób: wykonano krzyżowy przekład definicji między językiem polskim i angielskim, przy czym w wielu wypadkach podano więcej niż jedną definicję, gdyż okazuje się, że także w obrębie jednego kraju pojawia się kilka koncepcji definiowania. Dzięki temu użytkownik danego języka może porównać sposoby rozumienia pojęcia w innym obszarze kulturowym. Rozwiązanie to zastosowano tylko w odniesieniu do dwóch języków, jako eksperyment, który może w miarę potrzeb być kontynuowany.

Podczas prac nad tłumaczeniem opisów okazało się, że niezwykle ważnym elementem jest wskazanie kierunku przekładu. W zależności od tego kierunku możemy bowiem otrzymać różne wyniki, dlatego wprowadzono jeszcze jedno nowatorskie rozwiązanie, a mianowicie oznaczenie języka źródłowego (strzałka ze skrótem: eng, ger, pol. fr. ital.). Przy okazji wykryto wielu „fałszywych przyjaciół", np. terminy *renowacja* i *renovation*, które nie mają ze sobą wiele wspólnego i nie powinny być stosowane jako ekwiwalenty. Wykryto także niezwykle istotne różnice w przekładach jednego z najważniejszych konserwatorskich dokumentów – Karty Weneckiej [6].

Na zakończenie jeszcze kilka słów na temat terminologii historycznej. W słowniku wprowadzono osobną językową zakładkę poświęconą właśnie terminom historycznym (i nazwom binominalnym). Pochodzą one z różnych języków i nie są od dawna

[16] Species Fungorum, [on-line] http://www.speciesfungorum.org/Names/Names.asp, dostęp: 7 03 2017
[17] Rozdział opracowali konserwatorzy Iwona i Andrzej Komodzińscy, konsultacja dr inż. Helena Bis.
[18] Rozdział opracował chemik – profesor Roman Kozłowski.

stosowane, można jednak je spotkać w dawnych traktatach poświęconych technikom wykorzystywanym w sztuce. Podczas prac natrafiono na niezwykle interesujące źródło, a mianowicie przekłady włoskiego traktatu Aleksego Pademontana (wyd. 1555) na angielski (przekład 1595), niemiecki (1605) i polski (1568), które stanowią bardzo cenne zbiory dawnej terminologii. Wiele terminów z tej rozprawy (i jej przekładów) wprowadzono do słownika IMD w ramach rozdziału poświęconego pozłotnictwu, skądinąd pierwszemu, trzyjęzycznemu opracowaniu terminologii tej techniki[19].

Nie udało się uzyskać zadowalającego rezultatu w przekładzie podstawowych terminów prawniczych. Twórcom słownika zależało na ich zebraniu dla potrzeb np. konserwatorów prowadzących własne firmy. Osoby te często mają problem ze znalezieniem podstawowych słów ułatwiających choćby wymianę oficjalnej korespondencji. Stąd przygotowaliśmy rozdział „Praktyka zawodowa" i podrozdział „Przydatna terminologia ekonomiczno-prawna". Zestawienie terminologii prawnej w trzech językach okazało się niemożliwe. Owszem, dla pewnych, bardzo ogólnie wykorzystywanych pojęć istnieje obszar ekwiwalencji, ale dla większości podawanie innojęzycznych odpowiedników w kilku językach mogłoby wprowadzać w błąd ze względu na róznice w systemach prawnych.

5. Podsumowanie

Przedstawiony powyżej opis najważniejszych założeń badawczych i przyjętej metodologii jest z uwagi na zakres artykułu niepełny. Pozwala jednak na przedstawienie pewnych uwag i konkluzji.

Przede wszystkim prace nad słownikami specjalistycznymi należy powierzyć specjalistom określonych dziedzin. Nie można liczyć na to, że zadanie to wykonają tłumacze, filolodzy czy leksykografowie, choć ich udział w pracach należy uznać – przywołując nasze doświadczenia – za niezbędny. Optymalnym sposobem pozyskiwania terminologii są badania korpusowe. Ważne jest, aby opierały się one na nie budzących najmniejszych zastrzeżeń – tak co do zawartości merytorycznej, jak i językowej – publikacjach.

Istnieją obszary, dla których wielojęzyczne zestawienie terminów nie jest możliwe. Mamy wstępną koncepcję rozwiązania tego problemu, wymaga ona jednak weryfikacji w praktyce.

W przypadku opracowań wielojęzycznych konieczne jest wskazywanie kierunku przekładu, o czym dotychczas nie pomyślał żaden z autorów podobnych słowników.

Na zakończenie chcemy zwrócić uwagę na następującą kwestię. W obecnych czasach mamy wyjątkowe możliwości tworzenia baz danych, a przede wszystkim komunikacji i współpracy opartej na sieci. Pionierski projekt IMD może być kontynowany, a baza danych rozbudowywana o nowe gałęzie wiedzy i kolejne języki. Błędem byłoby nie wykorzystać tych doświadczeń dla ułatwienia wymiany myśli i idei między naukowcami, zarówno w kraju, jak i za granicą.

[19] Autorem rozdziału jest mistrz pozłotnictwa Ludomir Domański. Współpraca: konserwatorzy Andrzej Komodziński (współtwórca słownika) i Małgorzata Sawicki. Część przekładów na niemiecki wykonała Magdalena Duś.

Literatura

1. Getty Panel Paintings Initiative *The Conservation of Panel Piantings and Related Objects, Research Agenda 2014-2020*, praca zbiorowa, Netherlands Organisation for Scientific Research (NWO), Rijksmuseum Amsterdam, pdf.
2. Seidel-Grzesińska A., Stanicka-Brzezicka K. *Wielojęzyczne słowniki hierarchiczne w dokumentacji muzealnej w Polsce*, Muzealnictwo, 2014 (55), s. 169-179, pdf .
3. Piotrowicz G. (red.), *Wykorzystanie nowoczesnych technologii i mediów cyfrowych w Bibliotece*, Uniwersyteckiej we Wrocławiu. Stan na rok 2015, Biblioteka Uniwersytecka we Wrocławiu, 2015, pdf
4. Gurtner C., Hilbert G., Hughes D., Kozłowski R. J. *Manual of best practice in the application of r.oman cements. Roman cement, past and present. Conservation theory and practice*, Weber (opr.), program "Rocare", EU-Project No 226898, pdf.
5. Bogdanowska M., *Taylor M. IMD interdisciplinary multilingual dictionary a new online tool for communication*, s. 723-728, [w:] International Journal of Conservation Science. Vol. 4, spec. issue 2013.
6. Bogdanowska M. *Słowa, pojęcia i terminy – o zawiłościach konserwatorskiego słownictwa*, s. 65-73, [w:] *Karta Wenecka 1964-2014*, Bukowska W., Krawczyk J. (red). Wydział Sztuk Pięknych Uniwersytetu Mikołaja Kopernika, Toruń 2015.

Interdyscyplinarny Słownik Wielojęzyczny on-line - możliwości i ograniczenia

Streszczenie

W 2015 ukończono projekt poświęcony badaniom różnojęzycznej ekwiwalencji terminologicznej w wybranej dyscyplinie nauki, którą była ochrona dziedzictwa – konserwacja zabytków. W wyniku prac powstała baza danych zawierająca terminologię specjalistyczną z kilku obszarów nauki. Zasób został udostępniony on-line na stronie www.imd.pk.edu.pl. Tym samym powstał nowy typ słownika, który z założenia tworzony jest od podstaw przez specjalistów i naukowców. Aktualnie baza zawiera po 10 000 haseł w językach głównych tj. polskim, angielskim i niemieckim, oraz po około 4000 haseł w językach francuskim oraz włoskim. Zakres tematyczny obejmuje takie obszary, jak: teoria konserwacji, chemia, biologia, architektura, technika, technologie stosowane w sztuce i inne zgrupowane w 460 rozdziałach i podrozdziałach ułożonych hierarchicznie. Dzięki temu narzędziu i opracowanej metodologii możliwe jest rozbudowywanie bazy o kolejne dziedziny i języki, a tym samym, gromadzenie wiedzy w jednym miejscu. W artykule przedstawiono metodologię przyjętą podczas realizacji zadania, ograniczenia wynikające zarówno z problemów technicznych, jak i translatorskich, jak również możliwości, które stwarza kontynuacja prac nad rozbudową terminologicznej bazy danych nauki.

Słowa kluczowe: słownik specjalistyczny, baza danych, terminologia fachowa, ochrona dziedzictwa.

Interdisciplinary Multilingual Dictionary on-line – prospects and limitations

Summary

In 2015, we completed the latest phase of a research project investigating terminological equivalence in heritage studies. The resulting database, containing multilingual specialist terminology, was developed and published on our website as an open access data source. The dictionary currently contains over 10000 entries in three main languages; Polish, English and German and over 4000 in French and Italian. The topics include theory of conservation, conservation-restoration treatments, chemistry, biology, architecture and art technique and technology, amongst many others. They are grouped into 460 chapters allowing sequential reading, but also indexed hierarchically. This tool, along with the methodology and system we developed during the project, was designed to be expanded by including further scientific fields and languages. The following paper discusses the methodology, together with the limitations which resulted from the technical and linguistic difficulties we encountered.

Key words: multilingual dictionary, heritage studies, database

Sylwester Korga[1], Edyta Jakubczak[2]

Badanie wpływu technologii informacyjnych na efekty w nauce uczniów Szkoły Podstawowej w Tarnogórze

1. Wstęp

Dynamiczny proces rozwoju technologii multimedialnych ma istotny wpływ na życie codzienne, społeczne i zawodowe wielu ludzi. Niezależnie od miejsca ich przebywania za pomocą współczesnych środków przekazu mogą oni oglądać telewizję, słuchać radia, komunikować się, przesyłać pliki multimedialne oraz dokonywać transakcji bankowych. Wykorzystane w ten sposób urządzenia multimedialne tworzą wirtualną rzeczywistość.

W dobie postępującej globalizacji i rozwoju multimediów oraz urządzeń nie sposób nie docenić korzyści płynących z umiejętności posługiwania się nimi. Korzystanie ze zdobyczy współczesnej cywilizacji oparte jest na zdobywaniu oraz poszerzaniu wiedzy z zakresu technologii informacyjnych, które umożliwiają realizację procesów kształcenia oraz niwelują problem wyobcowania technologicznego.

Kształtowanie tego typu postaw społecznych opartych o dziedzinę IT (z ang. Information Technology) odbywa się w wyspecjalizowanych placówkach i jednostkach oświatowych. Jednym z zadań tego typu jednostek jest przygotowanie uczniów do życia w społeczeństwie informacyjnym. Zgodnie z aktualną podstawą programową kształcenia ogólnego dla szkół podstawowych z dnia 27 sierpnia 2012 roku Rozporządzenie MEN (Dz. U. z 2012r. poz. 977) w sprawie podstawy programowej kształcenia ogólnego w poszczególnych typach szkół [1] uczeń powinien nabyć umiejętność wyszukiwania, porządkowania i wykorzystywania informacji z różnych źródeł z zastosowaniem technologii informacyjno – komunikacyjnych na zajęciach szkolnych.

Rzeczywistość, która od wczesnych lat otacza człowieka wymusza na nim poznawanie wszystkiego, co znajduje się w jego otoczeniu w tym również technologii informatycznych. Dlatego też bardzo ważna jest rola osób dorosłych np.: rodziców, opiekunów i nauczycieli aby wskazali pozytywne i negatywne wartości nośników przetwarzania informacji oraz sposoby ich prawidłowego wykorzystywania w nauce i rozrywce.

[1] s.korga@pollub.pl; Katedra Podstaw Techniki, Wydział Podstaw Techniki, Politechnika Lubelska, www.pollub.pl
[2] jakubczak-edyta@wp.pl; Wydział Podstaw Techniki, Politechnika Lubelska, www.pollub.pl

2. Technologia informacyjna i jej zastosowanie w szkolnictwie

Technologia Informacyjna jest to technologia wykorzystywana do przetwarzania danych przy użyciu komputerów oraz oprogramowania do przekształcania, przekazywania i przechowywania oraz odzyskiwania informacji [2].

Szeroko rozumiana technologia informacyjna podzielona jest ze względu na dziedzinę wiedzy na następujące obszary:

- informatyka dyscyplina naukowa, która wytworzyła pojęcia, metody i techniki budowania złożonych systemów gromadzenia, przetwarzania, przedstawiania i przekazywania informacji i wiedzy w postaci symbolicznej [3],
- telekomunikacja dziedzina nauki i techniki zajmująca się transmisją danych na odległość przy użyciu środków łączności [4],
- telematyka rozwiązania telekomunikacyjne, informatyczne i informacyjne oraz rozwiązania automatycznego sterowania dostosowane do potrzeb obsługiwanych systemów fizycznych wynikających z ich zadań, infrastruktury, organizacji, procesów utrzymania oraz zarządzania [5].

Techniki multimedialne ze względu na swą uniwersalność są powszechnie stosowanym środkiem przekazu informacji. Znajdują one coraz większe zastosowanie w szkolnictwie. Współczesne nauczanie wpływa na komfort życia społecznego oraz pobudza wielozmysłowy odbiór za pomocą różnych bodźców i zachęca do aktywności poznawczej. Pomaga zapewnić łatwiejsze przyswajanie wiedzy przez uczniów oraz usprawnia procesy nauczania i uczenia się. Kształcenie z wykorzystaniem urządzeń multimedialnych umożliwia użycie metod nauczania i właściwego tempa pracy ucznia do jego indywidualnych możliwości [6].

Według wielu autorów literatury z zakresu kształcenia multimedialnego A. Komeński uważany jest za prekursora zasady poglądowości w nauczaniu. Priorytetem jest konieczność zdobywania wiedzy ogólnej poprzez bezpośrednie poznawanie rzeczy oraz zjawisk przyrodniczych, fizycznych, społecznych. Zawiera również zasadę poglądowości w nauczaniu, która opiera się na wykorzystaniu technicznych środków dydaktycznych oraz środków masowego przekazu (telewizja, radio, komputer, Internet). Koncepcja kształcenia multimedialnego zapewnia wszechstronny rozwój osobowości ukierunkowany na przyswajanie wiedzy, odkrywanie, przeżywanie i działanie [7-12].

Multimedia stanowią inspirację do poszukiwań oraz przemyśleń. Rozbudzają chęć odkrywania oraz zdolność efektywnego uczenia się. Pobudzają wyobraźnię oraz uatrakcyjniają proces nauczania, który wpływa na skuteczność uczenia się oraz zwiększenie motywacji do nauki [13]. Ze względu na rodzaj odbieranego bodźca można podzielić je na [6]:

- słuchowe (płyty gramofonowe, taśmy magnetofonowe, płyty CD, radioodbiorniki, instrumenty muzyczne),
- wzrokowo-słuchowe (projektory filmowe, aparaty telewizyjne, kasety wideo),
- częściowe (maszyny dydaktyczne, laboratoria językowe, preparaty, narzędzia, schematy, symbole, teksty pisane oraz drukowane).

3. Badania

3.1. Problematyka badania

Celem procesu badawczego jest określenie zależności pomiędzy zastosowaniem technologii informacyjnych, a postępami w nauce uczniów klas IV, V i VI Szkoły Podstawowej w Tarnogórze pod kątem stosowania różnych narzędzi przetwarzania informacji, aplikacji oraz programów komputerowych wykorzystywanych przez respondentów.

3.2. Metodologia badań

W celu przeprowadzenia badań dotyczących wpływu technologii na postępy w nauce uczniów zastosowano metodę sondażu diagnostycznego przy pomocy ankiety, do której kwestionariusz został opracowany w oparciu o przegląd literatury dotyczącej możliwości wykorzystania komputera przez uczniów, jako pomocy w nauce szkolnej, odrabianiu prac domowych, wpływu na osiągane wyniki.

Główne informacje dotyczące wpływu technologii multimedialnych na postępy w nauce zebrano za pomocą kwestionariusza ankiety złożonego z 22 pytań, na które odpowiedzieli uczniowie klas IV, V i VI Szkoły Podstawowej w Tarnogórze.

4. Analiza wyników

Na podstawie analizy zebranych wyników określono zależności zachodzące pomiędzy wpływem technologii multimedialnych na postępy w nauce jednostki statystycznej Szkoły Podstawowej w Tarnogórze.

Do zbadania tych zależności zastosowano obliczenia współczynników korelacji Pearsona pomiędzy badanymi zmiennymi. Współczynniki te obliczane były według wzoru [14]:

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} \qquad (1)$$

gdzie:

x_i - wartość pierwszej cechy

$\bar{x}$ - przeciętny poziom (średnia arytmetyczna) pierwszej cechy

y_i - wartość drugiej cechy

$\bar{y}$ - przeciętny poziom (średnia arytmetyczna) drugiej cechy

Współczynnik korelacji Pearsona przyjmuje wartości od - l do l. Jeżeli wartość tego współczynnika jest ujemna to znaczy, że między dwoma zjawiskami występuje korelacja ujemna, a więc ze wzrostem wartości jednej cechy, maleją wartości cechy drugiej. Wartość dodatnia współczynnika oznacza korelację dodatnią, czyli wraz ze wzrostem wartości jednej cechy wzrasta proporcjonalnie wartości drugiej cechy.

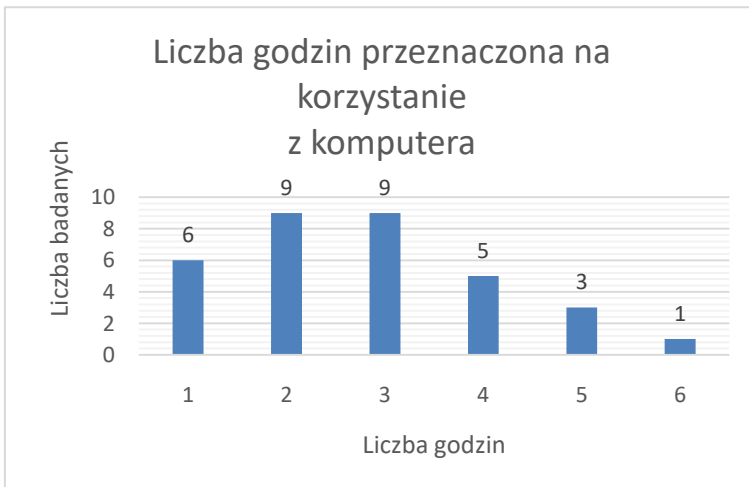
Współczynnik określa także siłę korelacji, im jest on mniejszy, tym związek jest słabszy, im większy tym korelacja silniejsza. Tak, więc brak korelacji ma miejsce, wtedy, gdy współczynnik osiągnął poziom "0", a pełna korelacja (zależność funkcyjna)

ma miejsce wtedy, gdy współczynnik osiągnął poziom "1" (korelacja dodatnia) lub -l (korelacja ujemna).

Skala korelacji, czyli przyporządkowanie konkretnym wartościom współczynnika stopni siły współzależności, ma charakter umowny. Może być różnie ujmowana przez różnych autorów opracowań statystycznych np. J. G. Guilford podaje dla współczynnika korelacji miarowej następujące określenia poziomów korelacji [15]:

- poniżej 0,20 słaba, przeciętna lecz nieznaczna,
- 0,20 – 0,40 niska, wyraźna, lecz mała,
- 0,40 – 0,70 umiarkowana, zależność istotna,
- 0,70 – 0,90 wysoka, zależność znaczna,
- 0,90 – 1 bardzo wysoka, zależność bardzo pewna.

4.1. Określenie statystycznego współczynnika korelacji -korzystanie z komputera względem osiąganych wyników jednostek badanych



Wykres 1. Liczba godzin przeznaczona na korzystanie z komputera [Źródło: opracowanie własne]

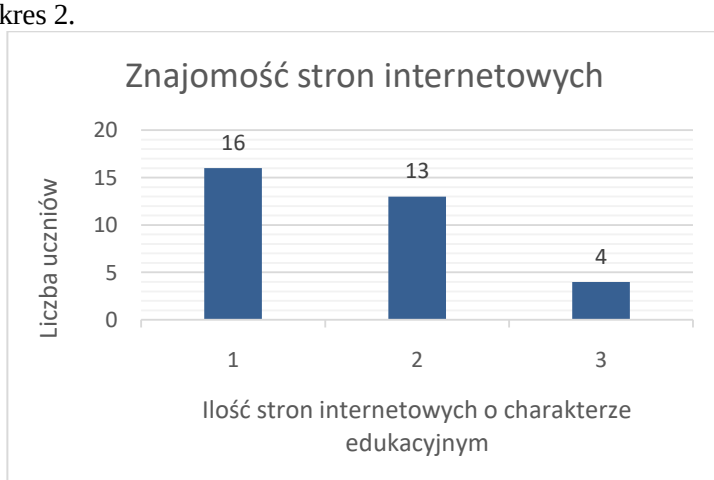
Zbadano ile godzin w ciągu doby ankietowani spędzają przy komputerze. Z odpowiedzi uzyskanych na podstawie ankiety można stwierdzić, że w ciągu dnia ankietowany spędza około 4,15 godziny. Korzystanie z komputera, a wyniki w nauce osiągnęły poziom korelacji 0,34, co oznacza niską zależność według przyjętej skali. Korelacja jest dodatnia wynika z tego, iż ze wzrostem ilości godzin korzystania z komputera w ciągu doby u ankietowanych widoczny jest wzrost ocen szkolnych, co przekłada się na średnią semestralną. Dane przedstawiono na wykresie 1.

Jeden respondent zdeklarował, że korzysta z komputera przez 6 godzin w ciągu doby. Natomiast 5 wybrało, że 3 godzin przeznacza na tę formę spędzenia wolnego czasu, 5 badanych wykorzystuje komputer przez 4 godzin, 9 ankietowanych przez 3 lub 2 godziny korzysta z komputera, najmniej czasu – 1 godzina pracuje z komputerem 6 badanych.

Wykres 2. Częstotliwość korzystania z programów edukacyjnych
[Źródło: opracowanie własne]

Z ankiety przeprowadzonej autorka dowiedziała się ile godzin w ciągu tygodnia jednostka statystyczna – uczeń korzysta z programów edukacyjnych. Największą populację stanowią ankietowani, którzy korzystają z programów raz w tygodniu jest ich, aż 13 osób. Na drugim miejscu wykorzystanie programów 4 razy w tygodniu wybrało 7 badanych, natomiast 5 ankietowanych wybrało, że korzysta z programów 5 razy w tygodniu, 3 badanych potwierdza, że korzysta 6 razy w tygodniu, natomiast 2 oraz 3 respondentów wskazało dwukrotne korzystanie z programów edukacyjnych. Natomiast tylko jedna jednostka statystyczna wskazała odpowiedź, że stosuje program dziesięciokrotnie w ciągu tygodnia.

Korelacja współczynnika krotności do średniej semestralnej wynosi 0,42, co oznacza umiarkowaną zależność, która jest istotna dla badanej cechy. Stwierdza się, że współczynnik jest dodatni i skutkuje to wskazaniem zależności ze wzrostem wykorzystania programów edukacyjnych, średnia semestralna uczniów rośnie. Wyniki ilustruje wykres 2.



Wykres 1. Znajomość stron internetowych [Źródło: opracowanie własne]

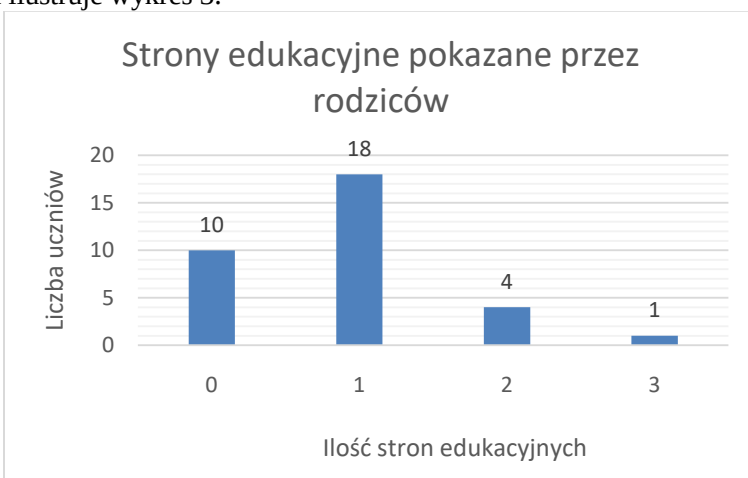
Ankietowani na pytanie ile znają stron internetowych odpowiadali w różny sposób. Najczęstszą odpowiedzią padającą była znajomość tylko jednej strony edukacyjnej wybrało ją 16 ankietowanych. Na drugim miejscu znalazło się 13 respondentów, którzy potrafią wymienić 2 strony. Najmniejszą populacją okazali się badani, którzy są w stanie podać przykłady 3 stron edukacyjnych, które znają.

Współczynnik korelacji wynosi 0,47 (korelacja umiarkowana, zależność istotna). Ponadto jest dodatnia, co wskazuje na wzrost jednej cechy przy wzroście drugiej. Z przeprowadzonej analizy pytania: „Ile znasz stron internetowych o charakterze edukacyjnym?" Wnioskuje się wzrost średniej semestralnej uczniów przy wzroście znajomości stron o charakterze edukacyjnym.

Najczęściej badani wymieniali takie strony edukacyjne jak:

- http://www.sieciaki.pl/
- http://www.anglomaniacy.pl/
- https://nauczyciel.planetaenergii.pl/
- https://www.wikipedia.org/
- http://www.detektywhoracy.operon.pl/
- http://www.kurnik.pl/
- http://www.matzoo.pl/
- http://www.polalech.pl/

Wyniki ilustruje wykres 3.

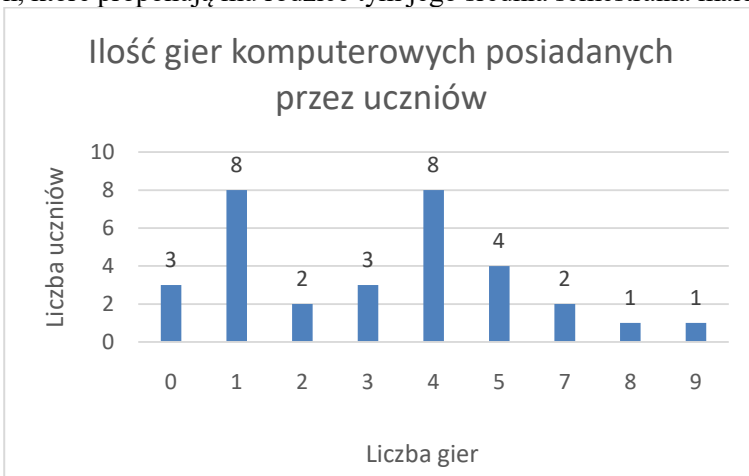


Wykres 2. Strony edukacyjne pokazane przez rodziców [Źródło: opracowanie własne]

Z wyników przeprowadzonej ankiety wynika, że rodzice badanych w większości są czynni zawodowo, a ich praca nie ma ścisłego związku z dokształcaniem się. Ponadto niektórzy zajmują się tylko domem, a ich obecność w sieci dotyczy głównie robienia zakupów, przeglądania stron kulinarnych i innych stron o charakterze praktycznym. Większość rodziców nie stara się pokazać jednostce badanej stron o charakterze edukacyjnym, ponieważ sami ich nie szukają lub nie zdają sobie sprawy, że istnieją takie sposoby pomocy dzieciom w pokonywaniu trudności lub utrwaleniu i ugruntowaniu zdobytej wiedzy. Stosują komputer, jako medium własnej rozrywki.

W grupie badanej populacji 10 ankietowanych zadeklarowało, że rodzice nie pokazali im żadnej strony internetowej. 18 respondentów wybrało odpowiedź w której zawarta jest tylko 1 strona, natomiast 4 badanych wskazało, że zna 2 strony, które pokazali im rodzice. Jeden respondent zna 3 strony internetowe. Wyniki przedstawia wykres 4.

Powyższe wyniki badań zestawiono i wykonano obliczenia statystyczne, dzięki którym obliczono współczynnik korelacji, który wynosi -0,31 świadczy to o korelacji niskiej i wyraźnej. Proponowane strony edukacyjne przez rodziców nie wpływają w istotny sposób na zależność pomiędzy średnią semestralną, a znajomością stron edukacyjnych. Współczynnik korelacji jest ujemny, więc stwierdzono, że im uczeń zna więcej stron, które proponują mu rodzice tym jego średnia semestralna maleje.

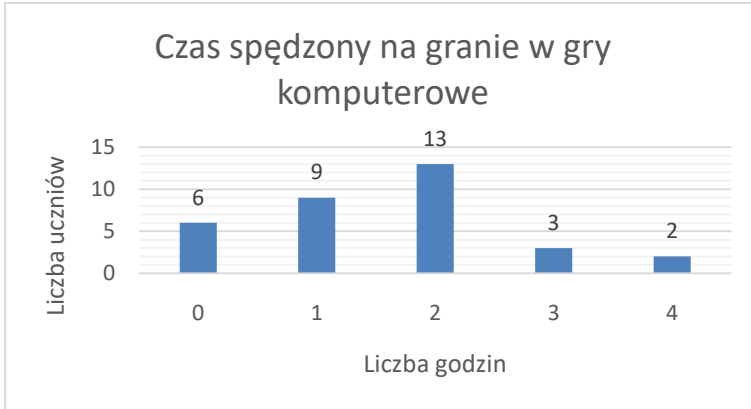


Wykres 3. Ilość gier komputerowych posiadanych przez jednostkę statystyczną
[Źródło: opracowanie własne]

Z przeprowadzonego wywiadu wśród badanych oraz nauczycieli. Wynika, że badani najchętniej spędzają wolny czas grając w gry komputerowe. Są dla nich ucieczką od rzeczywistości oraz mogą tam odreagować emocjonalne i stresujące sytuacje.

Pytanie 14 nie sprawiało trudności podczas udzielenia odpowiedzi. Granie w gry komputerowe jest częstym sposobem spędzania wolnego czasu, ale również ucieczką od obowiązków szkolnych oraz domowych. Rozpiętość posiadanych gier jest różnorodna sięga ona od 0 do 9. Średnio na jednostkę badanej populacji przypada 2,96 gier komputerowych. Taka sama liczba badanych (8) podała, że posiada 1 lub 4 gry komputerowe. Na drugim miejscu znalazło się 4 ankietowanych, którzy mają 5 gier. Trzecie miejsce zajmują respondenci z 3 grami lub nieposiadający ich wcale. Przedostatnie miejsce to badani, którzy odpowiedzieli, że posiadają 2 lub 7 gier. Na ostatnim miejscu plasują się ankietowani, którzy dysponują 8 lub 9 grami komputerowymi, co ilustruje wykres 5.

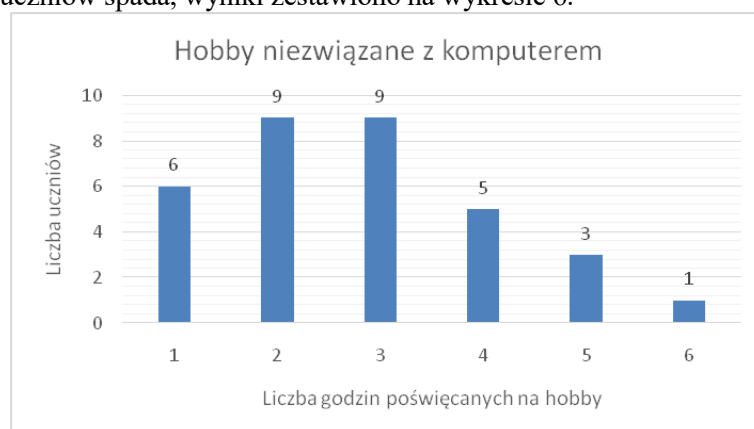
Współczynnik korelacji Poissona wynosi -0,51 co świadczy o umiarkowanej zależność, która jest istotna dla badanych oraz jej wpływie na ich średnie semestralne. Współczynnik jest ujemny tym samym wywnioskowano, że mniejsza liczba gier komputerowych wpływa pozytywnie na wyniki w nauce szkolnej.



Wykres 4. Liczba godzin poświęcana na granie w gry z rodzeństwem
[Źródło: opracowanie własne]

Ankietowani odpowiadali najczęściej, ze spędzają 2 godzinny dziennie grając z rodzeństwem takich odpowiedzi padło 13. Dziewięciu badanych zdeklarowało, że gra z rodzeństwem w gry komputerowe przez godzinę dziennie. W sondażu diagnoztycznym ankietowali odpowiedzieli sześciokrotnie, że nie posiadają rodzeństwa. Nieliczni, 2 lub 3 ankietowanych udzieliło odpowiedzi, że czas spędzony z rodzeństwem wynosił od 3 do 4 godzin.

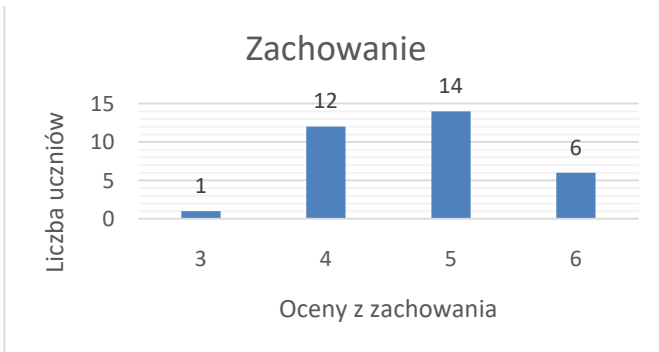
Współczynnik korelacji między czasem spędzonym na granie w gry komputerowe z rodzeństwem, a średnią semestralną badanego wynosi -0,48. Oznacza to, że zależność pomiędzy ilością godzin poświęconą na granie w gry komputerowe z rodzeństwem, a średnią semestralną jest umiarkowana. Autorka zauważyła, że współczynnik jest ujemny. Daje to możliwość wyciągnięcia wniosku: przy wzroście spędzonego czasu z rodzeństwem na granie w gry komputerowe średnia semestralna badanych uczniów spada, wyniki zestawiono na wykresie 6.



Wykres 5. Liczba godzin poświęcanych na hobby niezwiązane z komputerem
[Źródło: opracowanie własne]

Na podstawie wykresu 7. można wnioskować, że największą grupę stanowi – 18 badanych, którzy 2 lub 3 godziny tygodniowo przeznaczają na hobby niezwiązane z komputerem. Godzinę w ciągu tygodnia 6 ankietowanych wykorzystuje na swoje zainteresowania. W sondażu 5 respondentów podało, że 4 godziny poświęca na swoje pasje. Trzykrotnie wskazano odpowiedź, że przeznacza 5 godzin tygodniowo. Jedna jednostka statystyczna wskazała, że tygodniowo przez 6 godzin oddaje się swoim hobby.

Poświecenie czasu na hobby niezwiązane z komputerem ma słabą korelację z wynikami w nauce o wartości 0,28 i jest to korelacja dodatnia, zatem pozytywnie wpływa na średnią końcową badanej populacji.



Wykres 6. Oceny z zachowania [Źródło: opracowanie własne]

Ocena z zachowania rośnie wprost proporcjonalnie do średniej semestralnej. Z danych przedstawionych na wykresie wynika, że 6 ankietowanych ma ocenę wzorową, 14 bardzo dobrą, 12 ocenę dobrą oraz 1 ocenę poprawną z zachowania.

Ocena z zachowania jest najmocniejszą korelacją spośród wszystkich badanych zależności. Jej współczynnik wynosi 0,78.

Wnioskuję, że przy wzroście oceny z zachowania średnia ucznia rośnie wprost proporcjonalnie. Przedstawia wykres 8.

5. Podsumowanie

Przeprowadzona analiza sondażu diagnostycznego ukazuje zależności zachodzące pomiędzy ocenami semestralnymi grupy badawczej, a znajomością stron edukacyjnych proponowanych przez rodziców. Analizowano, czy wyniki w nauce zależą od ilości gier komputerowych. Wskazano wpływ oceny z zachowania na wysokość średniej ocen. Sprawdzono, czy czas spędzony na granie w gry z rodzeństwem, własne zainteresowania i ilość urządzeń elektronicznych wpływa na wyniki szkolne.

Na podstawie przeprowadzonych badań i wykorzystanych obliczeń statystycznych uzyskano następujące zależności pomiędzy badanymi czynnikami:

- Znajomość stron edukacyjnych polecanych badanym przez ich rodziców jest nieproporcjonalna do ich wyników szkolnych i nie ma wpływu na średnią symetralną.
- Wzrost liczby gier komputerowych posiadanych przez respondentów w domu wpływa pozytywnie na ich wyniki w nauce i znacząco podnosi średnią symetralną.

- Ocena z zachowania ma ścisły związek z osiągnięciami dydaktycznymi badanej populacji. Wysokość średniej semestralnej jest wprost proporcjonalny i zależna od rodzaju oceny z zachowania. Wyższą ocenę z zachowania mają ci badani, których średnia ocen mieści się w granicach 4,5 i powyżej 5.
- Wzrost czasu ankietowanego spędzonego z rodzeństwem przy komputerze przeznaczonego na granie w gry negatywnie wpływa na wywiązywanie się z obowiązków szkolnych powoduje obniżenie średniej ocen. Biorąc pod uwagę inny aspekt tego zagadnienia, ma pozytywny wpływ na rozwój społeczny i emocjonalny, czy budowane są w ten sposób emocje, więzi rodzinne, czy bardziej chęć zdobycia lepszej pozycji w rankingu. Nie możliwe jest stwierdzenie jakich gier to dotyczy, ponieważ nie pytano o to w ankiecie.
- Średnia ocen semestralnych jest wyższa im więcej czasu badani przeznaczyli w ciągu tygodnia na własne zainteresowania.
- Liczba komputerów w domu respondentów nieznacząco wpływa na ich wyniki w nauce.
- Wyniki osiągane przez ankietowanych nie do końca zależą tylko od nich samych. Na pewno są ściśle związane z ich środowiskiem domowym, rówieśniczym, bagażem genetycznym, poradnością wychowawczą rodziców oraz innymi czynnikami zewnętrznymi.
- Dostępność urządzeń elektronicznych i Internetu ma pośredni wpływ na wyniki dydaktyczne i wychowawcze. Technologia elektroniczna pomaga w życiu codziennym i nauce, pod warunkiem, że każdy korzysta z niej racjonalnie.

Zależności zachodzące pomiędzy wpływem technologii multimedialnych, a postępami w nauce jednostki statystycznej, współczynnik i poziom korelacji Pearsona przedstawiono w tabeli 1.

Tabela 1. Najbardziej istotne badane determinanty z punktu widzenia wychowania i nauczania ucznia

Zależności zachodzące pomiędzy wpływem technologii multimedialnych, a postępami w nauce jednostki statystycznej	Współczynnik korelacji Pearsona	Poziom korelacji
Zachowanie	0,78	wysoki
Ilość posiadanych gier komputerowych przez ucznia	-0,51	umiarkowany
Czas spędzony na gry komputerowe z rodzeństwem	-0,48	umiarkowany
Znajomość stron internetowych	0,47	umiarkowany
Krotność korzystania z programów edukacyjnych	0,42	umiarkowany
Liczba godzin przeznaczona na korzystanie z komputera	0,34	niski
Ilość stron edukacyjnych pokazanych przez rodziców	-0,31	niski
Hobby niezwiązane z komputerem	0,28	słaby

Źródło: Opracowanie własne

Literatura

1. www.men.gov.pl
2. http://uriasz.am.szczecin.pl/dydaktyka/gik/bazy_danych.pdf,
3. https://pl.wikipedia.org/wiki/Informatyka
4. https://pl.wikipedia.org/wiki/Telekomunikacja
5. https://pl.wikipedia.org/wiki/Telematyka
6. Żuk-Suska E., Suski A., Kowalczyk M., Kuwałek M. *Technologie Informacyjne i Edukacja Multimedialna w Praktyce Szklonej*, EDUCOM, Lublin,2005.
7. Bednarek J. *Multimedia w kształceniu*, Warszawa 2006.
8. Gajda J. *Media w edukacji*, Wydanie VIII, Kraków 2010.
9. Gołaszewska M. *Multimedia – krytyka i obrona. Esej o estetycznym statusie nowych mediów*, In: Piękno w sieci. Estetyka a nowe media, red. K. Wilkoszewska, Kraków 1999.
10. Monet D. *Multimedia*, Wydawnictwo Książnica, Katowice 1999.
11. Steinbrink B. Multimedia u progu technologii XXI wieku, Wrocław 1993.
12. Szkudlarek T. Media. Szkic z filozofii i pedagogiki dystansu, Kraków, 1999.
13. http://www.mwmadej.republika.pl/madej2.html
14. Komosa A. Musiałkiewicz *Statystyka*, Ekonomik,1999.
15. Juszczyk S. *Statystyka dla pedagoga*, Wydawnictwo Adam Marszałek, Toruń 2001.

Badanie wpływu technologii informacyjnych na efekty w nauce uczniów Szkoły Podstawowej w Tarnogórze

Streszczenie

W niniejszej pracy zbadano występowanie korelacji pomiędzy wykorzystaniem technologii multimedialnych, a wynikami w procesie nauczania uczniów szkoły podstawowej. Proces badawczy dotyczył analizy statystycznej odpowiedzi i wyników, które uzyskano na podstawie wypełnionych przez uczniów ankiet. W pracy przygotowano zestawienia cech statystycznych, jako determinantów badanych populacji i zbadano występowanie zależności między nimi, jak również obliczono moc korelacji. Wykorzystanie obliczeń statystycznych dotyczących korelacji Pearsona pozwoliło uzyskać odpowiedź na pytanie – Które czynniki dotyczące wykorzystania urządzeń multimedialnych są korzystne, a które są niekorzystne w procesie nauczania osób ankietowanych. Analiza statystyczna na podstawie, której przebadano grupę uczniów wykazała, że technologia informacyjna pozytywnie wpływa na wyniki w nauce, zachowanie, kontakty interpersonalne, więzi rodzinne oraz percepcje bodźców z otoczenia ucznia.

Słowa kluczowe: Technologia informacyjna, korelacja Pearsona, techniki multimedialne, efekty w nauce, badania statystyczne

The study of information technology affecting students learning progressin Tarnogóra Primary School

This documentation was examined the occurrence of the correlation between the use of multimedia technologies and grades in the process of teaching elementary school pupils. The research process concerned the statistical analysis of the answers and the results that were obtained from questionnaires filled out by students. The study prepared statement statistical characteristics as determinants of the populations studied and examined the occurrence relationships between them as well as the calculated correlation power. The use of statistical calculations on the Pearson correlation allowed to answer the question - Which factors relating to the use of multimedia devices are beneficial and which are adverse to the process of teaching people surveyed. The statistical analysis used to test a group of students showed that information technology has a positive effect on learning process, behavior, interpersonal relationships, family bonds and perceptions of stimuli from the environment pupils.

Keywords: multimedia technologies, Pearson correlation, statistical analysis, IT, process of teaching

Stanisław Skulimowski[1], Tomasz Szymczyk[2]

Przykłady wykorzystania wybranych stanowisk laboratoryjnych Instytutu Informatyki Politechniki Lubelskiej do celów dydaktycznych

1. Wprowadzenie

Specyfiką dziedziny nauki jaką jest informatyka, jest jej nieustanna zmienność. Zmieniają się nie tylko technologie, języki programowania, frameworki ale także obszary wykorzystania informatyki. Pociąga to za sobą konieczność nieustannego dokształcania się zarówno wykładowców jak i studentów.

Ciągły postęp technologiczny wymusza nieustanną naukę w celu doskonalenia procesów wytwórczych. Proces uczenia jest tym trudniejszy im większa jest złożoność zagadnienia oraz koszt technologii. Złożoność oznacza nie tylko zakres i poziom wymaganych umiejętności twardych (np. znajomość języków programowania, wiedza specjalistyczna) oraz miękkich (np. zarządzanie czasem, komunikatywność, kreatynność, innowacyjność), ale też konieczność znajomości kontekstów procesu wytwórczego – czemu mają służyć te technologie, kto będzie z nich korzystał itp. Brak wiedzy na temat kontekstów często prowadzi do tzw. „rozmijania się teorii z rzeczywistością". Dodatkowo, rosnąca złożoność projektów informatycznych, prowadzi do wzrostu czynników ryzyka, które przy nieodpowiednim zarządzaniu (w tym: zarządzaniu zespołem, zarządzaniu ryzykiem) mogą prowadzić do dezorientacji, spadku motywacji, a w skrajnych przypadkach do tworzenia paniki.

Autorzy opisali, w jaki sposób opisane powyżej problemy są zmniejszane i nowelowane w czasie zajęć dydaktycznych prowadzonych przy wykorzystaniu aparatury dostępnej w Laboratorium Programowania Systemów Inteligentnych i Komputerowych Technologii 3D. W artykule opisano metody dydaktyczne stosowane w pracy ze studentami w laboratorium a także przedstawiono przykładowe możliwości adaptacji opracowanych rozwiązań. W dalszej części artykułu zamieszczono także przegląd projektów studenckich, stanowiących rezultat działań mających na celu usamodzielnienie studentów i wzbudzenie w nich zdolności kreatywnych.

Artykuł stanowi rozszerzenie do analizy potencjału badawczego wybranych stanowisk laboratoryjnych Instytutu Informatyki, rozpoczętej w 2016 roku [1].

[1] s.skulimowski@pollub.pl, Instytut Informatyki, Wydział Elektrotechniki i Informatyki, Politechnika Lubelska, www.pollub.pl
[2] t.szymczyk@pollub.pl. Instytut Informatyki, Wydział Elektrotechniki i Informatyki, Politechnika Lubelska, www.pollub.pl

1.1. Laboratorium Programowania Systemów Inteligentnych i Komputerowych Technologii 3D

Laboratorium Programowania Systemów Inteligentnych i Komputerowych Technologii 3D, w skrócie LAB 3D, jest jedną z pracowni Instytutu Informatyki, Wydziału Elektrotechniki i Informatyki Politechniki Lubelskiej. Laboratorium zostało otwarte w 2015 roku. Mieści się w budynku Centrum Innowacji i Zaawansowanych Technologii.

W LAB 3D prowadzone są prace badawczo-wdrożeniowe, projekty badawcze nad wykorzystaniem metod grywalizacyjnych w edukacji.

Laboratorium pełni także rolę promocyjną dla edukacji i naukowości. Odbywają się tam pokazy z udziałem dzieci i młodzieży, uczniów szkół i laureatów konkursów, którzy chcieliby sprawdzić jak wygląda praca nauczyciela akademickiego, oraz jak wygląda proces badawczy z wykorzystaniem takich technologii jak interfejsy wirtualnej i rozszerzonej rzeczywistości, skanery 3D oraz drukarki 3D. Nierzadko w laboratorium goszczą także przedstawiciele z innych uczelni lub przemysłu zarówno z Polski, jak i z zagranicy.

2. Opis metod dydaktycznych stosowanych w laboratorium

Regularny tryb nauczania zakłada pozyskiwanie wiedzy deklaratywnej (ang. declarative knowledge), którą można zastosować w wielu przypadkach i scenariuszach [2]. Często jednak faza wdrażania opracowanych, teoretycznych rozwiązań ujawnia brak zdolności w posługiwaniu się określonymi urządzeniami i brak zrozumienia dla ograniczeń tych technologii. Część technologii znajdujących się w LAB 3D jest udostępniana studentom, jako narzędzie do realizacji projektów programistycznych. Studenci mają możliwość skorzystać i wypróbować (ang. hands-on learning) często dość kosztowne urządzenia. W ten sposób ogranicza się barierę dostępności technologii, która często powoduje rezygnację z projektów jeszcze przed ich rozpoczęciem.

W skład zadań przeznaczonym studentom wchodzą: zapoznanie się ze specyfikacją urządzeń, przetestowanie ich możliwości w środowisku laboratoryjnym, testowanie i prowadzenie eksperymentów z wykorzystaniem własnych programów i algorytmów. Przeniesienie problemów informatycznych z przestrzeni programowej do przestrzeni materialnej i praca z technologiami przemysłowymi, umożliwia zdobywanie wiedzy w praktyce i pozwala lepiej zrozumieć możliwości i ograniczenia dotychczas stosowanych i znalezionych rozwiązań. W ten sposób możliwe jest tworzenie wiedzy proceduralnej (ang. procedural knowledge) [2].

Posługiwanie się dodatkowymi interfejsami, wymaga od studentów wyjścia poza sferę teoretyczną i naukę przez praktykę (ang. learning-by-doing) na drodze implementacji i wdrażania tworzonych rozwiązań informatycznych [3]. Dodatkowo, w czasie realizacji projektów grupowych wymagających znajomości wielu pojęć i posiadaniu wielu zdolności, ujawnia heterogeniczną naturę natężenia cech i ich nie regularną dystrybucję w obrębie grupy. Oznacza to konieczność rozdzielania zadań według posiadanych umiejętności. W ten sposób nauka (w tym nauka programowania [4]) staje się przygodę (ang. adventure learning), w czasie której studenci nie tylko poznają wcześniej nieznane technologie bez ingerencji prowadzącego zajęcia [3], ale także poznają swoje własne ograniczenia i zdolności.

3. Przykłady projektów

Celem prowadzenia działań dydaktycznych w LAB 3D jest udostępnienie studentom technologii, jako narzędzi do tworzenia niekonwencjonalnych aplikacji, realizujących aktualne, społeczne potrzeby. Większość przedstawionych poniżej projektów wykorzystuje interfejsy wirtualnej (ang. VR - Virtual Reality) oraz rozszerzonej (ang. AR - Augmented Reality) rzeczywistości, w tym:

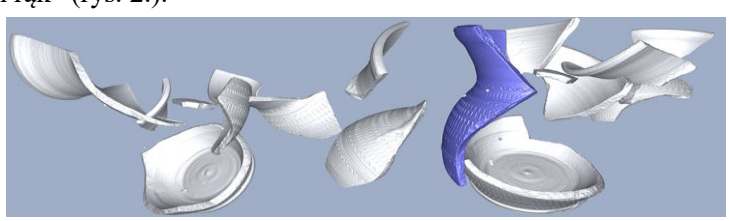
- zSpace (rys. 1. A, B) – ekran interaktywny 3D,
- Oculus Rift (rys. 1. C, D) – okulary wirtualnej rzeczywistości,
- Leap Motion (rys. 1. E, F) – czujnik rejestrujący ruchy rąk,
- 5DT Data Glove Ultra (rys. 1. G) – pomiar wychylenia palców.

Rysunek 1. Wykaz urządzeń użytych w przedstawionych przykładach projektów programistycznych: A oraz B – zSpace, C oraz D – Oculus Rift, E oraz F – Leap Motion, G – 5DT Data Glove Ultra [opracowanie własne]

3.1. Projekty z wykorzystaniem Oculus Rift

Okulary do wirtualnej rzeczywistości pozwalają na projekcję wirtualnego trójwymiarowego świata [5]. W celu ubogacenia procesu dydaktycznego, tematyka projektów zespołowych została wzbogacona o projektowanie rozwiązań zorientowanych na rekonstrukcję, ochronę i promocję dziedzictwa kulturowego, wpisując się tym samym w panujący trend digitalizacji obiektów zabytkowych i wykorzystanie ich w celach edukacyjnych [6-8]. Z przeprowadzonych doświadczeń i eksperymentów ze studentami wynika, że wzrost poziomu interakcji poparty wzrostem poziomu immersji w przestrzeń wirtualną pomaga w zrozumieniu zjawisk i zależności prezentowanych z wykorzystaniem wirtualnej rzeczywistości.

Jednym ze zrealizowanych projektów wykorzystującym Oculus Rift jako domyślny widok oraz Leap Motion do sterowania, jest rekonstrukcja (układanie) zeskanowanych skanerem 3D artefaktów muzealnych w wirtualnej rzeczywistości z użyciem „wirtualnych rąk” (rys. 2.).



Rysunek 2. Wizualizacja rozbitego naczynia [element dokumentacji studenckiego projektu programistycznego]

Innym przykładem zrealizowanego projektu wykorzystującego Oculus Rift jest wirtualne odtworzenie zniszczonych obiektów historycznych. W ramach tego projektu przebadano możliwości układania fragmentów w prostym oprogramowaniu Blender 3D zarówno z użyciem zwykłej laserowej myszki komputerowej jak również z użyciem zaawansowanej, dedykowanej myszy 3D [9].

Wyniki skanowania 3D obiektów muzealnych, przeznaczonego do celów dydaktycznych [10], umieszczono w środowisku graficznym Unity w postaci odłamków przedmiotu, które użytkownik może starać się złożyć. Użytkownik ma możliwość sprawdzenia elementów eksponatu z każdej strony zmieniając kąt widzenia pochylając i przemieszczając swoją głowę. Interpretację tego rodzaju ruchu jako zmianę parametrów projekcji zapewnia kamera referencyjna Oculus Rift śledząca ruch głowy. Z racji wykorzystania środowiska graficznego Unity (wersja 5.3.3), możliwe jest dodanie mechanik grywalizacyjnych oraz efektów wizualnych, co może zwiększyć zainteresowanie tematyką muzealnictwa.

3.2. Projekty z wykorzystaniem zSpace

Kolejnym interfejsem umożliwiającym zwiększenie immersji poprzez wzbogacenie wrażeń wizualnych jest stół do wirtualnej rzeczywistości zSpace (rys. 3.). Urządzenie to jest w stanie wyświetlać obraz w perspektywie uwzględniającej położenie głowy operatora, bez konieczności korzystania z hełmu VR takiego jak Oculus Rift – zamiast tego zSpace wykorzystuje zestaw diod referencyjnych umieszczonych przy ramie wyświetlacza oraz lekkie okulary.

ZSpace to kolejny przykład interfejsu, dzięki któremu możliwe jest znoszenie barier technologicznych i skracanie czasu nauki pojęć o wysokim poziomie abstrakcji w dziedzinie nauk ścisłych.

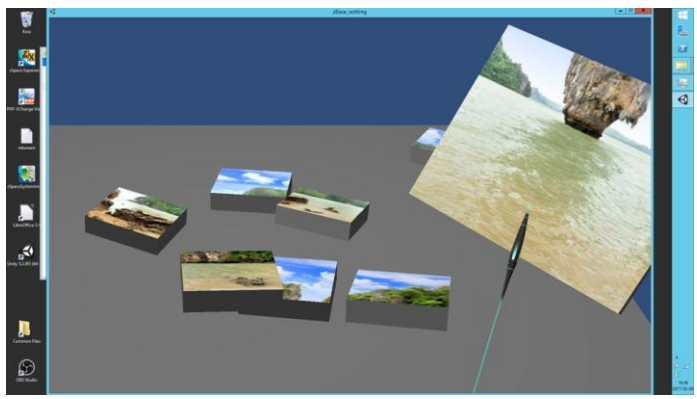
Rysunek 3. Stanowisko badawcze zSpace [opracowanie własne]

Projekty studenckie tworzone z użyciem zSpace zakładają tworzenie ekspozycji edukacyjnych w oparciu o technologię inżynierii odwrotnej 3D. Projekty te mogą służyć także jako zestaw ćwiczeń dla osób, które mają problem z orientacją w przestrzeni, problemy z poczuciem perspektywy oraz w fizjoterapii [11]. Każdy z tych projektów miał na celu wykorzystanie integralnego środowiska graficznego

Unity, w celu stworzenia wrażeń osadzonych w „nieograniczonej" przestrzeni. Ze względu na ograniczoną materiałów pomocniczych, projekty studenckie wykorzystujące zSpace należy traktować jako pionierskie.

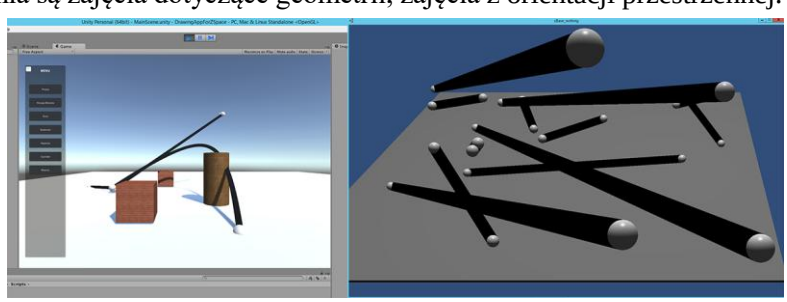
Wirtualne puzzle pozwalają na tworzenie układanki na podstawie dostępnych zestawów elementów, z możliwością zmiany perspektywy obserwowanej planszy (rys. 4.). Projekt ten powstał z myślą o osobach mając problem z widzeniem przestrzennym.

Manipulowanie elementami w przestrzeni 3D czy też rysowanie w niej bardzo ułatwia interfejs Stylus. Ten, podobny do długopisu wskaźnik, wyposażony w akcelerometr, idealnie nadaje się do naturalnego i intuicyjnego manipulowania elementami 3D.



Rysunek 4. Projekt wirtualnych puzzli [opracowanie własne]

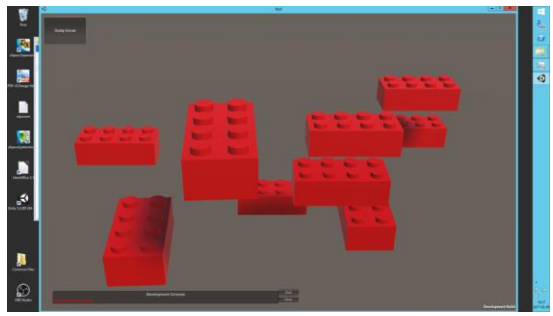
Projekt wirtualnego rysowania z użyciem krzywych umożliwia tworzenie przestrzennych brył wektorowych (rys. 5.). Potencjalnym miejscem aplikacji takiego rozwiązania są zajęcia dotyczące geometrii, zajęcia z orientacji przestrzennej.



Rysunek 5. Projekt wirtualnego rysowania krzywych [opracowanie własne]

Wirtualne budowanie obiektów zakłada możliwość tworzenia własnych modeli 3D w oparciu o zestaw dostępnych komponentów (rys. 6.). W przyszłości, rozwinięta wersja takiego narzędzia mogłaby stanowić grę edukacyjną angażującą zdolności manualne oraz kreatywnego myślenia, w podobny sposób jak ma to miejsce w przypadku klocków materialnych. Przewagą stosowania wirtualnych komponentów jest możliwość ich dowolnej duplikacji, modyfikacji (np. kolor, skala), tworzenie

i dodawanie nowych klocków czy elementów. Ponadto możliwe jest zapisanie stanu ułożonej budowli czy np. przesłanie jej na odległość do innego użytkownika. Możliwe jest także jej wczytanie i dowolna dalsza modyfikacja lub łączenie wielu projektów w jeden całościowy.



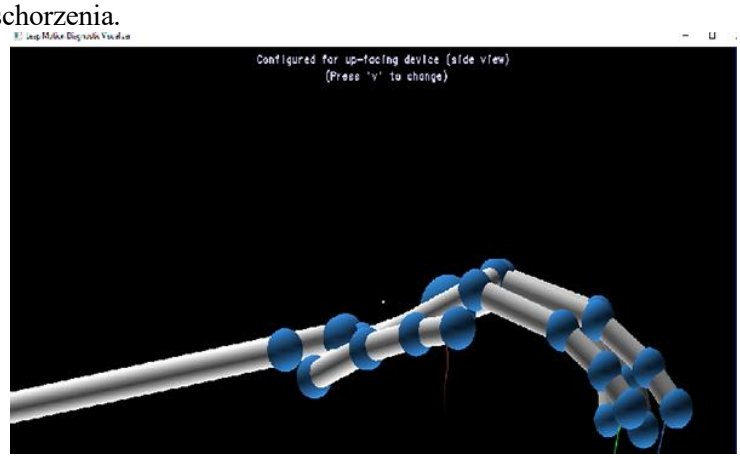
Rysunek 6. Projekt wirtualnego budowania z użyciem komponentów [opracowanie własne]

Rezultatem każdego z projektów stworzonych z użyciem zSpace było zdobycie wiedzy w zakresie tworzenia aplikacji 3D, oraz zrozumienie i wykorzystanie w praktyce pojęcia „pętli gry".

3.3. Projekty z wykorzystaniem Leap Motion

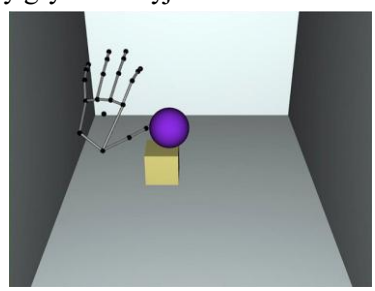
Leap Motion jest bezdotykowym urządzeniem wejściowym (ang. input device) umożliwiającym rejestrację i analizę położenia rąk oraz dłoni operatora. Dzięki temu możliwe jest posługiwanie się Leap Motion w zastępstwie za myszkę komputerową, jednak dłuższe sesje mogą okazać się bardziej męczące [12].

Urządzenie to pozwala także na określanie zgięcie palców (śledzenie statyczne) i rozpoznawanie gestów (śledzenie dynamiczne) [13] – precyzja działania urządzenia pozwala na stwierdzenie różnicy pomiędzy prawidłowo wyprostowaną dłonią, a taką posiadającą schorzenia.



Rysunek 7. Projekt pomiaru zwyrodnień dłoni [element dokumentacji studenckiego projektu programistycznego]

Projekty studenckie wykorzystujące Leap Motion do odczytu ruchu i położenia rąk oraz dłoni, skupiają się na wykorzystaniu tego urządzenia przy pomiarze zwyrodnień dłoni, jako element wstępnej weryfikacji dla przedstawicieli środowiska medycznego. Projekt przedstawiony na rysunku 7. pozwala zmierzyć przedstawić braki w zdolnościach motorycznych. Natomiast projekt prezentowany na rysunku 8. wykorzystuje do wstępnego badania elementy grywalizacyjne.



Rysunek 8. Projekt ćwiczeń zręcznościowych wykorzystujących obiekty wirtualne [element dokumentacji studenckiego projektu programistycznego]

Projekty te pozwoliły na zwiększenie wiedzy w zakresie biologii, a w szczególności wiedzy z zakresu anatomii i ortopedii. Dodatkowo, studenci mieli możliwość pracy w zakresie: integracji zewnętrznych bibliotek do projektów, nauka obsługi programów do zarządzania wersjami oprogramowania.

Opracowane aplikacje, po rozwinięciu mogą posłużyć jako element rehabilitacji przy zapaleniu kości i stawów dłoni lub reumatycznym zapaleniu stawów. Zastosowanie Leap Motion do śledzenia ruchu dłoni przy jednoczesnym użyciu hełmu VR, może stanowić alternatywę dla zajęć z wykwalifikowanym rehabilitantem – dane z czujników mogą posłużyć do weryfikacji skuteczności terapii [14], a dodatkowo gamifikacyjny aspekt uczestnictwa w wirtualnej rzeczywistości zachęca do samodzielnego podejmowania ćwiczeń. Pomimo że Leap Motion nie jest wciąż rozpoznawalnym urządzeniem na rynku, jego obsługa nie sprawia problemu nowym użytkownikom – nowe doświadczenia płynące z tego naturalnego interfejsu wręcz skłaniają do samodzielnego podejmowania próby jego opanowania.

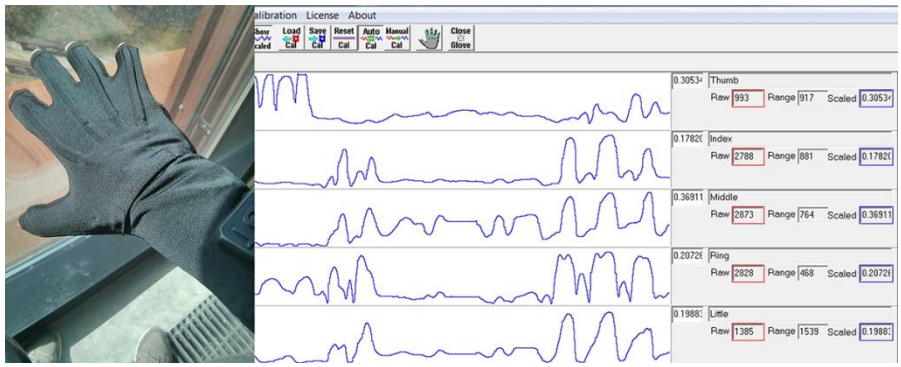
Na podstawie doświadczeń prowadzonych ze studentami udało się stwierdzić, że interfejsy opierające się na kinetyce operatora, mogą w znaczący sposób wpłynąć na wzrost poziomu koncentracji oraz mogą wpłynąć na zwiększenie koordynacji.

3.4. Projekty z wykorzystaniem rękawic 5DT

Podobnie jak Leap Motion, rękawice 5DT pozwalają na pomiar zgięcia palców u operatora, jednak rękawice potrafią to robić z większą dokładnością [15]. Wynika to z zastosowanej technologii światłowodowej i pobieranie sygnałach analogowych.

W czasie prowadzonych doświadczeń z udziałem studentów udało się wykazać ograniczenia tej technologii (m.in. błędna interpretacja przeprostów), jednak w dalszym ciągu może być użyteczna w procesie wykrywania mikro drgań i mikro skórczy.

Projekty realizowane z użyciem rękawic 5DT umożliwiają prezentację zakresu ruchu palców, do badania potencjalnych schorzeń zwyrodnień dłoni (rys. 9.) [16].



Rysunek 9. Projekt dokładnego pomiaru ruchu palców dłoni [opracowanie własne]

W przypadku tych aplikacji dużą rolę odgrywa zdolność uczenia się nowych języków, ponieważ głównym środowiskiem pracy jest Microsoft Visual Studio i język C#.

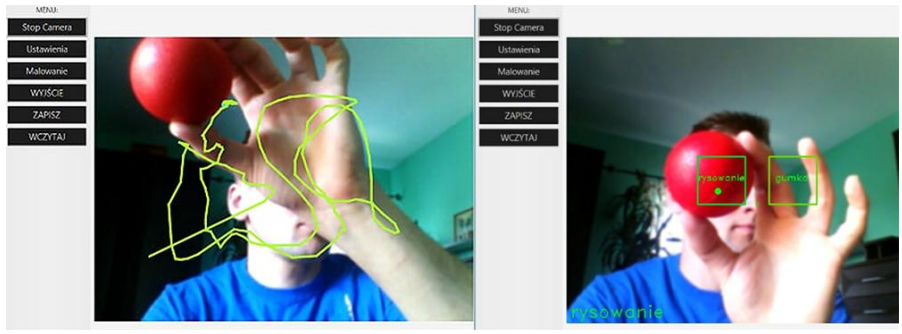
3.5. Inne projekty VR i AR

Wiele z projektów programistycznych prowadzonych w LAB 3D oraz w samym Instytucie Informatyki, wykorzystujących VR lub AR, skupia się na minimalizowaniu nakładów technologicznych, finansowych oraz na korzystaniu z szeroko dostępnych rozwiązań [17]. Projekty tworzone przez studentów działają w oparciu o kamery internetowe i komputery stacjonarne klasy PC, urządzenia przenośne typu tablet lub smartphone z systemem Android. Przykłady tego typu projektów przedstawiono w kolejnych podrozdziałach.

Podobnie jak w przypadku technologii wirtualnych rzeczywistości takich jak Oculus Rift czy zSpace, czynnik w postaci ruchomego obrazu na którym można prowadzić interakcje, potrafi być dobrym bodźcem do podjęcia samodzielnej nauki w nieznanym dotąd obszarze.

3.5.1. Wirtualne rysowanie z użyciem kamery i kolorowych markerów

Projekt wirtualnego rysowania z użyciem kamery, zakłada wykorzystanie obiektów o jednolitym kolorze wyraźnie wyróżniających się na tle innych obiektów w otoczeniu (rys. 10.). Na podstawie obrazu przechwyconego z kamery podłączonej do komputera, program pozwala na rysowanie kształtów, definiowanie gestów wykonywanych przez użytkownika i wykonywanie określonych operacji na komputerze a także pozwala na zastąpienie kursora myszki.



Rysunek 10. Projekt wirtualnego rysowania z użyciem kamery i kolorowych markerów [element dokumentacji studenckiego projektu programistycznego]

Tego typu rozwiązanie może okazać się w przyszłości przydatne dla osób o ograniczonej zdolności ruchowej oraz osób które nie są w stanie obsługiwać klasycznych urządzeń wejściowych – klawiatury i myszki komputerowej.

3.5.2. Meblowanie mieszkania z wykorzystaniem znaczników AR

Planowanie zagospodarowania przestrzeni zawsze stanowi wyzwanie dla osób, które mają do dyspozycji jedynie wymiary dostępnych mebli oraz pomieszczeń (rys. 11.). Aplikacja wirtualnego meblowania, umożliwia tworzenie wirtualnych prezentacji mebli i ich rozmieszczenia w przestrzeni.

Rysunek 11. Projekt modelowania mieszkania z wykorzystaniem znaczników AR [element dokumentacji studenckiego projektu programistycznego]

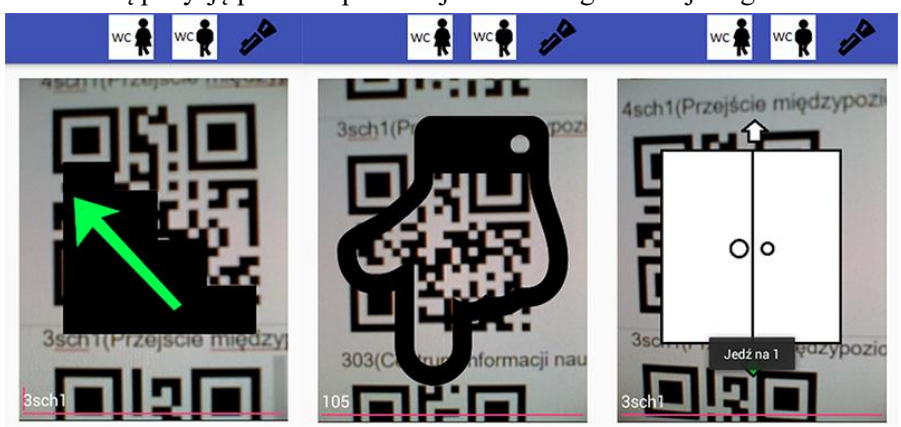
Stworzona aplikacja zakłada wykorzystanie urządzenia przenośnego z systemem Android. Urządzenie to powinny posiadać kamerę która może pracować w trybie nieprzerwanym. Użyte kody QR pozwalają na wykrycie obiektów w przestrzeni i na tej podstawie właściwe skalowanie i odwrócenie modeli mebli których położenie należy sprawić i ustalić.

3.5.3. Wirtualna nawigacja po rozległym budynku z użyciem telefonu oraz znaczników AR

Duże i rozległe budynki o skomplikowanym planie architektonicznym wymagają dużej ilości oznaczeń w celu określenia aktualnej pozycji osób wewnątrz (rys. 12,). Dotyczy to zarówno muzeów z wieloma możliwymi trasami przejścia, urzędów w których różne departamenty i biura są rozmieszczone na wielu piętrach, oraz szpitali

z długimi korytarzami, wieloma gabinetami i salami zabiegowymi i salami dla pacjentów [18].

Projekt programistyczny wirtualnej nawigacji zakłada wykorzystanie urządzenia przenośnego oraz znaczników QR rozmieszczonych w obszarze budynku, w procesie określania aktualnej pozycji i wyznaczania drogi do celu w rozległym budynku. W pierwszej kolejności użytkownik podaje w aplikacji cel do którego chce dotrzeć (np. gabinet pediatry, biuro administracyjne itp.), a następnie skanuje najbliższy kod QR znajdujący się na jednej ze ścian budynku. Następnie, na podstawie numeru zeskanowanego kodu i jego położenia względem planu budynku, system informatyczny określa aktualną pozycję petenta i prezentuje w formie graficznej drogowskaz do celu.



Rysunek 12. Projekt wirtualnej nawigacji po rozległym budynku z użyciem telefonu oraz znaczników AR [element dokumentacji studenckiego projektu programistycznego]

Projekt ten może znaleźć swoje zastosowanie w wielu placówkach publicznych, placówkach opieki zdrowotnej czy obiektach kulturalnych, odwiedzanych przez turystów i może się przyczynić do zmniejszenia ryzyka zabłądzenia oraz skrócenia czasu potrzebnego na dotarcie do konkretnego miejsca w budynku [19].

4. Podsumowanie i wnioski końcowe

Przedstawione projekty informatyczne realizowane w Laboratorium Programowania Systemów Inteligentnych i Komputerowych Technologii 3D przyczyniają się do wzrostu poziomu wiedzy, nie tylko z dziedziny informatyki czy zarządzania projektami, ale także pełnią rolę informacyjną na temat aktualnych problemów i potrzeb społecznych. Projekty realizowane w LAB 3D stanowią przykład realizacji badań naukowych, naukowo-dydaktycznych o charakterze interdyscyplinarnym i pozwalają uczestnikom na zgłębianie wiedzy w zakresie psychologii, socjologii, medycyny, wiedzy o kulturze.

Dzięki stosowaniu metod uczenia przez praktykę oraz uczenia problemowego, studenci są w stanie zdobywać wiedzę i doświadczenie na temat technologii 3D bez ograniczeń klasycznego, hermetycznego procesu dydaktycznego. Przekazanie

odpowiedzialności studentom w dziedzinie planowania i opracowania projektów informatycznych wykorzystujących VR oraz AR, pozwoliło na ujawnienie wcześniej nie rozpatrywanych trudności, przeszkód oraz możliwych celów. Dzięki ciągłemu procesowi kreowania nowych zdolności, możliwy jest także rozwój kadry naukowej i dydaktycznej Instytutu Informatyki, działających w obrębie LAB 3D.

Dalsze prace dydaktyczne w Laboratorium Programowania Systemów Inteligentnych i Komputerowych Technologii 3D będą obejmowały wykorzystanie pozostałych dostępnych urządzeń, takich jak skanery 3D i drukarki 3D oraz zestawy programowalnych i konfigurowalnych robotów Lego Mindstorms.

W artykule przedstawiono tylko kilka wybranych projektów z użyciem części dostępnego sprzętu. Pod opieką Instytutu Informatyki jest także Laboratorium Analizy Ruch i Ergonomii Interfejsów, w którym prowadzone są prace z wykorzystaniem technologii do motion capture oraz eye tracking.

Literatura

1. Skulimowski S., Szymczyk T., Montusiewicz J. *Analiza możliwości wykorzytania potencjału badawczego wybranych stanowisk laboratoryjnych Instytutu Informatyki Politechniki Lubelskiej*, Postępy w naukach technicznych oraz współczesne metody nauczania, 2016, s. 291-305.
2. Masuod M. *Declarative and procedural knowedge,* Curriculum planning knowledge & research in educational sciences, 12 (2015),s. 131-139.
3. Carlos L.E., Sullivan J.F., *Hands-on Engineering: Learning by Doing in the Integrated Teaching and Learning Program*, International Journal of Engineering Education, 15(1999), s. 20-31.
4. Miles J.C., Priest S., *Adventure Programming*, (1999).
5. Szymczyk T. *Wykorzystanie rozszerzonej rzeczywistości we współczesnych systemach informatycznych*, Proceedings Of Electrotechnical Institute, 2013 vol. 261, s. 27-43.
6. Fabola A., Miller A. *Virtual Reality for Early Education: A Study*, Communications in Computer and Information Science, 621(2016), aef6@st-andrews.ac.uk..
7. Andreoli R., Corolla A., Faggiano A., Malandrino, Pirozzi D., Ranaldi M., Santangelo G., Scarano V. *Immersivity and Playability Evaluation of a Game Experience in Cultural Heritage*, Lecture Notes in Computer Science, 10058(2016), dmalandrino@unisa.it.
8. Szymczyk T., Skulimowski S. *The use of virtual and augmented reality in the teaching process*, INTED 2017 Proceedings, s. 6570-6577.
9. Szymczyk T., Montusiewicz J., Kęsik J. *Interactive 3d environment for conducting demonstrations and training in the reconstruction of archaeological objects* ,: EduLearn Proceedings, 2016, s. 1278-1287 .
10. Montusiewicz J., Czyż Z., Kęcik J. *Using 3D replication technology in preparing didactic aid sets in the area of cultural heritage*, EduLearn Proceedings, (2015), s. 1861-1871.
11. Postolache G., Oliveira R., Postolache O. Designing Digital Tools for Physiotherapy, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 196 (2017), s.74-88.
12. Bachmann D., Weichert F., Rinkenauer G. *Evaluation of the Leap Motion Controller as a New Contact-Free Pointing Device*, Sensors, 15(2015), s. 214-233.
13. Guna J., Jakus G., Pogačnik M., Tomažič S., Sodnik J. *An Analysis of the Precision and Reliability of the Leap Motion Sensor and Its Suitability for Static and Dynamic Tracking*, Sensors, 14(2014), s.3702-3720.

14. Grubišić I., Kavanagh H.S., Grazio S.,*Novel approaches in hand rehabilitation*, Periodicum Biologorum, 117(2015), s.139-145.
15. Arkenbout E.A., Winter J.C.E., Breedveld P.,*Robust Hand Motion Tracking through Data Fusion of 5DT Data Glove and Nimble VR Kinect Camera Measurements*, Sensors, 15(2015), s.31644-31671.
16. Dipietro L., Sabatini A.M., Dario P.,*A Survey of Glove-Based Systems and Their Applications*, IEEE Transactions on Systems, Man, and Cybernetics, Part C - Applications and Reviews, 38 (2008), s. 461-482.
17. Brown A., Green T.,*Virtual Reality: Low-Cost Tools and Resources for the Classroom*, TechTrends, (2016), brownab@ecu.edu.
18. Klpwbup- Montusiewicz J., Szymczyk T. *Koncepcja lokalizacji pasywnej w budynkach użyteczności publicznej*, LOGISTYKA, 2015, nr 4, s. 8018-8027
19. Szymczyk T., Montusiewicz J., Gutek D.,*Navigation in large-format buildings based on rfid sensors and Qr and Ar markers*, Advances in Science and Technology Research Journal, 31(2016), s.263-273.

Przykłady wykorzystania wybranych stanowisk laboratoryjnych Instytutu Informatyki Politechniki Lubelskiej do celów dydaktycznych

Streszczenie
Coraz węższa specjalizacja technologiczna oraz stałe doskonalenie dostępnych narzędzi wytwórczych, prowadzi do ciągłego zwiększania potencjału badawczego danej jednostki naukowej. Zdaniem autorów w celu uwolnienia pełnego potencjału naukowego i badawczego dostępnej technologii, wskazane jest jej udostępnienie w ręce studentów, którzy będą mogli ją spożytkować bez ograniczających stereotypów. W ten sposób studenci mogą brać czynny udział w procesie badawczym. Dodatkowe połączenie procesu badawczego z procesem dydaktycznym, pozwala zintensyfikować działania w obrębie obu tych procesów, a w konsekwencji zapewnia odkrycie nowych i alternatywnych rozwiązań dotychczasowych problemów badawczych oraz weryfikację hipotez.
Artykuł stanowi przegląd projektów realizowanych w zespołach studentów, we współpracy z pracownikami naukowo-dydaktycznymi Instytutu Informatyki. Wybrane projekty i prace badawcze powstały dzięki wykorzystaniu narzędzi i urządzeń dostępnych w przestrzeni laboratoryjnej Instytutu i dotyczą między innymi możliwości wykorzystania narzędzi wirtualnej i rozszerzonej rzeczywistości.
Słowa kluczowe: nauka przez praktykę, interdyscyplinarność, wirtualna rzeczywistość, LAB 3D, Unity

Examples of the didactical use of selected laboratory positions of the Institute of Computer Science of the Lublin University of Technology

Abstract
Increased technological specialization and continuous improvement of available tools Manufacturing, leads to a continuous increase of the research potential of a scientific unit. According to the authors, in order to unleash the full potential of scientific and researched technology, it is advisable to make it available to students who will be able to use it without limiting stereotypes. In this way students can take an active part participant in the research process. The additional combination of the research process with the didactic process allows for increased activity within both processes, and consequently provides the discovery of new and alternative solutions to existing research problems and the verification of hypotheses.
This article reviews the projects implemented in the student teams, in collaboration with the academic staff of the Institute of Computer Science.
Selected projects and research works were created using the tools and equipment available in the Institute's laboratory space. These include the ability to use virtual and augmented reality tools.
Keywords: Learning-By-Doing, Interdisciplinarity, Virtual Reality, LAB 3D, Unity

Indeks Autorów