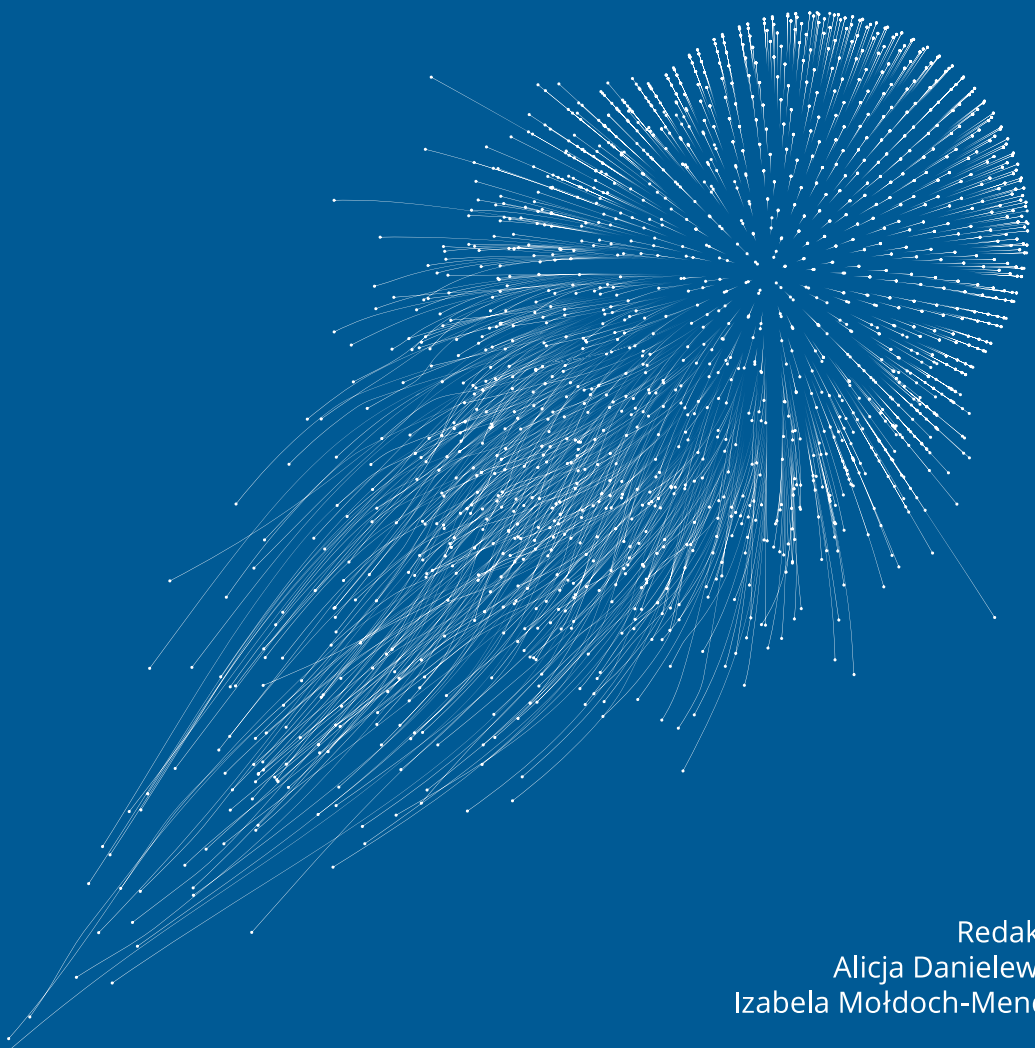
TECHNOLOGIE XXI WIEKU

AKTUALNE PROBLEMY I NOWE WYZWANIA

TOM 3



Redakcja:
Alicja Danielewska
Izabela Mołdoch-Mendoń

Lublin 2021

Technologie XXI wieku – aktualne problemy i nowe wyzwania.
Tom 3

Technologie XXI wieku – aktualne problemy i nowe wyzwania. Tom 3

Redakcja:
Alicja Danielewska
Izabela Mołdoch-Mendoń

Lublin 2021

Wydawnictwo Naukowe TYGIEL składa serdecznie podziękowania dla zespołu Recenzentów za zaangażowanie w dokonane recenzje oraz merytoryczne wskazówki dla Autorów.

Recenzentami niniejszej monografii byli:

- dr hab. Jan Chadam
- dr hab. inż. Arkadiusz Parus
- dr hab. inż. Krzysztof Regulski
- dr hab. inż. Bohdan Stawiski, em. prof. PWr
- dr hab. inż. Artur Wiatrowski
- dr Kornelia Batko
- dr Sławomir Gułkowski
- dr inż. Krzysztof Lichy
- dr inż. Paweł Mazurek
- dr inż. Łukasz Sobaszek

Wszystkie opublikowane rozdziały otrzymały pozytywne recenzje.

Skład i łamanie:
Monika Maciąg

Projekt okładki:
Marcin Szklarczyk

Korekta:
Ewelina Chodźko

© Copyright by Wydawnictwo Naukowe TYGIEL Sp. z o.o.

ISBN 978-83-67104-07-4

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

Spis treści

Katarzyna Ungeheuer, Konstanty Marszałek

Pozyskiwanie energii słonecznej w XXI wieku – bezkrzemowe, cienkowarstwowe ogniwa fotowoltaiczne 7

Bartosz Partyka, Piotr Zwierzykowski

Analiza możliwości zastosowania mechanizmów uczenia maszynowego w interpretacji danych zbieranych przez systemy monitoringu sieci i jakości usług ISP na przykładzie INEA 29

Magdalena Piwowarczyk

Przegląd algorytmów sterowania zawieszeniem pojazdu 43

Andrzej Tutaj, Paweł Król

Dobór suboptymalnych czasów realizacji elementarnych ruchów manipulatora w urządzeniu przygotowującym leki cytostatyczne 57

Danuta Miedzińska, Radosław Ciepielewski

Rozwój nowych technologii materiałowych – materiały porowate i cellulary w przemyśle XXI wieku 86

Barbara Mrzygłód, Krzysztof Regulski, Andrzej Opaliński

Uczenie maszynowe w materiałoznawstwie 105

Grzegorz Słowiński

Uczenie maszynowe z wykorzystaniem języka Python i jego bibliotek 117

Paweł Andruloniw, Piotr Zwierzykowski

W jaki sposób metody uczenia maszynowego mogą pomóc w wykrywaniu degradacji i szacowaniu jakości usług wideo w sieciach telekomunikacyjnych? 132

Krzysztof Rutkiewicz, Julia Płonka

System zarządzania przedsiębiorstwem funkcjonującym na rynku dystrybucji helu ciekłego i gazowego 142

Andrzej Kępys, Ernest Syska

VEO Weather Routing – algorytm wyznaczania tras w oparciu o dane pogodowe, mapowe i żeglowne jachtów 162

Mariusz Bedla, Adam Krechowicz, Stanisław Deniziak, Grzegorz Łukawski, Paweł Pięta, Daniel Kaczmarski, Mateusz Pawełkiewicz, Mirosław Borkowski

Zastosowanie metodyki big data w diagnostyce pacjentów z mózgowym porażeniem dziecięcym 175

Aneta Lubańska, Szymon Kiełczawa, Michał Mazur

Badanie wpływu wybranych parametrów procesu rozpylania magnetronowego na morfologię powierzchni oraz właściwości optyczne i elektryczne cienkich warstw tlenków wolframu 191

Indeks Autorów 206

Katarzyna Ungeheuer[1], Konstanty Marszałek[2]

Pozyskiwanie energii słonecznej w XXI wieku – bezkrzemowe, cienkowarstwowe ogniwa fotowoltaiczne

1. Wprowadzenie

Działalność człowieka i gwałtowny rozwój cywilizacyjny spowodowały nadmierną emisję gazów cieplarnianych do atmosfery, w znaczącym stopniu przyczyniając się do globalnego ocieplenia. Ten proces wiąże się z destabilizacją znanego nam klimatu i pogorszeniem warunków życia na naszej planecie [1]. Jednym z głównych źródeł gazów cieplarnianych jest spalanie paliw kopalnych, których zasoby są ograniczone. Produkcja energii elektrycznej w dużej mierze oparta jest obecnie na użytkowaniu nieodnawialnych źródeł energii. Potrzeby generacji elektrycznej rosną wraz ze zwiększającą się liczbą ludzi na Ziemi [2]. Prognozowane jest, że w 2050 roku produkcja energii elektrycznej będzie wynosiła 44,26 PWh, z czego blisko 50% będą stanowiły źródła odnawialne [3]. Istnieje zatem potrzeba poszukiwania nowych źródeł energii, odnawialnych i bardziej ekologicznych. Wśród takich źródeł energii możemy wyróżnić energię wiatrową, geotermalną, energię wody oraz energię słoneczną. To właśnie Słońce jest głównym źródłem energii dla Ziemi i całego istniejącego na niej życia, a także powoduje powstawanie wiatrów. Już wcześniej ludzie potrafili korzystać z energii wiatrowej czy wodnej, a w obecnych czasach w pełni rozumiemy znaczenie energii słonecznej, do tego stopnia, że potrafimy zamieniać ją w sposób bezpośredni na użyteczną dla naszej cywilizacji energię elektryczną [4]. Zostało to zapoczątkowane odkryciem efektu fotowoltaicznego w 1839 roku przez A.E. Bequerel'a. Niecałe 40 lat później powstało pierwsze proste ogniwo fotowoltaiczne wykorzystujące selen, jego konstruktorami byli W.G. Adams i R.E. Day. Natomiast dopiero w 1953 roku w laboratorium Bella powstało pierwsze krzemowe ogniwo słoneczne o sprawności 4%, a rok później już 6% [5]. Dalszy rozwój fotowoltaiki był napędzany początkowo przez zastosowania kosmiczne, gdzie wysoki koszt nowej technologii nie stanowi bariery, tak jak na etapie komercjalizacji. Ogniwa stopniowo osiągały coraz lepsze sprawności i ich cena malała, a w XXI wieku korzystanie z paneli fotowoltaicznych stało się rzeczą bardzo popularną, zwłaszcza po dołączeniu Chin do rynku fotowoltaicznego i spadku ceny produkcji krzemu. Cena energii słonecznej za watt zmalała od 76,67$ w 1977 roku do 0,21$ w roku 2020 [5]. National Renewable Energy Laboratory of the Unated States Department of Energy (Narodowe Laboratorium Energii Odnawialnej Departamentu Energii Stanów Zjednoczonych) prowadzi od 1976 roku coroczny zapis nowych osiągnięć w dziedzinie ogniw fotowoltaicznych. Takie zestawienie z 2020 roku przedstawiono na rysunku 1. Ogniwa oparte na krzemie krystalicznym w ciągu 34 lat zwiększyły sprawność od około 14% do 27,6%, co jest już bardzo blisko limitu sprawności

[1] ungeheue@agh.edu.pl, Wydział Informatyki, Elektroniki i Telekomunikacji, Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie.
[2] marszale@agh.edu.pl, Wydział Informatyki, Elektroniki i Telekomunikacji, Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie.

dla złącza opartego na krzemie wynoszącego 32% [6]. Warte uwagi są ogniwa wielozłączowe pozwalające na wykorzystanie szerszego zakresu widma słonecznego niż stosowanie pojedynczych złącz półprzewodnikowych, w której to technologii w 2020 roku udało się osiągnąć sprawność 47,1% przy koncentracji promieniowania, a dla natężenia 1000 W/m^2 uzyskano sprawność 39,2% [7]. Obecnie istotne są rozwijające się technologie, takie jak ogniwa bazowane na perowskitach, cienkich warstwach czy ogniwa polimerowe, które wykazują znacznie bardziej dynamiczny wzrost sprawności niż konwencjonalne ogniwa krzemowe.

Dla ogniw fotowoltaicznych istnieje szereg parametrów, które pozwalają na ich porównywanie. Ponadto zostały ustalone standardowe warunki, w których przeprowadza się pomiary tych parametrów. Standardowe warunki testowe (STC – Standard Test Conditions) to widmo promieniowania słonecznego AM1.5, natężenie oświetlenia na jednostkę powierzchni o wartości 1000 W/m^2 oraz temperatura ogniwa 25°C. Podstawowymi parametrami ogniwa są współczynnik wypełnienia *FF* (ang. *fill factor*) (1) oraz sprawność η (2):

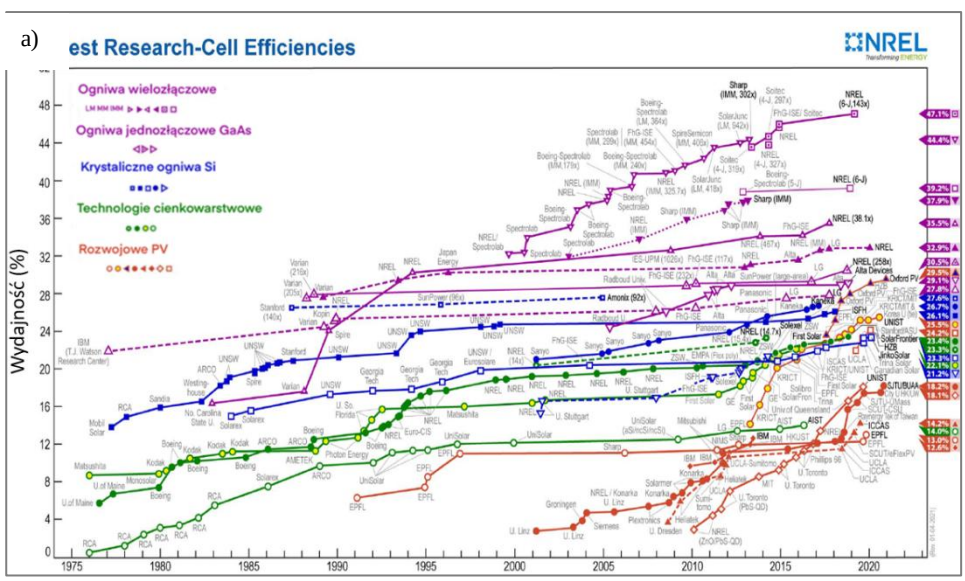
$$FF = \frac{P_m}{V_{oc} I_{sc}} \tag{1}$$

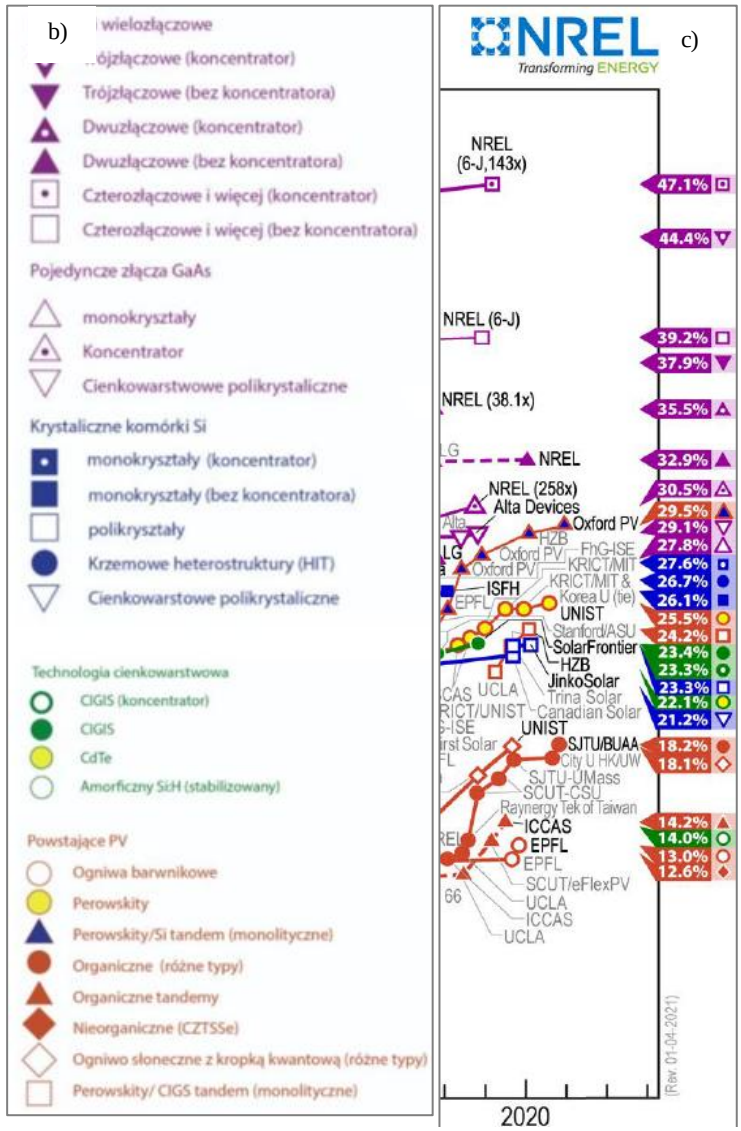
gdzie: V_{oc} – napięcie przy obwodzie otwartym, I_{sc} – natężenie prądu zwarcia, P_m – maksymalna moc ogniwa.

$$\eta = \frac{P_m}{P_s} = \frac{FFV_{oc}I_{sc}}{P_s} \tag{2}$$

gdzie: P_m – maksymalna moc ogniwa, P_s – moc dostarczona przez promieniowanie słoneczne, *FF* – współczynnik wypełnienia, V_{oc} – napięcie przy obwodzie otwartym, I_{sc} – natężenie prądu zwarcia.

a)





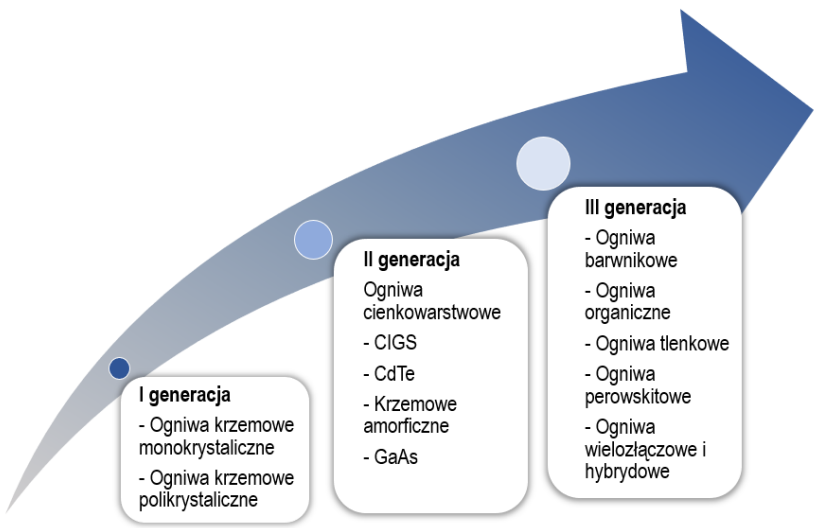
Rysunek 1. a) Zestawienie sprawności ogniw fotowoltaicznych różnych technologii z lat 1976-2019, b) legenda rodzajów ogniw, c) przybliżenie najnowszych wyników zestawienia [8]

Wyznacza się je poprzez pomiar charakterystyki prądowo-napięciowej ogniwa w standardowych warunkach testowych uwzględniając wielkość i powierzchnie ogniwa podczas wyznaczania tych parametrów. Sprawność jest najważniejszym z tych parametrów i mówi o tym, jaka część energii dostarczonej do ogniwa jest zamieniana na energię elektryczną oraz pozwala w prosty sposób porównywać ogniwa fotowoltaiczne.

W tym rozdziale opisane są podstawowe rodzaje ogniw fotowoltaicznych, takich jak krzemowe oraz cienkowarstwowe, a także technologie najnowszych generacji, opracowywane obecnie w laboratoriach badawczych, z wyróżnieniem ogniw organicznych, perowskitowych, barwnikowych i tlenkowych.

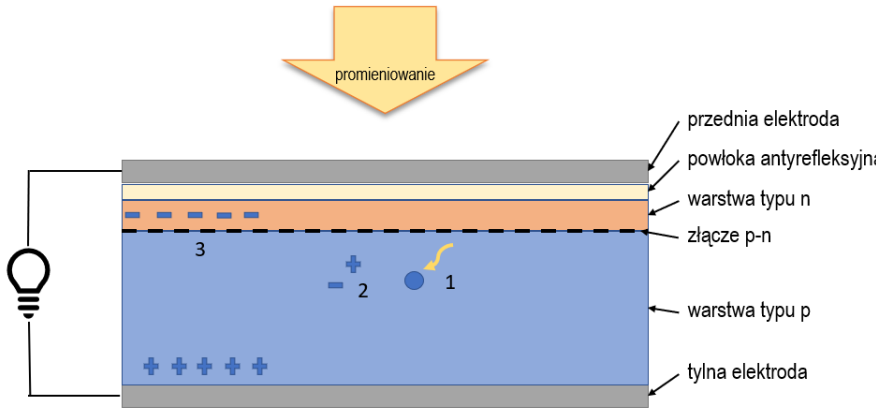
1.1. Ogniwa fotowoltaiczne pierwszej generacji

Rodzaje ogniw fotowoltaicznych są klasyfikowane ze względu na chronologiczną kolejność, w jakiej się pojawiały oraz na podobieństwa technologiczne. Wyróżnia się tak zwane generacje ogniw fotowoltaicznych: pierwszą generację, do której należą ogniwa oparte na krzemie mono- i polikrystalicznym, w drugiej generacji znajdują się ogniwa cienkowarstwowe, trzecią generację stanowią najnowsze technologie, które są w trakcie badań. Podział technologii ogniw fotowoltaicznych na generacje został przedstawiony na rysunku 2.



Rysunek 2. Generacje ogniw fotowoltaicznych [opracowanie własne]

Choć efekt fotowoltaiczny został odkryty na innych materiałach niż krzem, to właśnie ogniwa oparte na tym pierwiastku stanowią pierwszą generację oraz dominują na rynku fotowoltaicznym. Krzem jest półprzewodnikiem o przerwie energetycznej 1,12 eV. Swoje szerokie zastosowanie krzem zawdzięcza właściwościom półprzewodnikowym oraz możliwościom domieszkowania, co pozwala na otrzymywanie materiałów o półprzewodnictwie dziurowym i elektronowym i następnie złącz półprzewodnikowych, a dalej różnych elementów elektronicznych, np. diod, tranzystorów. Krzemowe ogniwo słoneczne jest oparte właśnie na złączu półprzewodnikowym p-n. Wafle krzemowe są domieszkowane w taki sposób, by otrzymać warstwę typu-p o grubości 180-300 μm, a na niej warstwę krzemu typu-n o grubości ok. 2 μm [5]. Na powstałym złączu dochodzi do rozdzielenia ładunków (dziur i elektronów) powstałych w wyniku absorpcji fotonu padających na materiał. Rozdzielone ładunki powodują powstanie napięcia, które możemy wykorzystać w postaci energii elektrycznej. Budowa krzemowego ogniwa fotowoltaicznego oraz schemat jego działania zostały przedstawione na rysunku 3. Powłoki antyrefleksyjne zmniejszają straty energii i sprawiają, że większa jej część dociera do warstw półprzewodnikowych. Warstwa typu n jest często modyfikowana poprzez trawienie, tak aby zyskać teksturowaną powierzchnię zwiększającą absorpcję promieniowania przez materiał.



Rysunek 3. Budowa ogniwa krzemowego oraz zasada działania: 1 – foton wzbudza elektron w atomie, 2 – powstaje wolny elektron oraz dziura, które po dotarciu do złącza p-n zostają rozdzielone, 3 – rozdzielone ładunki gromadzą się w przeciwnych stronach ogniwa, tworząc napięcie, które można wykorzystać do zasilania urządzeń [opracowanie własne]

Krzemowe ogniwa słoneczne są wykonane z czystego materiału. Ogniwa monokrystaliczne są produkowane z wafli krzemowych, które uzyskuje się metodą Czochralskiego. Odkrycie tej metody produkcji czystego krzemu było kamieniem milowym w zastosowaniu tego materiału, nie tylko w dziedzinie energetyki, ale przede wszystkim elektroniki. Czysty krzem otrzymany chemicznie z SiO_2 w procesie Siemensa jest umieszczany w tyglu i stapiany. Wszystko odbywa się w kontrolowanych warunkach, tak aby do materiału nie dostały się zanieczyszczenia. Do stopionego krzemu opuszczany jest zarodek monokrystaliczny, a następnie rosnący monokryształ wyciągany jest do góry z kontrolowaną prędkością. Proces hodowli kryształu jest długotrwały i kosztowny.

Krzem polikrystaliczny jest produkowany poprzez spiekanie czystego chemicznie wytworzonego krzemu w prostopadłościenne bloki. Powstałe po pocięciu wafle mają widoczne ziarna krystaliczne o rozmiarach od kilku milimetrów do kilku centymetrów. Gotowe bloki w kształcie zbliżonym do walca w przypadku ogniw monokrystalicznych oraz prostopadłościenne polikrystaliczne są cięte na wafle o grubości minimalnie 160 μm. Podczas tego etapu występują straty materiału. Gotowe wafle są trawione, domieszkowane, pokrywane odpowiednimi powłokami (np. antyrefleksyjne), nanoszone są ścieżki elektrod i wszystko jest uszczelniane za pomocą odpowiednich folii polimerowych. Ogniwa łączone są w panele, które wymagają sztywnej konstrukcji ze względu na kruchość krzemu. Wykorzystuje się w tym celu szyby oraz ramy aluminiowe.

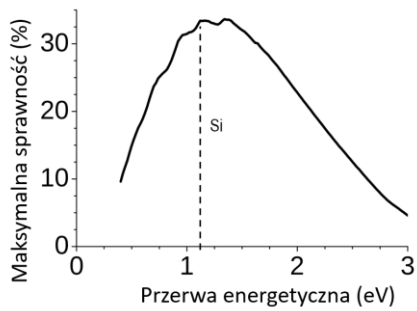
Obecnie dostępne na rynku panele pierwszej generacji dają sprawność średnio 25-27% i 23-25% odpowiednio dla ogniw mono- i polikrystalicznych [9].

Oczywiście prace nad zwiększaniem wydajności ogniw opartych na krzemie nadal trwają. Jedną z podstawowych metod jest stosowanie powłok antyrefleksyjnych oraz teksturowanie [5, 10]. Wykorzystuje się różne konstrukcje samych ogniw, np. Passivated Emitter and Rear Cell (PERC), Heterojunction with Intrinsic Thin Layer (HIT), Passivated Emitter with Rear Totally Diffused (PERT), Passivated Emitter with Rear Locally Diffused (PERL), w których odpowiednie umieszczenie domieszkowanych obszarów oraz różne sposoby rozmieszczenia elektrod pozwalają zmniejszyć straty zachodzące

przy konwersji energii [4, 11]. Ponadto istnieją prace nad hybrydowymi ogniwami, gdzie na warstwę krzemu nakładane są warstwy należące do nowszych generacji ogniw fotowoltaicznych [12].

1.2. Ogniwa krzemowe a cienkowarstwowe

Krzem przez swoją przerwę energetyczną ma ograniczone możliwości wydajności konwersji energii słonecznej na elektryczną. Zgodnie z limitem Schockleya-Quisslera maksymalna wydajność dla przerwy energetycznej krzemu wynosi ok. 32% [6]. Zależność maksymalnej możliwej sprawności konwersji energii w zależności od wielkości przerwy wzbronionej została przedstawiona na rysunku 4. z oznaczeniem przerwy energetycznej krzemu. Możliwości stosowania ogniw opartych na złączu krzemowym są zatem ograniczone. Z tego względu istotne jest poszukiwanie nowych materiałów, które pozwolą na pozyskiwanie energii bardziej sprawnie i z wykorzystaniem mniej kosztownych metod.



Rysunek 4. Limit wydajności złącza p-n w zależności od wartości przerwy wzbronionej materiału [13]

Jednymi z wad ogniw pierwszej generacji są straty materiałowe podczas procesu produkcji oraz masa samych paneli. Rozwiązaniem dla tych problemów jest stosowanie układów cienkowarstwowych jako ogniw fotowoltaicznych. Ponadto w produkcji ogniw krzemowych główny koszt to produkcja czystego krzemu. Stosowanie innych materiałów, umożliwiających wykorzystanie mniej kosztownych technologii jest niewątpliwą zaletą. Cienkie warstwy można otrzymywać z wykorzystaniem różnych metod, takich jak rozpylanie jonowe, elektrodepozycja, osadzanie warstw atomowych, metody mokre np. zol-żel, osadzanie laserem pulsacyjnym. Dodatkowo układy cienkowarstwowe zużywają mniejszą ilość surowców podczas produkcji, ponieważ samo ogniwo jest dużo cieńsze.

Ponadto ogromną zaletą ogniw cienkowarstwowych jest ich niewielka masa i możliwość nanoszenia na elastyczne podłoża. Otwiera to nowe możliwości zastosowania ogniw fotowoltaicznych, nie tylko jako wolnostojące elektrownie czy instalacje na dachach budynków, ale także fotowoltaika zintegrowana z budynkami. Inne z zalet ogniw cienkowarstwowych to mniejszy wpływ wysokiej temperatury na generowaną moc, sprawna praca przy niewielkiej ilości światła lub świetle rozproszonym, mniejszy wpływ zacienienia na pracę paneli oraz estetyczny wygląd.

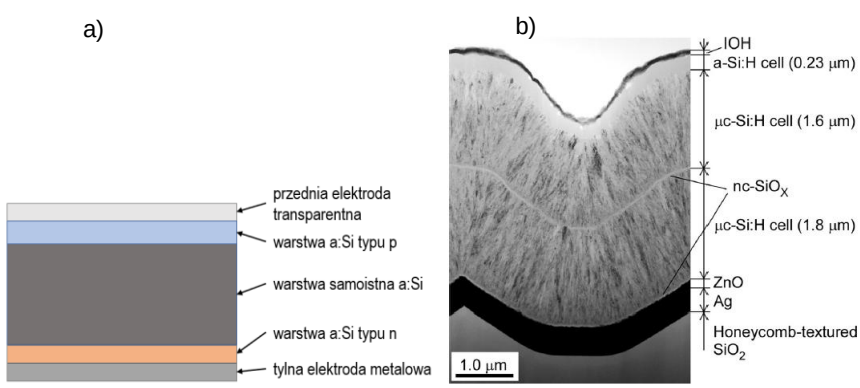
W zasadzie wszystkie technologie, poza tymi należącymi do pierwszej generacji, można nazwać cienkowarstwowymi. W drugiej generacji są to ogniwa krzemowe amorficzne oraz ogniwa oparte na innych półprzewodnikach nieorganicznych: tellurku kadmu (CdTe), selenku indu galu miedzi (CIGS), arsenku galu (GaAs).

2. Ogniwa fotowoltaiczne drugiej generacji

Do drugiej generacji zaliczane są ogniwa cienkowarstwowe oparte na amorficznym krzemie, selenku indu galu miedzi, tellurku kadmu oraz arsenku galu. Chronologicznie jako pierwsze w latach 70. pojawiły się ogniwa krzemowe amorficzne. CIGS i CdTe to polikrystaliczne półprzewodniki typu p i są wykorzystywane w ogniwach słonecznych w połączeniu z warstwą typu n, nieprzezroczystą elektrodą tylną i przezroczystym tlenkiem przewodzącym jako elektroda przednia. Oba te materiały są badane od końca lat 70. W zastosowaniu jako panele, ogniwa cienkowarstwowe są nanoszone na tafle szklane, lecz możliwe jest ich nanoszenie na podłoża elastyczne, folie metaliczne lub polimerowe [14].

2.1. a-Si – krzem amorficzny

Pierwsze ogniwo wykonane z amorficznego krzemu powstało w 1976 roku. Miało grubość 1 μm, wydajność 1,1%, a warstwa amorficznego krzemu została naniesiona z wykorzystaniem wyładowania jarzeniowego. [15] Co bardzo istotne krzem amorficzny ma większy współczynnik absorpcji niż krzem krystaliczny, co umożliwia zastosowanie cienkich warstw. Ponadto nieuporządkowana struktura sprawia, że energia wzbroniona dla krzemu amorficznego wynosi 1.7 eV, a więc jest wyższą niż dla krzemu krystalicznego, co pozwala ogniwom a-Si absorbować widzialną część widma słonecznego silniej niż podczerwoną część widma. Choć według limitu S-Q oznacza to niższą możliwą sprawność, daje pole do tworzenia ogniw wielozłączowych [4]. I tak najwyższa osiągnięta sprawność wynosi 14% dla trójzłączowego cienkowarstwowego ogniwa opartego na krzemie amorficznym i mikrokrystalicznycm (a-Si:H/μc-Si:H/μc-Si:H) [16]. Typowa budowa ogniwa a:Si jest podobna do ogniw krystalicznych (rys. 5a). Rekordowe ogniwo było naniesione na heksagonalnie teksturyzowany substrat w kilku krokach metody wspomaganego plazmą chemicznego osadzania z fazy gazowej. Elektrody stanowił tlenek cyny domieszkowany fluorem oraz tlenek cynku domieszkowany galem. Pomiędzy substratem a elektrodą nałożona została warstwa srebra, by zapewnić odpowiednie odbicie. Budowa wykonanego ogniwa jest widoczna na rysunku 5b. Pojedyncze złącza amorficznego krzemu nie pozwalają na osiąganie takich wyników. Maksymalna osiągnięta wydajność to około 12%, lecz ogniwo ulega degradacji pod wpływem światła słonecznego, a sprawność stabilizuje się na poziomie 4-8% [17].



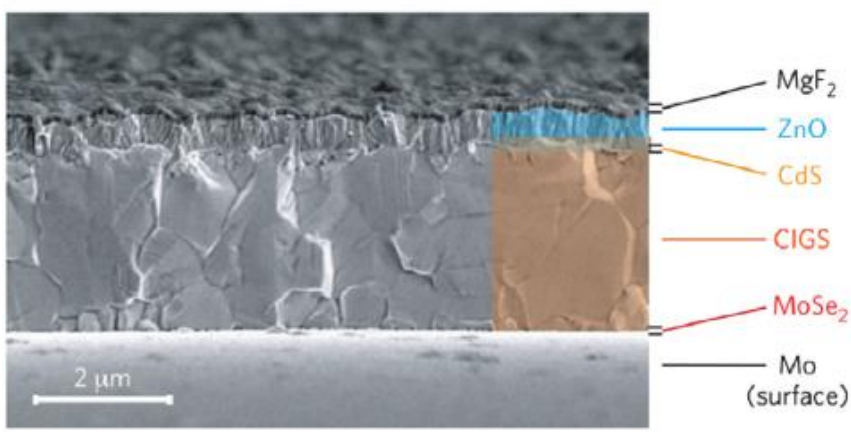
Rysunek 5. a) Typowy przekrój budowy amorficznego ogniwa krzemowego [opracowanie własne] b) obraz z transmisyjnego mikroskopu elektronowego rekordowego ogniwa amorficznego o sprawności 14% [16]

Amorficzne panele mają estetyczną ciemną barwę i można je spotkać w prostych użytkowych zastosowaniach, np. w kalkulatorach czy solarnych lampkach ogrodowych, gdzie nie jest konieczne uzyskiwanie dużych mocy czy sprawności.

2.2. CIGS

Za skrótem CIGS kryją się pierwiastki: miedź, ind, gal i selen, które tworzą związek krystalizujący w grupie chalkopirytu. Ma on wysoki współczynnik absorpcji optycznej i pozwala na sterowanie wielkością przerwy wzbronionej poprzez zmiany składu chemicznego od wartości 1 eV do 1,7 eV. Ogniwa tego typu mogą być nanoszone na podłoża sztywne, jak i elastyczne. W skali laboratoryjnej ogniwa typu CIGS osiągają bardzo dobre wyniki – sprawność 23,35% w roku 2019 udało się osiągnąć przez firmę Solar Frontier [18]. Trudniejszym zadaniem zdaje się być produkcja modułów i paneli o wysokiej sprawności, dla których rekordowe wartości to około 19%, a dla paneli dostępnych na rysunku są to sprawności bliżej 15%.

Podstawowa budowa ogniwa typu CIGS to podłoże szklane (lub stalowe jako folia elastyczna) [19], warstwa dolnej elektrody Mo, $MoSe_2$ tworzący się na granicy Mo i warstwy CIGS, który poprawia przewodnictwo pomiędzy tymi warstwami. Następnie warstwa typu p CIGS, warstwa typu n najczęściej CdS, nazywana warstwą okienną ze względu na to, że przepuszcza część promieniowania do niżej położonej warstwy absorbera. Na górze ogniwa znajduje się transparentna elektroda przewodząca najczęściej w postaci ZnO. Ogniwo może być pokryte powłoką antyrefleksyjną, na rysunku 6. jest to powłoka z MgF_2 [20, 21].

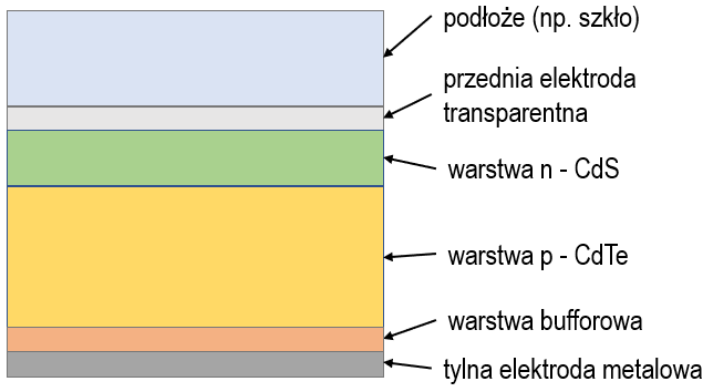


Rysunek 6. Przekrój ogniwa typu CIGS na zdjęciu z mikroskopu elektronowego przedstawiający budowę ogniwa [20]

Dużą zaletą ogniw cienkowarstwowych typu CIGS jest fakt, że zacienienie ogniwa nie ma tak negatywnego wpływu na jego działanie, jak w przypadku ogniw krzemowych. Jednym z problemów tej technologii jest cena i dostępność indu, co w przyszłości negatywnie wpłynie na kosztowność tego typu ogniw [22]. Dalsze prace nad ogniwami typu CIGS obejmują przede wszystkim rozwijanie samej technologii produkcji, tak aby zmniejszyć koszty oraz otrzymywać lepsze sprawności. Istnieją również prace wykorzystujące warstwy CIGS w ogniwach hybrydowych w połączeniu z ogniwami perowskitowymi [23] czy typu GaAs [24].

2.3. CdTe

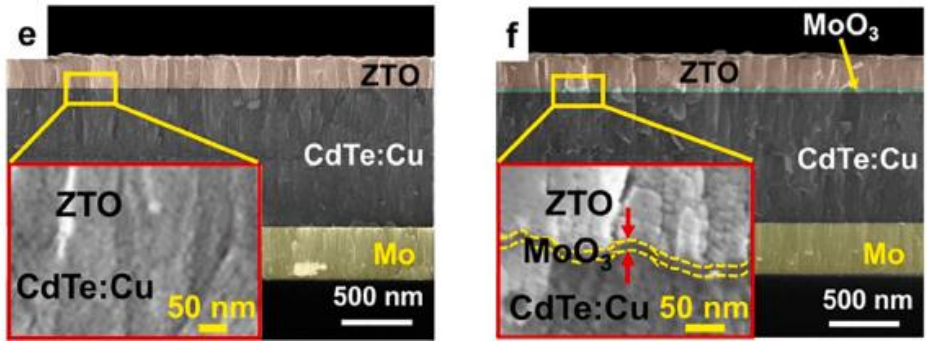
Tellurek kadmu jest związkiem półprzewodnikowym typu p o przerwie energetycznej 1,45 eV i wysokiej absorpcji. Najwyższa otrzymana sprawność ogniwa w laboratorium to 22,1% przez First Solar w 2014 roku [8, 18]. Równi ten producent dominuje w produkcji paneli CdTe o sprawności około 17% [21]. Toksyczność kadmu stanowi problem dla tych ogniw. Firma First Solar opracowała program recyklingu dla paneli oraz posiada wydajną technologię produkcji, dzięki czemu panele CdTe są konkurencyjne cenowo. Problem w przyszłości może stanowić dostępność telluru, który ma ograniczone zasoby i z pewnością wpłynie to niekorzystnie na ceny [22]. Podstawową budowę ogniw typu CdTe przedstawiono na rysunku 7.



Rysunek 7. Budowa ogniwa typu CdTe [opracowanie własne]

Badanie naukowe nad ogniwami CdTe skupiają się na poszczególnych elementach ogniwach [23, 24]. A. Hu i inni badali wpływ grubości warstwy okiennej typu n CdS:O na pracę ogniwa. W swoich ogniwach zastosowali ITO, warstwę buforową MgZnO, emiter typu n CdS:O, absorber CdTe i tylną elektrodę metaliczną z pastą węglową. Pasta węglowa z dodatkiem miedzi była po nałożeniu wygrzewana, co spowodowało dyfuzję miedzi i polepszenie kontaktu elektrycznego [25]. Absorber CdTe jest często poddawany obróbce termicznej z wykorzystaniem $CdCl_2$, co poprawia właściwości całego ogniwa [21]. Odpowiedni dobór procesu wygrzewania gotowego ogniwa, zastosowanie powłoki antyrefleksyjnej i warstwy CdS:O o grubości 30 nm pozwoliło na uzyskanie sprawności wielkości 18,05% [23].

W innym badaniu R. Singh i inni studiowali wpływ barier na interfejsach pomiędzy warstwami ogniwa oraz stosowanie warstw blokujących przepływ dziur. Na podłożu krzemowym naniesiono z wykorzystaniem naparowania jonowego elektrodę metaliczną (Mo lub Al), warstwę CdTe domieszkowane Cu oraz ZnO_2:SnO_2 (TZO) jako emitera i elektrodę transparentną. Dla części ogniw pomiędzy warstwę CdTe a TZO naniesiono cienką warstwę MoO_3 jako blokującą przepływ dziur. Zdjęcie mikroskopowe przekroju wytworzonych ogniw jest widoczne na rysuku 8. Zastosowanie warstwy MoO_3 polepszyło parametry pracy ogniwa i najlepszą sprawność osiągnięto dla ogniwa z tą warstwą przygotowanego na teksturyzowanym krzemie [26]. Własności złącz i wpływa warstwy blokującej były badane za pomocą mikroskopii sił atomowych (AFM) z sondą Kelvina.



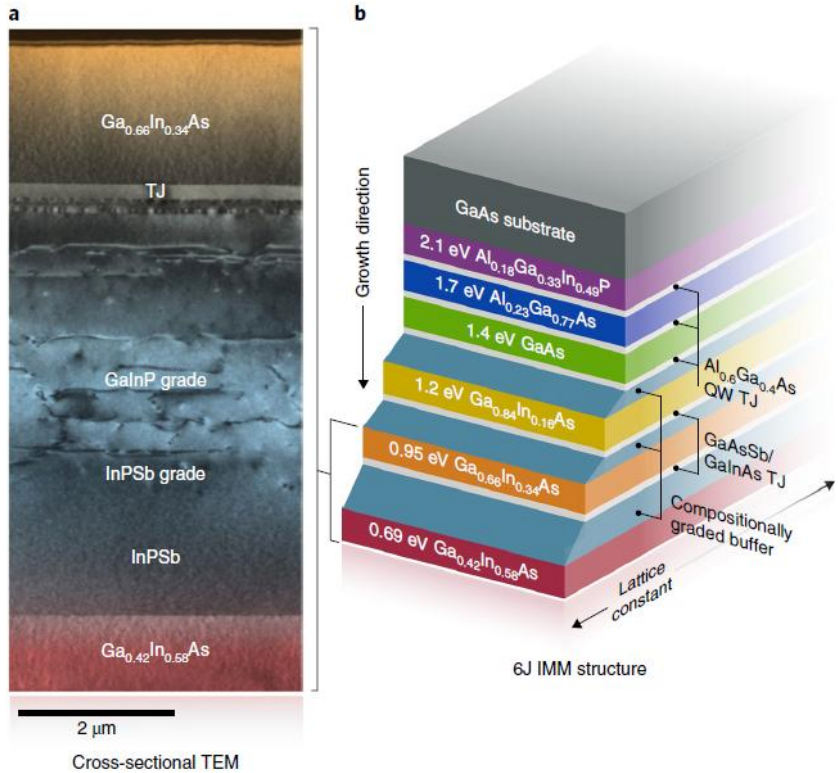
Rysunek 8. Przekrój badanych ogniw z i bez dodatkowej warstwy blokującej MoO_3 [26]

Prowadzone są także badania nad otrzymywaniem transparentnych ogniw CdTe, również z wykorzystaniem modyfikacji za pomocą $CdCl_2$ [27], a także badania nad samym procesem chlorowania CdTe [28]. Naukowcy prowadzą też prace nad nowymi metodami otrzymywania tych ogniw, np. z wykorzystaniem metody zol-żel [28]. Sama warstwa CdTe może być również modyfikowana różnymi domieszkami np. Cu [29] lub Se [30, 31].

2.4. GaAs

Arsenek galu (GaAs) jest jednym z najbardziej rozpowszechnionych półprzewodników III-V w zastosowaniach fotowoltaicznych. Wynika to z wielu czynników, głównie z wysokiej ruchliwości elektronów, prostej przerwy energetycznej oraz dobrze opanowanych mechanizmów wzrostu. Sprawność pojedynczego złącza GaAs może osiągać wartości bliskie 30%. Ogniwa słoneczne GaAs zostały dokładnie przebadane i szybko stały się układem odniesienia dla cienkowarstwowych ogniw słonecznych ze względu na ich dobre właściwości elektryczne, odporność na ciepło, wysoką sprawność i wydajność. Arsenek galu jest używany w procesie produkcji wielu zaawansowanych urządzeń, takich jak monolityczne mikrofalowe układy scalone, diody emitujące światło podczerwone, okna optyczne, diody laserowe i najważniejsze ogniwa słoneczne. GaAs jest stosowany jako materiał podłożowy do epitaksjalnego wzrostu innych półprzewodników III-V, w tym arsenku galu domieszkowanego aluminium, arsenku galu domieszkowanego indem i innych [32].

Ogniwa oparte na tym związku osiągają rekordowe sprawności w konfiguracjach wielozłączowych jak obecnie najwyższa wartość sprawności w warunkach standardowych 39,2% [8]. Domieszkowanie arsenku galu innymi pierwiastkami, takimi jak ind, glin, antymon oraz stosowanie związku InGaP pozwala na projektowanie i wytwarzanie wielozłączowych struktur o wysokiej sprawności konwersji energii. Sześciozłączowy układ warstw o obecnie najwyższej sprawności, został wykonany technologią organometalicznej epitaksji z fazy gazowej [7]. Budowa tego skomplikowanego układu została przedstawiona na rysunku 9.



Rysunek 9. Budowa wielozłączowego ogniwa o rekordowej sprawności [7]

2.5. Cienkowarstwowe ogniwa dostępne na rynku

W 2019 roku światowa produkcja energii z paneli fotowoltaicznych wynosiła między 129 a 137 GWp, z czego ok. 5,5% stanowiły ogniwa cienkowarstwowe, 29% ogniwa krzemowe polikrystaliczne i 65,5% ogniwa monokrystaliczne. Za udział w rynku ogniw cienkowarstwowych w 76% odpowiadają panele typu CdTe, 21% CIGS i zaledwie 3% (a więc zaledwie 0,016% całego rynku) to ogniwa krzemowe amorficzne [33].

Według [14] krzemowe krystaliczne ogniwa słoneczne dominują na rynku fotowoltaicznym, stanowiąc jego 92%, następnie plasują się ogniwa z tellurku kadmu (CdTe), zajmując 5%, selenku miedzi i galu (CIGS) – 2% i ogniwa z krzemu amorficznego (a-Si:H) – ok. 1%.

Wśród producentów paneli cienkowarstwowych można wymienić takie firmy, jak Solibro, Manz (Wurth), Solar Frontier, MiaSole dla ogniw typu CIGS oraz firmy First Solar i Calypso dla ogniw typu CdTe [33].

W Polsce wśród ofert instalacji fotowoltaicznych zdecydowanie dominują ogniwa krzemowe monokrystaliczne. Koszt samej instalacji zależy od rodzaju paneli i ich ilości. Cena zawiera również montaż i opodatkowanie.

Ogniwa oparte na GaAs są stosowane tylko w zastosowaniach kosmicznych ze względu na ich wysoką cenę. Ogniwa z krzemu amorficznego można kupić jako elementy innych produktów np. lampek ogrodowych czy zabawek.

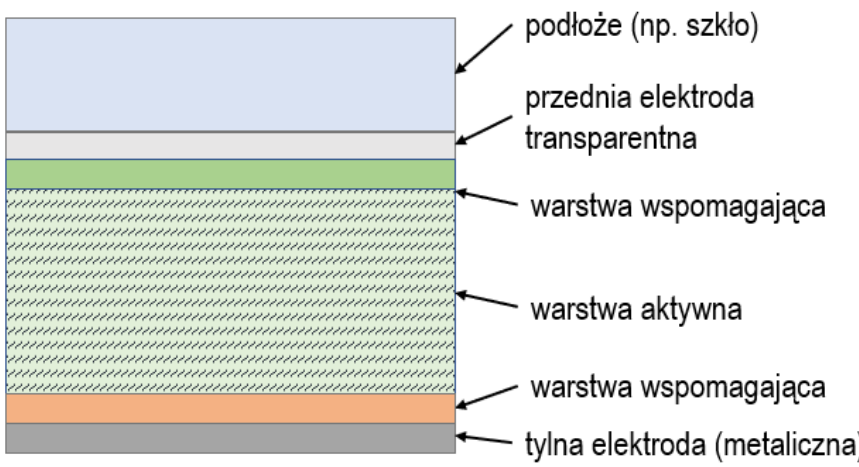
3. Najnowsza generacja ogniw fotowoltaicznych

Poza znanymi materiałami i technologiami pozyskiwania energii słonecznej wciąż poszukuje się nowych rozwiązań. W najmłodszej, trzeciej generacji ogniw fotowoltaicznych wyróżniamy ogniwa polimerowe (organiczne), barwnikowe, perowskitowe oraz inne np. tlenkowe. Do tych poszukiwań nowych rozwiązań motywują naukowców kosztowny proces produkcji ogniw pierwszej generacji, toksyczność kadmu stosowanego szeroko w ogniwach drugiej generacji, ograniczone zasoby pierwiastków, takich jak ind czy tellur, cel znalezienie nowych materiałów, których produkcja będzie możliwie prosta, a osiągi wysokie. Na uwagę zasługują zdecydowanie ogniwa perowskitowe, które bardzo szybko osiągnęły wysokie sprawności.

3.1. Ogniwa polimerowe

Organiczne ogniwa słoneczne i polimerowe ogniwa słoneczne zbudowane są z cienkich warstw (typowo 100 nm) organicznych półprzewodników, takich jak polimery i związki małocząsteczkowe. Stosuje się związki, takie jak pentacen, polifenylen winylenu, ftalocyjanina miedzi (niebieski lub zielony pigment organiczny) oraz fulereny węglowe. Wytwarzanie organicznych ogniw nie wymaga warunków próżniowych lub wysokotemperaturowych, co oznacza niskie koszty produkcji takich paneli [5].

Ogniwa polimerowe lub też inaczej nazywane organicznymi działają na nieco innej zasadzie niż ogniwa I i II generacji. W konstrukcji ogniwa nie pojawia się złącze p-n na granicy dwóch różnych materiałów, lecz procesy rozdzielania ładunków zachodzą pomiędzy cząsteczkami polimeru. Zatem pomiędzy elektrodami znajdują się cząsteczki będące donorami oraz akceptorami elektronów, a złącza półprzewodnikowe powstają pomiędzy tymi cząsteczkami. Taką strukturę nazywamy złączem objętościowym [5]. Budowa ogniwa polimerowego została przedstawiona na rysunku 10. Wykorzystywane polimery oczywiście muszą być przewodzące, aby ładunki mogły się w nich przemieszczać.



Rysunek 10. Budowa ogniwa organicznego [opracowanie własne]

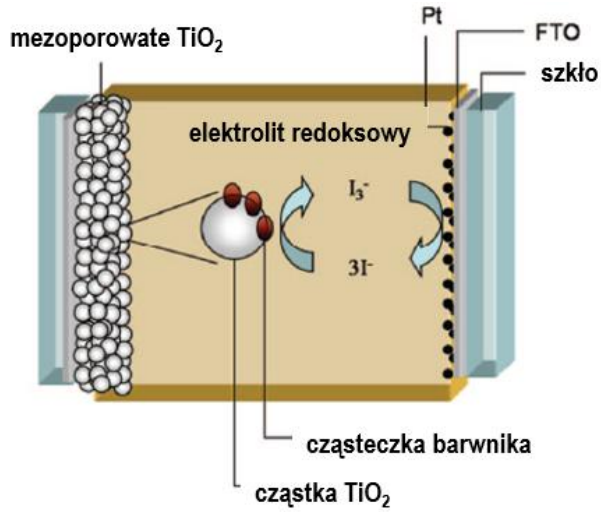
Działanie ogniwa organicznego składa się z kilku etapów podobnych jak w innych ogniwach: absorpcja fotonu i generacja ekscytonu, dyfuzja ekscytonu, rozdzieleni ładunku, transport ładunku, zbieranie ładunków na elektrodach. Jednak tutaj elektrony nie są wzbudzane do pasma przewodnictwa, o których można mówić w krystalicznych ciałach stałych. Elektron jest wzbudzany tak, że przechodzi z najwyżej obsadzonego orbitala molekularnego donora do najniżej nieobsadzonego orbitala molekularnego akceptora. Powstaje tym samym ekscyton, który dyfunduje aż do rozdzielenia ładunków na granicy cząstek. Takich granic w ogniwach organicznych jest znacznie więcej niż w konwencjonalnych rodzajach ogniw, więc ten etap jest bardziej wydajny. Niestety inne procesy zachodzące w całej strukturze ogniwa zmniejszają w bardzo dużym stopniu tę wydajność [5].

Ze względu na stosunkowo słabe oddziaływanie pomiędzy cząsteczkami organicznymi, półprzewodniki organiczne trudno formują uporządkowaną strukturę. Zaburzenia strukturalne są zazwyczaj nieuniknione w materiałach organicznych, dlatego są one słabo krystaliczne lub amorficzne i łatwo ulegają zginaniu i transformacji. Właściwości te prowadzą do łatwości przetwarzania w roztworze, elastyczności i sterowalności materiałów organicznych, które w ciągu ostatnich 30 lat wzbudziły ogromne zainteresowanie badaczy. Możliwym mechanizmem transportu w nieuporządkowanych ośrodkach organicznych jest hoping nośników, którego prędkość jest znacznie wolniejsza niż szybkość transportu nośników w kryształach nieorganicznych, przez co wykazują one małą ruchliwość. Gdy materiały organiczne tworzą cienkie warstwy, nie da się oczywiście uniknąć pułapek, na których nośniki mogą rekombinować, dodatkowo zmniejszając efektywność [35].

Pierwsze zastosowanie polimerów jako materiał fotowoltaiczny sięga lat 80., gdzie poliacetylen był tym związkiem. Sprawność złącza poliacetylenu z Al wyniosła wtedy zaledwie 0,3% [36]. Obecny rekord sprawności ogniwa organicznego wynosi 18,2% [8, 37]. Obecnie, organiczne ogniwa słoneczne z niefullerenowymi akceptorami elektronów zapewniają najwyższą wydajność spośród wszystkich badanych OSC [38, 39] W chemii organicznej istnieją ogromne możliwości w poszukiwaniu i tworzeniu nowych związków do różnych zastosowań i tak też dzieje się w przypadku ogniw fotowoltaicznych pozwalając na osiąganie sprawności 14,4% [40], 15,34% [41] bliskich wartości rekordowej oraz coraz bliższych sprawnościom ogniw generacji pierwszej i drugiej.

3.2. Ogniwa barwnikowe

Ogniwa barwnikowe powstały w 1991 roku i są nazywane ogniwami Grätzela od nazwiska jednego z twórców [42]. Ich działanie jest inspirowane barwnikami występującymi w naturze biorących udział w procesie fotosyntezy i opierają się na reakcjach typu redoks. Poglądowy rysunek zasady działania tego typu ogniwa zamieszczono na rysunku 11. Ogniwo barwnikowe składa się z elektrody z naniesioną warstwą półprzewodnika, zwykle TiO_2 w formie porowatej struktury lub nanocząstek, na której zaadsorbowane są cząsteczki barwnika. Pod wpływem promieniowania w barwniku dochodzi do wzbudzenia elektronów, które przenoszą się do pasma przewodnictwa półprzewodnika (TiO_2). Stamtąd elektron wędruje do elektrody i obwodu. Na przeciwnej elektrodzie (katodzie) zachodzi reakcja redukcji jonów elektrolitu (zwykle jony jodu $3I^{-1}$ oraz I^{-3}). Jony reagują z barwnikiem, ulegając utlenieniu i regenerując tym samym barwnik [5].



Rysunek 11. Schemat budowy i działania ogniw barwnikowych [43]

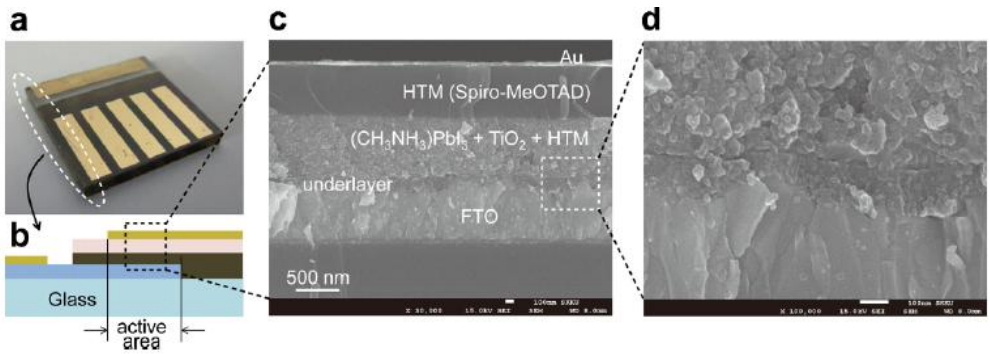
Obecnie rekord sprawności dla ogniw barwnikowych wynosi 13% [44]. Zostały one wytworzone przez naukowców z Ecole Polytechnique Federale de Lausanne (EPFL) w 2014 roku. Od tamtej pory nie udało się pobić tej wartości. Ten rodzaj ogniw posiada wiele zalet: są wytwarzane prostymi, niskokosztowymi metodami, nie wymagają stosowania toksycznych lub mało dostępnych materiałów, mogą być nanoszone na podłoża elastyczne, wykazywać transparentność lub różne kolory. Ponadto działają dobrze przy niskim natężeniu promieniowania, np. w pomieszczeniach gdzie światło jest rozproszone [45]. Do szerokiego komercyjne zastosowania ogniwom barwnikowym brakuje wydajności oraz stabilności [46]. Choć ogniwa barwnikowe nie dają tak dobrych sprawności jak polimerowe czy perowskitowe, to również nad nimi prowadzone są badania, np. nad modyfikacją warstw TiO_2 [47, 48] lub wykorzystywaniem innych półprzewodników [49] i projektowaniem nowych barwników [50]. Ogniwa perowskitowe pośrednio zostały wynalezione poprzez prace nad udoskonalaniem ogniw barwnikowych.

3.3. Ogniwa perowskitowe

Perowskitowe ogniwa słoneczne wywodzą się z ogniw barwnikowych, lecz mają bardzo obiecującą strukturę półprzewodnikową i wykazują szybki wzrost wydajności. Struktura perowskitu (pokazana na rysunku 12 a) pozwala na tworzenie molekularnych kompozytów organiczno-nieorganicznych. Ze względu na ich wysoki współczynnik absorpcji optycznej, przestrajalne właściwości, długą drogę dyfuzji nośników ładunku, i co najbardziej zdumiewające, ich zdolność do tolerowania ogromnych gęstości defektów strukturalnych z minimalnymi konsekwencjami elektronicznymi stanowią bardzo dobry materiał do zastosowań fotowoltaicznych. Natomiast ze względu na ich wysoką ruchliwość nośników ładunku, w tego typu ogniwach słonecznych stosowane są organiczno-nieorganiczne materiały perowskitowe. Wysoka ruchliwość jest ważna, ponieważ wraz z wysokimi czasami życia nośników ładunku oznacza, że generowane przez światło elektrony i dziury mogą przemieszczać się na wystarczająco duże odległości, aby zostać pobrane w postaci prądu [51]. Ponadto, perowskitowe ogniwa słoneczne mogą być osadzane metodami niskotemperaturowymi, takimi jak metody roz-

tworowe, nanoszenie wirówkowe (spin-coating), osadzanie natryskowe i metody odparowywania termicznego [21].

Pierwsze ogniwa oparte na perowskitach metalohalogenkowych zostały opracowane w 2012 roku. Dwie grupy badawcze opracowały ogniwa o sprawności 9,7% [52] i 10,9% [53]. W pierwszej z nich wykorzystano perowskit $(CH_3NH_3)PbI_3$ jako elektrod użyto złota i tlenku przewodzącego FTO. Perowskit został otrzymany w roztworze i naniesiony na mezoporowatą warstwę TiO_2. Zdjęcia gotowego ogniwa są przedstawione na rysunku 12.

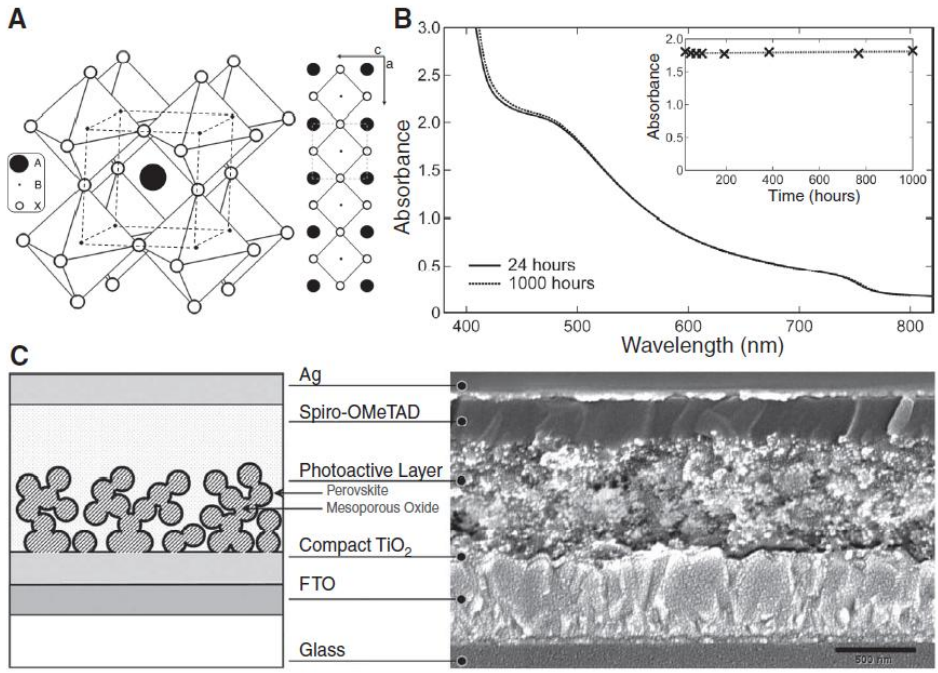


Rysunek 12. Urządzenie półprzewodnikowe i jego przekrojowa mezo-struktura. (a) Rzeczywiste urządzenie półprzewodnikowe. (b) Struktura przekroju poprzecznego urządzenia. (c) Obraz SEM przekroju poprzecznego SEM urządzenia. (d) Struktura złącza międzyfazowego warstwa aktywna – podłoże – FTO [52]

Druga grupa wykorzystała perowskit $(CH_3NH_3)PbI_2Cl$, który został naniesiony na rusztowanie z Al_2O_3. Elektrodami były szkło z warstwą ITO i TiO_2, półprzewodnik typu n stanowił związek 2,2´,7,7´-tetrakis-(N,Ndi-p-methksyfenyloamina)9,9´-spirobifluoren (określany jako spiro-OMeTAD). Wygląd wytworzonego ogniwa przedstawiono na rysunku 13c.

Od kiedy pojawiły się ogniwa oparte na perowskitach ich wzrost sprawności jest bardzo dynamiczny. W roku 2012 było to około 10%, a teraz rekordową wartość stanowi 25,5% [54]. To ponad 15% różnicy w sprawności w ciągu zaledwie dekady [8]!

Ogniwa perowskitowe mają dwie podstawowe wady: wykorzystanie ołowiu, który jest pierwiastkiem toksycznym oraz ich słaba stabilność [55]. Obecnie tworzone ogniwa perowskitowe mogą pracować stabilnie nie dłużej niż dwa lata, co jest bardzo krótkim okresem w porównaniu do czasu pracy ogniw krzemowych, który wynosi 25, a nawet 30 lat. Pracuje się właśnie nad poszukiwaniem perowskitów o lepszej stabilności, np. nieorganiczny $CsPbBr_3$ [56, 57], wykorzystaniem perowskitów w formie nanokrystalicznej jako kropek kwantowych [49], badaniem samego procesu degradacji ogniw perowskitowych [58] oraz nad otrzymywaniem ogniw [59] także na elastycznych podłożach [60]. Stosowanie perowskitów niezawierających toksycznego ołowiu jest ważnym celem, w 2020 roku badaczom udało się otrzymać ogniwo o sprawności 13% oparte na perowskicie organiczno-nieorganicznym wykorzystującym cynę zamiast ołowiu [61, 62]. W ciągu krótkiego czasu udało się osiągnąć bardzo wiele w dziedzinie fotowoltaiki perowskitowej, lecz czeka badaczy jeszcze dużo pracy nad uzyskaniem ogniw stabilnych i zabezpieczonych przed wpływem wilgoci i tlenu, zanim ta technologia będzie skomercjalizowana.



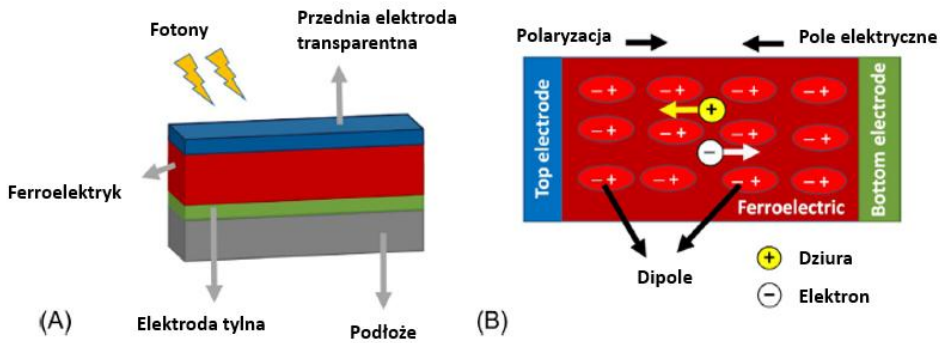
Rysunek 13. A) Po lewej: Trójwymiarowe schematyczne przedstawienie struktury perowskitu ABX_3 ($A = CH_3NH_3$, B = Pb, a X = Cl, I). Po prawej: dwuwymiarowy schemat ilustrujący komórkę elementarną perowskitu. B) Widma absorbancji od ultrafioletu do widzialnego (UV-Vis) warstwy fotoaktywnej oraz gęstość optyczna przedstawiona we wstawce. C) Po lewej: Schematyczne przedstawienie pełnej struktury urządzenia, w którym znajduje się również mezoporowaty tlenek Al_2O_3 lub anataz TiO_2. Po prawej: Przekrójowy obraz SEM pełnego urządzenia zawierającego mezoporowaty Al_2O_3. Skala, 500 nm [53]

3.4. Ogniwa tlenkowe

W poszukiwaniach nowych materiałów dla zastosowań w fotowoltaice część badaczy skupia się na tlenkach. Przykładowo własności światłoczułe tlenków miedzi zostały odkryte już w 1904 roku. Do tej pory najwyższa osiągnięta sprawność ogniwa wykorzystującego tlenek miedzi Cu_2O wynosi 8,1% i została osiągnięta dzięki dokładnemu doborowi warstwy typu n oraz domieszkowaniu tlenku miedzi sodem [63]. Wiele tlenków wykazuje własności przewodzące, jak np. tlenek indowo-cynowy i tlenek fluorowo-cynowy, które są szeroko wykorzystywane jako transparentne elektrody w ogniwach innych rodzajów. Tlenki są stosowane w ogniwach słonecznych jako izolatory, półprzewodniki, elektrody transparentne zapewniając niższy koszt i większą stabilność urządzenia. Zastosowanie tlenków daje możliwości spełnienia nietoksyczności, dostępności i trwałości materiałów [64]. Jednym z najczęściej stosowanych tlenków jako aktywne fotowoltaicznie są tlenki miedzi, zwłaszcza Cu_2O [65-67]. Przykładem jest zastosowanie układu nanodrutów Cu_2O typu p pokrytych warstwą ZnO typu n. Takie ogniwo osiągnęło tylko 0,39% sprawnością, ale pokazuje możliwości tworzenia ogniw opartych tylko na tlenkach [68]. Oprócz tlenków miedzi jako warstwę absorbera w ogniwach stosuje się także Co_3O_4 czy Fe_2O_3 [69].

Efekt fotowoltaiczny może być też wykorzystywany dzięki ferroelektrycznym właściwościom materiałów. Kilka tlenków jest badanych właśnie w kierunku budowy tego typu ogniw. W przypadku ferroelektrycznych ogniw fotowoltaicznych nie ma po-

trzeby stosowania złącza p-n, a separacja ładunków może pochodzić z polaryzacji wynikającej z ferroelektryczności (rys. 14). Fotoprąd jest napędzany przez wewnętrzne pole elektryczne indukowane polaryzacją ferroelektryczną. W rzeczywistości, odpowiedni dobór elektrod w zależności od ich funkcji roboczych może ułatwić uzyskanie fotoprądu. Takie tlenki to Bi_2FeCrO_6, $BiMnO_3$, $BiFeO_3$, $TbMnO_3$ [70].



Rysunek 14. Zasada działania ogniwa ferroelektrycznego [71]

4. Podsumowanie

Pozyskiwanie energii elektrycznej z odnawialnych źródeł stanowi naszą przyszłość. Ogniwa fotowoltaiczne pracują bez hałasu i wytwarzania odpadów czy gazów cieplarnianych, więc stanowią idealną technologię dla naszej cywilizacji. Należy jednak zwracać uwagę na proces produkcji samych ogniw, na jego koszt, dostępność materiałów i ich toksyczność. Choć na całym świecie jest coraz więcej instalacji fotowoltaicznych, to wciąż konieczny jest rozwój tej dziedziny w poszukiwaniu nowych, lepszych rozwiązań. W ciągu ostatnich kilkudziesięciu lat pojawiły się nowe pomysły na realizację ogniw fotowoltaicznych: organiczne, barwnikowe, tlenkowe czy perowskitowe, które mają ogromny potencjał. Ich ogromną zaletą są możliwości wykorzystywania prostych metod produkcji i nanoszenia na różne podłoża, co znacznie pozwala na rozszerzenia zastosowania ogniw od instalacji zakładanych na ziemi czy dachach, do zintegrowanych z budynkami czy nawet tekstyliami lub pojazdami. Fotowoltaika to bez wątpienia nasza przyszłość, a w XXI będziemy obserwować jej rozwój i zwłaszcza rozwój tych najnowszych technologii.

Literatura

1. Zandalinas I., Fritschi F.B., Mittler R., *Global Warming, Climate Change, and Environmental Pollution: Recipe for a Multifactorial Stress Combination Disaster*, Trends in Plant Science, tom 26, n. 6, 2021, s. 588-599.
2. Letcher T.M., *Introduction with a focus on atmospheric carbon dioxide and climate change*, Future Energy, 2020, s. 3-17.
3. Sönnichsen N., *Projected electricity generation worldwide from 2018 to 2050, by energy source*, 2019, https://www.statista.com/statistics/238610/projected-world-electricity-generation-by-energy-source.
4. El Chaar L., Lamont L.A., El Zein N., *Review of photovoltaic technologies, Renewable and Sustainable Energy Reviews*, tom 15, n. 5, 2011, s. 2165-2175.

5. Marszałek K., Lewińska G., Dyndał K., *Fotowoltaika*, Akademia Górniczo-Hutnicza im. St. Satszica w Krakowie, 2021.
6. Shockley W., Queisser H.J., *Detailed balance limit of efficiency of p-n junction solar cells*, Journal of Applied Physics, t. 32, n. 3, 1961, s. 510-519.
7. Geisz J.F., France R.M., Schulte K.L., Steiner M.A., Norman A.G., Guthrey H.L., Young M.R., Song T., Moriarty T., *Six-junction III-V solar cells with 47,1% conversion efficiency under 143 Suns concentration*, Nature Energy, t. 5, n. 4, 2020, s. 326-335.
8. National Renewable Energy Laboratory, Best Research-Cell Efficiency Chart, 2019, s. 1, https://www.nrel.gov/pv/assets/pdfs/best-research-cell-efficiencies.20190802.pdf.
9. Biernaciak E., *Sprawność paneli fotowoltaicznych 2021 – co warto wiedzieć?*, 2021, https://enerad.pl/aktualnosci/sprawnosc-paneli-fotowoltaicznych-2021-co-warto-wiedziec/.
10. Marszałek K., Małek A., Winkowski P., *The GdF3/MgF2 bilayer as an antireflective narrow-band ultraviolet filter*, Optica Applicata, t. 46, n. 2, 2016, s. 187-197.
11. Mishima T., Taguchi M., Sakata H., Maruyama E., *Development status of high-efficiency HIT solar cells*, Solar Energy Materials & Solar Cells, t. 95, n. 1, 2011, s. 18-21.
12. Andreani L.C., Bozzola A., Kowalczewski P., Liscidini M., Redorici L., *Silicon solar cells: Toward the efficiency limits*, Advances in Physics: X, t. 4, n. 1, 2019, s. 125-147.
13. https://commons.wikimedia.org/wiki/File:ShockleyQueisserFullCurve.svg.
14. Ramanujama J., Bishop D.M., Todorov T.K., Gunawanb O., Rath J., Nekovei R., Artegiani E., Romeo A., *Flexible CIGS, CdTe and a-Si:H based thin film solar cells: A review*, Progress in Materials Science, t. 110, 2020, s. 1-20.
15. Carlson D.E., Wronski C.R., *Amorphous silicon solar cell*, Applied Physics Letters, t. 28, n. 11, 1976, s. 671-673.
16. Sai H., Matsui T., Matsubara K., *Stabilized 14.0%-efficient triple-junction thin-film silicon solar cell*, Applied Physics Letters, t.109, n. 18, 2016, s. 1-6.
17. Stuckelberger M., Biron R., Wyrsch N., Haug F. J., Ballif C., *Review: Progress in solar cells from hydrogenated amorphous silicon*, Renewable and Sustainable Energy Reviews, t. 76, wrzesień 2017, s. 1497-1523.
18. Green M.A., Dunlop E.D., Hohl-Ebinger J., Yoshita M., Kopidakis N., Hao X., *Solar cell efficiency tables (version 56)*, Progress in Photovoltaics, t. 28, n. 7, 2020, s. 629-638.
19. Misra P., Atchuta S.R., Mandati S., Sarada B.V., Rao T.N., Sakthivel S., *A non-vacuum dip coated SiO2 interface layer for fabricating CIGS solar cells on stainless steel foil substrates*, Solar Energy, t. 214, 2021, s. 471-477.
20. Liu F., Zeng Q., Li J., Hao X., Ho-Baillie A., Tang J., Green M.A., *Emerging inorganic compound thin film photovoltaic materials: Progress, challenges and strategies*, Materials Today, t. 41, grudzień 2020, s. 120-142.
21. Sundaram S., Shanks K., Upadhyaya H., *A comprehensive guide to solar energy systems: 18 – Thin Film Photovoltaics*, Academic Press Elsevier Inc., 2018, s. 361-370.
22. Candelise C., Spiers J.F., Gross R.J.K., *Materials availability for thin film (TF) PV technologies development: A real concern?*, Renewable and Sustainable Energy Reviews, t. 15, n. 9, 2011, s. 4972-4981.
23. Nakamura M., Tada K., Kinoshita T., Bessho T., Nishiyama Ch., Takenaka I., Kimoto Y., Higashino Y., Sugimoto H., Segawa H., *Perovskite/CIGS Spectral Splitting Double Junction Solar Cell with 28% Power Conversion Efficiency*, iScience, t. 23, n. 12, 2020, 101817.
24. Gharibshahian I., Abbasi A., Orouji A.A., *Modeling of GaAsxP1-x/CIGS tandem solar cells under stress conditions*, Superlattices and Microstructures, t. 153, listopad 2020, 106860.

25. Hu A., Zhou J., Zhong P., Qin X., Zhang M., Jiang Y., Wub X., Yang D., *High-efficiency CdTe-based thin-film solar cells with unltrathin CdS:O window layer and processes with post annealing*, Solar Energy, t. 214, sierpień 2020, s. 319-325.
26. Singh R., Dutta A., Basu N., Lahiri J., Som T., *Improved photovoltaic performance of CdTe-based solar cells: Roles of using a hole-blocking layer and nanoscale imaging of barrier height at interfaces*, Solar Energy, t. 215, styczeń 2021, s. 1-11.
27. He F., Li J., Lin S., Long W., Wu L., Hao X., Zhang J., Feng L., *Semitransparent CdTe solar cells with CdCl2 treated absorber towards the enhanced photovoltaic conversion efficiency*, Solar Energy, t. 214, listopad 2020, s. 196-204.
28. Kim S., Song J., Kim D., Hong J., Joon Cho I., Hyub Kim Y., Jeong J., Sun Yoon M., Ahn S., Chung W., Namb T., Cho S., *Effect of novel double treatment on the properties of CdTe solar cells*, Energy Reports, t. 7, 2021, s. 1396-1403.
29. Moon M.M.A., Rahman M.F., Kamruzzaman M., Hossain J., Ismail A.B.M., *Unveiling the prospect of a novel chemical route for synthesizing solution-processed CdS/CdTe thin-film solar cells*, Energy Reports, t. 7, 2021, s. 1742-1756.
30. Wang T., Zhu X., Li W., Zhang J., Wang W., *Incorporation of metal selenide thin films as the secondary absorber in the CdTe solar cells*, Journal of Alloys and Compounds, t. 828, 2020, 154415.
31. Artegiani E., Gasparotto A., Punathil P., Kumar V., Barbato M., Meneghini M., Meneghesso G., Piccinelli F., Romeo A., *A new method for CdSexTe1-x band grading for high efficiency thin-absorber CdTe solar cells*, Solar Energy Materials and Solar Cells, t. 226, wrzesień 2020, 111081.
32. https://sinovoltaics.com/learning-center/solar-cells/gaas-solar-cells/.
33. Burger B., Kiefer K., Kost C., Nold S., Philipps S., Preu R., Rentsch J., Schlegl T., Stryi-Hipp G., Wirth H., Warmuth W., *Photovoltaics report*, Fraunhofer Institute for Solar Energy Systems, wrzesień 2020.
34. Dimmler B., *Overview of Thin-Film Solar Cell Technologies*, 1 ed., t. 90, Elsevier Inc., 2014.
35. Mao P., Wei Y., Li H., Wang J., *Junction diodes in organic solar cells*, Nano Energy, tom 41, październik 2017, s. 717-730.
36. Weinberger B.R., Akhtar M., Gau S.C., *Polyacetylene photovoltaic devices*, Synthetic Metals, t. 4, n. 3, 1982, s. 187-197.
37. Li Z., Ying L., Zhu P., Zhong W., Li N., Liu F., Huang F., Cao Y., *A generic green solvent concept boosting the power conversion efficiency of all-polymer solar cells to 11%*, Energy & Environmental Science, t. 12, n. 1, 2019, s. 157-163.
38. Mehboob M.Y., Khan M.U., Hussain R., Hussain R., Ayub K., Sattar A., Ahmad M.K., Irshad Z., Saira, Adnan M., *Designing of benzodithiophene core-based small molecular acceptors for efficient non-fullerene organic solar cells*, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, t. 244, 2021, s. 118873.
39. Ma X., Wang J., Gao J., Hu Z., Xu C., Zhang X., Zhang F., *Achieving 17.4% Efficiency of Ternary Organic Photovoltaics with Two Well-Compatible Nonfullerene Acceptors for Minimizing Energy Loss*, Advanced Energy Materials, t. 10, n. 31, 2020, s. 1-9.
40. Jia T., Zhang J., Zhong W., Liang Y., Zhang K., Dong S., Ying L., Liu F., Wang X., Huang F., Cao Y., *14,4% Efficiency All-Polymer Solar Cell With Broad Absorption and Low Energy Loss Enabled By a Novel Polymer Acceptor*, Nano Energy, t. 72, luty 2020, 104718.
41. Hu D., Yang Q., Chen H., Wobben F., Le Corre V.M., Singh R., Liu T., Ma R., Tang H., Jan Anton Koster L., Duan T., Yan H., Kan Z., Xiao Z., Lu S., *15,34% Efficiency All-Small-Molecule Organic Solar Cells With an Improved Fill Factor Enabled By a Fullerene Additive*, Energy & Environmental Science, t. 13, n. 7, 2020, s. 2134-2141.

42. O'Regan B., Graetzel M., *A low-cost, high-efficiency solar cell based on dye-sensiized colloidal TiO_2 films*, Nature, t. 353, 1991, s. 737-739.
43. Hagfeldt A., Boschloo G., Sun L., Kloo L., Pettersson H., *Dye-Sensitized Solar Cells*, Chemical Reviews, t. 110, n. 11, 2010, s. 6595-6663.
44. Mathew S., Yella A., Gao P., Humphry-Baker R., Curchod B.F.E., Ashari-Astani N., Tavernelli I., Rothlisberger U., Md. Khaja Nazeeruddin, Graetzel M., *Dye-sensitized solar cells with 13% efficiency achieved through the molecular engineering of porphyrin sensitizers*, Nature Chemistry, t. 6, n. 3, 2014, s. 242-247.
45. Aslam A., Mehmood U., Arshad M.H., Ishfaq A., Zaheer J., Ul Haq Khan A., Sufyan M., *Dye-sensitized solar cells (DSSCs) as a potential photovoltaic technology for the self-powered internet of things (IoTs) applications*, Solar Energy, t. 207, maj 2020, s. 874-892.
46. Omar A., Ali M.S., Abd Rahim N., *Electron transport properties analysis of titanium dioxide dye-sensitized solar cells (TiO_2-DSSCs) based natural dyes using electrochemical impedance spectroscopy concept: A review*, Solar Energy, t. 207, kwiecień 2020, s. 1088-1121.
47. Wang J., Nie X., Wang W., Zhao Z., Li L., Zhang Z., *Single-layer graphene-TiO2 nanotubes array heterojunction as photoanode to enhance the photoelectric of DSSCs*, Optik – International Journal for Light and Electron Optics, t. 242, maj 2021, 167245.
48. Megat Hasnan M.M.I., Nayan N., Hamed N.K.A., Azman Z., Ahmad M.K., Mohamed Ali M. S., Mohd Yusop M.Z., Noor I.M., *Improvement of facile hydrothermal TiO2 rutile nanorod-flower using second HiPIMS deposition for DSSC performance enhancement*, Optical Materials, t. 117, maj 2021, 111149.
49. Ramesh K., Gnanavel B., *Fabrication and characterization of RGO/WO3 nanocomposites based working electrode for dye-sensitized solar cells (DSSCs)*, Materials Today Proceedings, 2021.
50. Ilgün C., Sevim A.M., Cakar S., Özacar M., Gül A., *Novel Co and Zn-Phthalocyanine dyes with octa-carboxylic acid substituents for DSSCs*, Solar Energy, t. 218, styczeń 2021, s. 169-179.
51. Yuan J., Hazarika A., Zhao Q., Ling X., Moot T., Ma W., Luther J.M., *Metal Halide Perovskites in Quantum Dot Solar Cells: Progress and Prospects*, Joule, t. 4, n. 6, 2020, s. 1160-1185.
52. Kim H.S., Lee C.R. , Im J.H., Lee K.B., Moehl T., Marchioro A., Moon S.J., Humphry-Baker R., Yum J.H., Moser J.E., Graetzel M., Park N.G., *Lead iodide perovskite sensitized all-solid-state submicron thin film mesoscopic solar cell with efficiency exceeding 9%*, Scientific Reports, t. 2, 2012, s. 1-7.
53. Lee M.M., Teuscher J., Miyasaka T., Murakami T.N., Snaith H.J., *Efficient hybrid solar cells based on meso-superdtructures organometal halide perovskites*, Science, t. 338, 2012, s. 643-647.
54. Jung E.H., Jeon N.J., Park E.Y., Moon C.S., Shin T.J., Yang T.Y., Noh J.H., Seo J., *Efficient, stable and scalable perovskite solar cells using poly(3-hexylthiophene)*, Nature, t. 567, n. 7749, 2019, s. 511-515.
55. Wang D., Wright M., Elumalai N.K., Uddin A., *Stability of perovskite solar cells*, Solar Energy Materials & Solar Cells, t. 147, 2016, s. 255-275.
56. Lan H., Xiao H., Zhao J., Chen X., Fan P., Liang G., *All-inorganic CsPbBr3 thin-film solar cells prepared by single-source physical vapor deposition*, Materials Science in Semiconductor Processing, t. 132, luty 2021, 105869.
57. Li Z., Zhou F., Wang Q., Ding L., Jin Z., *Approaches for thermodynamically stabilized CsPbI3 solar cells*, Nano Energy, t. 71, listopad 2019, 104634.
58. Lim J., Kim M., Park H.H., Jung H., Lim S., Hao X., Choi E., Park S., Lee M., Liu Z., Green M.A., Seo J., Park J., Yun J.S., *Kinetics of light-induced degradation in semi-*

transparent perovskite solar cells, Solar Energy Materials and Solar Cells, t. 219, wrzesień 2020, 110776.
59. Ryu S., Seo J.W., Shin S.S., Kim Y.C., Jeon N.J., Noh J.H., Seok S.I., *Fabrication of metal-oxide-free CH3NH3PbI3 perovskite solar cells processed at low temperature*, Journal of Materials Chemistry A, t. 3, n. 7, 2015, s. 3271-3275.
60. Lei Y., Chen Y., Zhang R., Li Y., Yan Q., Lee S., Yu Y., Tsai H., Choi W., Wang K., Luo Y., Gu Y., Zheng X., Wang C., Wang C., Hu H., Li Y., Qi B., Lin M., Zhang Z., Dayeh S.A., Pharr M., Fenning D.P., Lo Y.H., Luo Y., Yang K., Yoo J., Nie W., Xu S., *A fabrication process for flexible single-crystal perovskite devices*, Nature, t. 583, n. 7818, 2020, s. 790-795.
61. Nishimura K., Kamarudin M.A., Hirotani D., Hamada K., Shen Q., Iikubo S., Minemoto T., Yoshino K., Hayase S., *Lead-free tin-halide perovskite solar cells with 13% efficiency*, Nano Energy, t. 74, styczeń 2020, 104858.
62. Mahajan P., Datt R., Chung Tsoi W., Gupta V., Tomar A., Arya S., *Recent progress, fabrication challenges and stability issues of lead-free tin-based perovskite thin films in the field of photovoltaics*, Coordination Chemistry Reviews, t. 429, 2021, 213633.
63. Minami T., Nishi Y., Miyata T., *Efficiency enhancement using a Zn1-xGex-O thin film as an n-type window layer in Cu_2O-based heterojunction solar cells*, Applied Physics Express, t. 9, n. 5, 2016.
64. Kohan M.G., Concina I., Vomiero A., *All-oxide solar cells*, Solar Cells and Light Management, 2019, s. 229-246.
65. Minami T., Nishi Y., Miyata T., *Heterojunction solar cell with 6% efficiency based on an n-type aluminum-gallium-oxide thin film and p-type sodium-doped Cu_2O sheet*, Applied Physics Express, t. 8, n. 2, 2015, 022301.
66. Kaura J., Bethge O., Wibowo R.A., Bansal N., Bauch M., Hamid R., Bertagnolli E., Dimopoulos T., *All-oxide solar cells based on electrodeposited Cu2O absorber and atomic layer deposited ZnMgO on precious-metal-free electrode*, Solar Energy Materials and Solar Cells, t. 161, grudzień 2016, s. 449-459.
67. Lee G.C., Jung Y.S., Kim K.H., *Properties of Cu_2O thin films for all-oxide solar cells*, Molecular Crystals and Liquid Crystals, tom 598, n. 1, 2014, s. 62-68.
68. Brittman S., Yoo Y., Dasgupta N.P., Kim S.I., Kim B., Yang P., *Epitaxially aligned cuprous oxide nanowires for all-oxide, single-wire solar cells*, Nano Letters, t. 14, n. 8, 2014, s. 4665-4670.
69. Dimopoulos T., *All-Oxide Solar Cells, The Future of Semiconductor Oxides in Next – Generation Solar Cells*, 2018, s. 439-480.
70. Pérez-Tomás A., Mingorance A., Tanenbaum D., Lira-Cantú M., *Metal Oxides in Photovoltaics: All-Oxide, Ferroic, and Perovskite Solar Cells, The Future of Semiconductor Oxides in Next-Generation Solar Cells*, 2018 s. 267-355.
71. Fix T., *Oxide and Ferroelectric Solar Cells*, Advanced Micro- and Nanomaterials for Photovoltaics, 2019.

Pozyskiwanie energii słonecznej w XXI wieku – bezkrzemowe, cienkowarstwowe ogniwa fotowoltaiczne

Streszczenie

Począwszy od pierwszego światowego kryzysu energetycznego w latach 70. XX wieku obserwujemy stały rozwój badań i zastosowań alternatywnych źródeł energii. W tej pracy przedstawione zostały dotychczasowe technologie stosowane w dziedzinie pozyskiwania energii elektrycznej z wykorzystaniem efektu fotowoltaicznego. Zaczynając od najlepiej poznanych krzemowych ogniw fotowoltaicznych do obecnie badanych i rozwijanych ogniw kolejnych generacji. Obecnie największą część rynku zajmują krzemowe monokrystaliczne ogniwa fotowoltaiczne, których cena w ciągu ostatnich kilkudziesięciu lat zmalała w ogromnym stopniu od 76$ za W w 1977 roku do 0,19$ w roku 2020, dzięki czemu są tak popularne.

Wykonane są z krzemu o wysokiej czystości, którego produkcja jest kosztownym i skomplikowanym procesem, w którym część materiału jest marnowana podczas wycinania wafli krzemowych. Także produkcja wafli z krzemu polikrystalicznego, mimo że tańsza, jest procesem wymagającym znacznych ilości energii elektrycznej. Wafle krzemowe stosowane obecnie w fotowoltaice mają grubość 180 mikrometrów. Zastąpienie tych materiałów układami cienkowarstwowymi, których grubość wynosi ok. 2 mikrometrów stanowi znaczące zmniejszenie masy instalacji fotowoltaicznej oraz zużycia surowców. Dodatkowo wykorzystywanie cienkich warstw umożliwia tworzenie modułów na elastycznych podłożach, rozszerzając znacznie obszary zastosowań ogniw fotowoltaicznych. Wśród cienkowarstwowych ogniw fotowoltaicznych dobrze znamy, takie jak CIGS, CdTe czy a-Si, których efektywności konwersji energii kiedyś była znacząco niższa od ogniw krzemowych. Obecnie uzyskiwane są już rezultaty zbliżone do wartości dla ogniw krzemowych. Ważne jest opracowanie technologii wykorzystujących nietoksyczne i łatwo dostępne materiały. W tym celu pracuje się nad nowymi technologiami, takimi jak ogniwa polimerowe, barwnikowe, perowskitowe czy tlenkowe. W każdej z nich najistotniejsza jest praca nad otrzymaniem wysokiej sprawności ogniwa oraz trwałości, co często stanowi problem dla wymienionych materiałów. Układy cienkowarstwowe z takich materiałów są nanoszone z wykorzystaniem technologii niskotemperaturowych, np. tzw. technik mokrych, tańszych niż konwencjonalne wytwarzanie kryształów krzemowych metodą Czochralskiego. Przegląd zarówno podstawowej, jak i najnowszej literatury specjalistycznej pozwolił na opisanie obecnego stanu rozwoju fotowoltaiki bezkrzemowej oraz możliwości jej zastosowań na złożonych powierzchniach, takich jak środki transportu, fasady budynków czy w przestrzeni kosmicznej. Badania w tej dziedzinie są szczególnie istotne w XXI wieku, kiedy to stosowanie odnawialnych źródeł energii staje się koniecznością, w obliczu zachodzących zmian klimatycznych oraz nadchodzącym wyczerpywaniem złóż surowców kopalnianych.
Słowa kluczowe: fotowoltaika, ogniwa bezkrzemowe, nowoczesne materiały

Solar energy harvesting in 21st century – non-silicon thin film photovoltaic modules

Abstract

Since the first global energy crisis in the 1970s, there has been a steady development of research and applications of alternative energy sources. In this paper we present current technologies used in the field of photovoltaic electricity generation. Starting from the best known silicon photovoltaic cells to currently researched and developed next generation cells. Today, silicon monocrystalline photovoltaic cells occupy the largest share of the market, and their price has dropped dramatically over the past few decades from $76 per W in 1977 to $0.19 in 2020, making them so popular. They are made of high-purity silicon, the production of which is an expensive and complicated process, where some of the material is wasted when cutting silicon wafers. Also, the production of wafers from polycrystalline silicon, although cheaper, is a process that requires significant amounts of electricity. Silicon wafers currently used in photovoltaics are 180 micrometers thick. Replacing these materials with thin-film systems that are about 2 micrometers thick represents a significant reduction in the weight of the photovoltaic system and in the consumption of raw materials. In addition, the use of thin films makes it possible to create modules on flexible substrates, greatly expanding the application area of photovoltaic cells. Among thin-film photovoltaic cells we know well such as CIGS, CdTe or a-Si, whose energy conversion efficiency used to be significantly lower than crystalline silicon cells. Nowadays, results close to the values for monocrystalline silicon cells are already obtained. It is important to develop technologies using nontoxic and readily available materials. To this end, new technologies such as polymer, dye-sensitized, perovskite or oxide cells are being worked on. In each of them the most important thing is to work on obtaining high cell efficiency and durability, which is often a problem for the mentioned materials. Thin-film systems of such materials are deposited using low-temperature technologies, e.g. so-called wet techniques, cheaper than the conventional Czochralski method for silicon crystal fabrication. A review of both basic and recent specialized literature allowed us to describe the current state of development of silicon-free photovoltaics and the possibilities of its applications on complex surfaces such as means of transport, building facades or in space. Research in this field is particularly important in the 21st century, when the use of renewable energy sources becomes a necessity in the face of climate change and the imminent depletion of fossil resources.
Keywords: photovoltaics, non-silicon solar cells, oxide solar cells, novel materials

Bartosz Partyka[1], Piotr Zwierzykowski[2]

Analiza możliwości zastosowania mechanizmów uczenia maszynowego w interpretacji danych zbieranych przez systemy monitoringu sieci i jakości usług ISP na przykładzie INEA

1. Wstęp

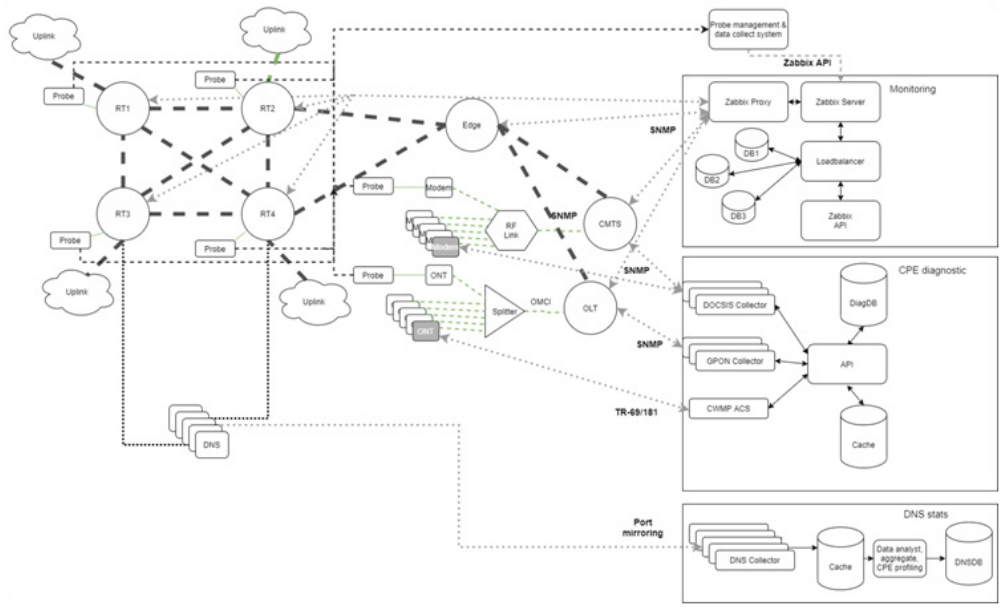
Na jakość usługi dostępu do Internetu, która jest świadczona klientowi przez operatora telekomunikacyjnego, ma wpływ działanie wielu systemów realizujących wewnętrze usługi oraz urządzeń sieciowych zapewniających komunikację. Działanie każdego z elementów począwszy od punktów styku z innymi operatorami, przez sieć szkieletową operatora, sieć dostępową i systemy realizujące usługi wewnętrzne, składa się na końcowe, subiektywne odczucie jakości usług postrzegane przez użytkownika – QoE (ang. *quality of experience*). Każdy system, urządzenie i usługa udostępnia wiele parametrów pracy za pośrednictwem różnych protokołów. Najistotniejsze parametry zbierane są przez systemy monitoringu i poddawane analizie (korelowane z danymi historycznym i innymi parametrami).

Liczba parametrów monitorowanych na przykładzie infrastruktury INEA przekracza 1,3 mln i 2,5 tys. nowych wartości na sekundę zbieranych z urządzeń sieciowych, systemów i usług oraz 843 mln odczytów parametrów z urządzeń terminujących usługi i CPE (ang. *customer premises equipment*) w ciągu doby. W obecnie stosowanych systemach monitoringu analiza parametrów odbywa się na podstawie szablonów i zależności określanych przez osoby projektujące monitoring. W systemie Zabbix zdefiniowanych jest ponad 600 tys. alarmów. Przekroczenie założonych warunków powoduje określone reakcje systemu monitoringu lub/i osób odpowiedzialnych za dane urządzenie lub rozwiązanie. Parametry określające jakość usługi pozyskiwane są za pomocą systemu symulującego działania użytkownika końcowego w oparciu o sondy rozmieszczone w sieci dostępowej. Dane z testów wykonywanych przez sondy przekazywane są do systemu Zabbix. Przy tak dużej liczbie parametrów i alarmów, niezwykle skomplikowane jest ich analizowanie określenie wpływu zmiany na jakość usługi dla klienta końcowego. Zastosowanie metod uczenia maszynowego do analizy danych typu big data i korelacji z oceną jakości usługi potencjalnie może usprawnić i przyśpieszyć proces identyfikacji obszarów oraz węzłów sieci rozległych, w których występują (lub wystąpią w określonym horyzoncie czasowym) problemy prowadzące do pogorszenia jakości świadczonych usług.

[1] mgr inż., bartosz.partyka@doctorate.put.poznan.pl, Szkoła Doktorska, Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, https://cat.put.poznan.pl/.
[2] dr hab. inż., piotr.zwierzykowski@put.poznan.pl, Instytut Sieci Teleinformatycznych, Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, https://cat.put.poznan.pl/.

2. Monitoring infrastruktury operatora telekomunikacyjnego INEA

Monitoring infrastruktury w INEA podzielony jest na dwa obszary obsługiwane przez odrębne systemy. Poglądowy uproszczony schemat monitorowanej infrastruktury sieciowej przedstawia rysunek 1.



Rysunek 1. Schemat poglądowy infrastruktury sieciowej ISP [opracowanie własne]

Sieć szkieletowa, punkty styku z innymi operatorami, urządzenia *edge*[3], urządzenia w sieci dostępowej i systemy monitorowane przez system Zabbix na licencji GNU GPL[4] version 2 zaprojektowanego do monitoringu czasu rzeczywistego [1]. Szerszy opis rozwiązania zawarto w punkcie 2.1.

Dane z urządzeń terminujących usługi i CPE zbierane są przez autorski system INEA pod nazwą Diagnostyka. Rozwiązanie zbiera dane również z urządzeń agregujących CPE, takich jak OLT[5] (ang. *Optical Line Terminal*) i CMTS[6] (ang. *Cable Modem Termination System*) jednak tylko z części związanej z transmisją z urządzeniami terminującymi.

[3] Routery, przełączniki realizujące transmisje danych pomiędzy urządzeniami sieci szkieletowej a urządzeniami w sieciach dostępowych.
[4] Licencja wolnego i otwartego oprogramowania.
[5] Koncentrator zainstalowany w centralnym punkcie sieci optycznej, agreguje urządzenia terminujące usługi.
[6] Koncentrator zainstalowany w centralnym punkcie sieci wykorzystującej koncentryczne kable sygnałowe, agreguje urządzenia terminujące usługi.

2.1. Monitoring systemów, sieci szkieletowej i dostępowej przy pomocy Zabbix

System Zabbix w INEA na rysunku 1. został oznaczony prostokątem „Monitoring". Rozwiązanie złożone jest z serwera WWW, serwera aplikacyjnego, dwunastu serwerów *proxy*[7], *load balancera HAproxy* rozkładającego zapytania pomiędzy bazami danych w klastrze multi-master opartym na MariaDB Galera Cluster [2]. Zabbix zbiera dane z 7,5 tys. urządzeń sieciowych, alarmów, UPS i wielu innych typów. Pobierane do zapisu i analizy jest 2,5 tys. wartości na sekundę, łączna liczba parametrów wynosi blisko 1,3 mln. Na parametrach skonfigurowanych jest 600 tys. alarmów, które informują pracowników działu utrzymania sieci o zmianach parametrów działania i/lub występujących problemach w infrastrukturze. Zabbix przez interfejs WWW umożliwia przeglądanie zebranych danych również w formie wykresów. Przygotowane są również mapy sieci odzwierciedlające infrastrukturę fizyczną [1]. Na podstawie tych danych użytkownicy systemu, których w INEA jest 280, mogą podejmować decyzje i reagować w celu utrzymania jakości usług na założonym poziomie. W dużej części monitoring oparty jest na protokole SNMP (ang. *Simple Network Management Protocol*), za pośrednictwem którego cyklicznie odpytywane są urządzenia przy użyciu OID (ang. *object identifier*). Zabbix może realizować proste odkrywanie indeksów OID przy użyciu funkcji SNMPWALK[8]. Bardziej złożone odkrywanie indeksów realizowane jest przez autorskie skrypty pracowników INEA i przekazywane do systemu za pośrednictwem narzędzia Zabbix Sender do serwera Zabbix. Część danych zbieranych jest formie zdarzeń za pomocą trapów SNMP. Urządzenia przesyłają dane w momencie zmiany wartości parametru. Zagadnienie zostało szerzej opisane w punkcie 5.1.1. W mniejszym zakresie używane są sprawdzenia parametrów przy użyciu protokołów SSH (ang. *Secure Shell*) i Telnet. Systemy i usługi monitorowane są głównie z wykorzystaniem usługi Zabbix Agent instalowanej w systemie operacyjnym, która zbiera dane i przekazuje je do serwera Zabbix bezpośrednio lub pośrednio przez proxy [1]. Do systemu Zabbix trafiają dane z systemu monitoringu jakości usług, który działa w oparciu o sondy symulujące zachowanie użytkowników usług. Parametry wysyłane są do centralnego systemu zarządzającego sondami i zbierającego wyniki testów. Dane z testów pobierane są przez REST API[9] i przekazywane do systemu Zabbix w celu dalszej analizy. Interpretacja wyników uzyskanych z sond pozwala oceniać jakość świadczonych usług i szacować QoE klientów [3]. Taka ocena uzupełnia dane z wszystkich elementów sieci o wynikową jakość usługi, na którą składają się wszystkie monitorowane elementy infrastruktury. Zagadnienie oceny jakości usług szerzej opisane zostało w punkcie 5.1.2.

2.2. Diagnostyka urządzeń terminujących usługi u klientów

Kolekcja danych z urządzeń terminujących usługi i CPE zlokalizowanych u klientów odbywa się za pomocą systemu stworzonego na potrzeby INEA przez pracowników. Rozwiązanie na rysunku 1. oznaczone jest prostokątem „CPE diagnostic" i składa się z kolektorów odpowiedzialnych za zbieranie logów, bazy danych i cache, w których

[7] Serwer zbierający dane z urządzeń i przekazujący do serwera aplikacyjnego Zabbix, może wykonywać wstępną obróbkę danych.
[8] Odczyt wartości drzewa OID od wskazanego wierzchołka do liści.
[9] Uniwersalny interfejs HTTP do komunikacji między oprogramowaniem klienta a serwerem za pośrednictwem sieci.

dane są przechowywane. Dodatkowym narzędziem do zbierania danych z CPE jest serwer ACS Maestro[10], którego podstawowym zadaniem jest *provisioning*[11] urządzeń terminujących usługi.

Zbieranie danych z modemów kablowych sieci HFC realizowane jest przez 5 kolektorów wykonujących odczyty z każdego aktywnego urządzenia co 10 minut. Odczyty odbywają się przy użyciu protokołu SNMP, najpierw pobrana zostaje lista włączonych modemów oraz parametry sygnału odbieranego przez CMTS, w następnym kroku odpytywane są modemy o parametry odbieranego sygnału. Dane trafiają do bazy danych PostgreSQL oraz do cache Memcached, z którego część parametrów pobierana jest przez kolektory do wyliczenia zmiany wartości w czasie. Łącznie z sieci HFC zbieranych jest 827 mln parametrów w ciągu doby z 81 tys. modemów i 16 CMTS.

W technologii GPON przy użyciu SNMP odczytywane są parametry urządzeń końcowych (ONT) z OLT, który przy użyciu protokołu OMCI odczytuje parametry sygnału odbieranego przez ONT. Taka architektura powoduje konieczność pośredniczenia przez OLT w odczytach parametrów ONT i zwiększa ich obciążenie. Rozwiązaniem jest monitoring w modelu hybrydowym – urządzenia OLT przesyłają do kolektora trapy SNMP w przypadku zmiany parametrów i przekroczenia zdefiniowanego progu parametrów optycznych oraz wykonywany jest cykliczny odczyt dwa razy na dobę. Trap SNMP zawiera informację o przekroczeniu bez jej wartości, więc w momencie odebrania trapu wykonywany jest SNMPGET pobierający wartość parametru, który uległ zmianie. Taki model jest bardziej skomplikowany od opartego o cykliczne odczyty, ale pozwala zaoszczędzić miejsce w bazie danych i analizować tylko zmiany parametrów. Dodatkowo urządzeni OLT są znacznie mniej obciążane odczytami związanymi z monitoringiem. Odczyty realizowane przez serwer ACS przy wykorzystaniu protokołu CMWP (ang. *CPE WAN Management Protocol*) i modelu TR-69 pozwalają odczytywać parametry bezpośrednio z CPE. Wymiana danych między CPE a serwerem ACS odbywa się przy użyciu protokołu HTTP z szyfrowaniem SSL w formacie XML [4, 5]. W INEA z sieci GPON odczytywanych jest na dobę 16 mln parametrów z 251 tys. aktywnych urządzeń ONT i 404 OLT. Statystyki zawarte w niniejszym punkcie są aktualne na dzień 09.07.2021, sieć podlega ciągłemu procesowi rozbudowy i ilość monitorowanych parametrów systematycznie rośnie.

Dane z powyższych systemów analizowane są przez pracowników NOC[12] (ang. *Network Operations Center*) w czasie rzeczywistym, przez pracowników Call Center i drugiej linii wsparcia w przypadku wystąpienia usterki u klienta. Bieżące i historyczne są wykorzystywane przez administratorów zarządzających urządzeniami w danej technologii dane do analiz i pracy operacyjnej.

3. Ograniczenia obecnego podejścia do monitoringu

Z szerokiego zakresu zagadnień optymalizacji i integracji (również tych już zrealizowanych), wybrano i w niniejszym punkcie opisano te, które w ocenie INEA są najbardziej istotne. Rozwój w tych obszarach jest najbardziej uzasadniony ekonomicznie.

Standardowe podejście do monitoringu sieci oparte o rozwiązania opisane w punkcie 2 niesie za sobą konsekwencje w postaci bardzo dużego zaangażowania osób tworzących

[10] Komercyjny system do zarządzania i monitoringu CPE dystrybuowany przez Seliro AB.
[11] Proces automatycznej konfiguracji urządzeń i sieci umożliwiający dostęp do usług.
[12] Dział utrzymania sieci.

alarmy w systemie Zabbix i konieczności zaangażowania w proces przygotowania alarmów adminstratorów odpowiedzialnych za dane urządzenie lub technologię. Na podstawie analiz danych historycznych i doświadczenia tych osób tworzone są często bardzo złożone alarmy. Przykładowy alarm na spadek liczby ONT na porcie OLT przedstawiono poniżej (1).

$$\{\text{OLT1: OntTotal. GPON. [GPON0/0/0]. last()}\} > 2 \; and \; \{OLT1: OntOnline.GPON.[GPON0/0/0].change()\} < 0 \; and \; \{OLT1: OntTotal.GPON.[GPON0/0/0].change()\} = 0 \; and \; (\{OLT1: OntOnline.GPON.[GPON0/0/0].avg(900)\} < 0.75 * \{OLT1: OntOnline.GPON.[GPON0/0/0].avg(900{,}86400)\} \; and \; \{OLT1: OntOnline.GPON.[GPON0/0/0].avg(900{,}86400)\} > 20) \; or \; (\{OLT1: OntOnline.GPON.[GPON0/0/0].last()\} = 0 \; and \; \{OLT1: OntOnline.GPON.[GPON0/0/0].avg(900{,}86400)\} < 20)) \quad (1)$$

gdzie: OLT1 – nazwa urządzenia, OntTotal.GPON.[GPON0/0/0] – unikalny klucz (w obrębie urządzenia) identyfikujący parametr, *last* – odczytana ostatnia wartość, *change* – różnica między ostatnią odczytaną wartością a poprzednią, *avg(x,y)* – średnia wartość w przedziale czasowym *x* z przesunięciem wstecz o czas *y*.

System Zabbix umożliwia tworzenie zależności tylko na poziomie alarmów. Większość parametrów monitorowanych i alarmów jest dodawana do systemu w oparciu o mechanizm *discovery*. Zależności alarmów w obrębie jednego hosta są utrzymywane automatycznie [1, 6]. Utrzymanie zależności między hostami, z uwagi na ich dużą liczbę i dynamicznie zmieniającą się konfigurację, jest trudne i czasochłonne. W obecnie dostępnych rozwiązaniach, z których część została przeanalizowana (punkt 4), nie jest możliwe tworzenie hierarchii urządzeń, która pozwala na tworzenie automatycznej zależności na poziomie alarmów. Zależności nie dotyczą tylko urządzeń sieciowych i serwerów, ale również infrastruktury energetycznej i klimatyzacji w stacjach czołowych i centrach danych. Elementy te są również monitorowane. W przypadku awarii klimatyzacji występuje wzrost temperatury pracy wielu urządzeń w danej lokalizacji. System monitoringu raportuje dużą liczbę przekroczeń zadanych progów. Duży napływ informacji utrudnia szybkie określenie przyczyny awarii w postaci uszkodzonej klimatyzacji. W przypadku awarii analiza zależności spoczywa w dużej części na pracownikach NOC i osobach odpowiedzialnych za dany obszar infrastruktury i pełniących dyżur w danym czasie. Rozwiązaniem ograniczającym wzajemny wpływ liczby alarmów różne obszary infrastruktury sieci jest tworzenie oddzielnych *dashboardów*[13], zawierających alarmy tylko z danego obszaru. *Dashboardy* z istotnych obszarów sieci wyświetlane są stale na *Video Wall*[14] i analizowane w czasie rzeczywistym przez pracowników NOC.

Z uwagi na obecną architekturę monitoringu w INEA, w której wyodrębnione są dwa obszary infrastruktury, z których dane zbierane są przez niezależne rozwiązania, niezbędna jest integracja systemów Zabbix i Diagnostyka. Integracje realizowane są przez dodatkowe skrypty pobierające dane i wprowadzające zmiany w Zabbix przez

[13] Definiowalny ekran w systemie Zabbix umożliwiający wyświetlanie wybranych informacji.

[14] Instalacja złożona z wielu ekranów logicznie działająca jako jeden ekran, umożliwiająca wyświetlanie obrazu na dużej powierzchni.

API oraz bezpośrednio w bazie. Do Zabbix dane pobierane są przez skrypty odpytujące bazę danych systemu Diagnostyka i wysyłające do systemu za pośrednictwem narzędzia Zabbix Sender. Dodatkowe narzędzia integracyjne tworzone w INEA i umieszczane w GUI Zabbixa przy każdej aktualizacji systemu muszą być dodawane ponownie i testowane, ponieważ są nadpisywane przez instalator. Często zmienia się również struktura API i klas PHP, które wykorzystują. Wobec powyższego proces integracji jest czasochłonny i utrudniony, przez co w wielu przypadkach konieczność określania wzajemnego wpływu zmian na przez pracowników NOC.

Wszystkie zagadnienia opisane w niniejszym punkcie składają się na duże obciążenie napływającymi informacjami i konieczność analizy zależności przez NOC. Są to zagadnienia, których optymalizacja pozwoli przyśpieszyć proces identyfikacji przyczyny awarii i przeciwdziałać degradacji jakości usług.

4. Dostępne rozwiązania i narzędzia

Po zidentyfikowaniu głównych ograniczeń i obszarów do optymalizacji opisanych w punkcie 3. rozpoczęto analizę dostępnych rozwiązań. Analiza dotyczy tylko tych obszarów i uwzględnia rodzaj licencjonowania. Pięć z analizowanych systemów jest aktualnie używanych w INEA. Listę analizowanych systemów zawiera tabela 1. Wybrane zostały narzędzia w ocenie autorów najbardziej uniwersalne, znane i adresujące najszerszy możliwy zakres związany z monitoringiem.

Tabela 1. Lista analizowanych systemów

System	Szkielet	Dostęp	CPE	W INEA
NCE(U2000) – Huawei	Nie	Tak	Tak	Tak
ACS Maestro	Nie	Nie	Tak	Tak
AMS – Nokia	Nie	Tak	Tak	Tak
eSight – Huawei	Tak	Tak	Tak	Tak
Zabbix	Tak	Tak (nie w pełni)	Nie	Tak
ManageEngine OpManager	Tak	Tak (nie w pełni)	Nie	
NetCrunch	Tak	Tak (nie w pełni)	Nie	
Nagios Core	Tak	Tak (nie w pełni)	Nie	
Cacti	Tak	Tak (nie w pełni)	Nie	
Prometheus	Tak	Tak (nie w pełni)	Nie	
SolarWinds NPM	Tak	Tak (nie w pełni)	Nie	

Źródło: Opracowanie własne

Systemy NCE [7], AMS [8] adresują monitoring tylko sieci dostępowej GPON i ONT, nie mogą monitorować urządzeń końcowych w sieciach HFC. ACS Maestro jest narzędziem dedykowanym do provisioningu i monitoringu tylko CPE i urządzeń IoT[15] z wykorzystaniem standardów TR-069 i TR-181 [9]. System Huawei eSight adresuje bardzo szeroki zakres monitoringu, systemów, sieci szkieletowych, dostępowych i urządzeń końcowych w sieci GPON. Monitoring CPE w sieciach HFC jest możliwy, jednak wymagałby zbudowania całej logiki monitoringu. Nie posiada on możliwości budowania hierarchii, hostów i alarmów oraz wymaga zakupu dodatkowych licencji [10].

[15] Internet rzeczy, przedmioty zbierające, przetwarzające i wymieniajace dane.

System Zabbix jest aktualnie używany w organizacji. Zagadnienia związane z pokryciem obszaru monitoringu zostały opisane w punkcie 3. Najnowsza dostępna wersja 5.4 nie wprowadza istotnych zmian [11].

Następną grupę stanowią narzędzia Open Source – Nagios Core, Cacti, Prometheus. Funkcjonalnie są zbliżone do systemu Zabbix i adresują te same obszary monitoringu. Nagios Core posiada możliwość konfiguracji zależności i usług oraz hostów jednak odnosi się to tylko do ich osiągalności. Mogą one przyjmować wartości UP, DOWN, lub/i UNREACHABLE [12-14].

Rozwiązania płatne ManageEngine OpManager, NetCrunch i SolarWinds NPM nie rozszerzają funkcjonalności w obszarach istotnych dla ISP. Pozwalają modelować procesy i posiadają szereg wtyczek i predefiniowanych szablonów do monitorowania szerokiego zakresu urządzeń. Są proste do wdrożenia w małych i średnich firmach, bo nie wymagają specjalistycznej wiedzy z zakresu administracji urządzeniami sieciowymi, serwerami i systemami. Zawierają skompilowane MIB dla większości urządzeń wiodących wendorów, co znacznie ułatwia i przyśpiesza uruchomienie monitoringu. Bardziej zaawansowany monitoring, taki jak realizowany jest w infrastrukturze ISP, wymaga jednak dużego nakładu pracy, jak w przypadku systemów Open Source, ze względu na specyficzne rozwiązania i wyspecjalizowane urządzenia [15-17].

Brak znaczących różnic w funkcjonalności analizowanych narzędzi i brak możliwości monitorowania całej infrastruktury jednym systemem nie daje zasadnych podstaw do zmiany systemu monitoringu. Zmiana wiązałaby się z koniecznością dużego zaangażowania we wdrożenie nowego narzędzia. Lepszym rozwiązaniem wydaje się być rozwój i integracja aktualnie używanych systemów w obszarze podlegającym analizie opisanej w niniejszym opracowaniu.

5. Elementy infrastruktury niezaadresowane przez dostępne rozwiązania

W wyniku analizy wyodrębniono trzy grupy zagadnień, które podzielone zostały na dwa etapy rozwoju monitoringu. W opracowaniu skupiono się głównie na etapie pierwszym, który polegał szerszej analizie. Opis zawarto w punkcie 5.1. W etapie drugim (punkt 5.2) hipotetycznie możliwe jest wykorzystanie metod uczenia maszynowego. Wszystkie zaproponowane rozwiązania są wynikiem wstępnych analiz i badań. Żadne z proponowanych rozwiązań opartych na uczeniu maszynowym nie zostało zaimplementowane. Część rozwiązań opisanych w punkcie 5.1.1 jest uruchomiona w produkcyjnym środowisku, ale wymaga dalszego rozwoju i optymalizacji.

5.1. Rozwój systemów monitoringu – etap 1

W tym etapie dokonano podziału zagadnień na dwie grupy. Pierwsza grupa opisana w punkcie 5.1.1 zawiera zagadnienia integracyjne, optymalizacyjne i wymagające rozbudowy istniejących funkcjonalności systemów monitoringu. W drugiej grupie opisanej w punkcie 5.1.2 znalazły się zagadnienia, w których w ocenie autorów możliwe jest wykorzystanie metod uczenia maszynowego w kilku wariantach.

5.1.1. Konwencjonalne podejście

Rozwiązanie integracyjne i optymalizacyjne wykraczają poza temat niniejszego opracowania, jednak są istotne z punktu widzenia całego monitoringu infrastruktury ISP. Zostały wskazane i krótko opisane.

Integracja informacji z monitoringu, diagnostyki i monitoringu jakości usług jest istotnym elementem z punktu widzenia NOC. Problemy występujące w sieci dostępowej powodują konieczność wyszukania w systemie Diagnostyka urządzenia agregującego np. OLT i urządzeń ONT, które są do niego podłączone, aby określić skalę degradacji usług. Podobna sytuacja ma miejsce w przypadku problemów na porcie PON wskazanych przez system Zabbix. Proponowane rozwiązanie pozwoli przejść z menu alarmu do listy CPE potencjalnie objętych degradacją. Informacje będą też przekazywane do działów Call Center i drugiej linii wsparcia. Na karcie urządzenia zamontowanego u klienta w systemie Diagnostyka dodana zostanie informacja pobrana z systemu Zabbix dotycząca potencjalnych degradacji spowodowanych przez urządzenia nadrzędne, informacje dotyczące jakości usług pobrane z systemu monitoringu jakości usług i indeks jakości usług opisany w punkcie 5.1.2. Takie rozwiązanie pozwoli szybciej reagować oraz określać grupy urządzeń końcowych, na jakie ma wpływ awaria lub degradacja. Zagadnienia integracji systemów Zabbix i Diagnostyka są częściowo zrealizowane w zakresie awarii masowych w narzędziach zewnętrznych, natomiast informacje te nie są prezentowane w tych systemach.

Monitoring hybrydowy jest koncepcją zakładającą dostarczenie do systemu Zabbix dla jednego klucza identyfikującego wartość w obrębie hosta, informacji z kilku źródeł. Urządzenia wysyłają trapy SNMP z informacją o wystąpieniu problemu, ale w sytuacjach utraty komunikacji z systemem monitoringu, administracyjnego wyłączenia interfejsu lub w wyniku restartu i wielu innych przypadkach nie przesyłają trapów kończących zdarzenie. Konieczne jest wtedy odczytanie informacji z urządzenia po określonym czasie w celu potwierdzenia lub zakończenia wyświetlania alarmu w systemie monitoringu. Natywnie Zabbix nie posiada takiej funkcjonalności, a utrzymywanie każdego rodzaju sprawdzenia odrębnie, zwiększa objętość bazy danych i komplikuje budowę alarmu. Wstępne rozważania zakładają wykorzystanie kolektora logów typu Splunk, w którym zostanie zrealizowana logika lub implementacja skryptu na kolektorze trapów SNMP, który wykona powiązanie.

Obecnie analizie podlega system tagowania interfejsów na urządzeniach sieciowych. Na tagach bazuje również system monitoringu, określając priorytet alarmów wyświetlanych na dashboard Zabbix. Tagi przechowują również wiele innych informacji określających zależności między urządzeniami. Liczba ta stale się zwiększa. Utrzymanie spójnej struktury tagowania wymaga dużego zaangażowania po stronie administratorów sieci. Koncepcja zakłada dodawanie tagów na interfejsach w momencie dodawania ich do systemu monitoringu. Administrator w systemie Zabbix określi priorytet i wszystkie niezbędne informacje oraz wskaże urządzenie, do którego będzie realizowana komunikacja przez ten interfejs. Formularz będzie generował w sposób ustandaryzowany tagi i poprzez SNMP konfigurował je na urządzeniu lub/i wyświetlał tag do wklejenia w opisie interfejsu na urządzeniu.

5.1.2. Wykorzystanie metod uczenia maszynowego

Działania na dużych zbiorach danych, konieczność definiowania progów i algorytmów do analizy powodują, że monitoring infrastruktury staje się coraz bardziej czasochłonny i skomplikowany. Wykorzystanie metod uczenia maszynowego wydaje się być dobrym kierunkiem rozwoju systemów monitoringu w celu zautomatyzowania budowy zależności, oceny parametrów i ograniczenia zakresu informacji, na których działy

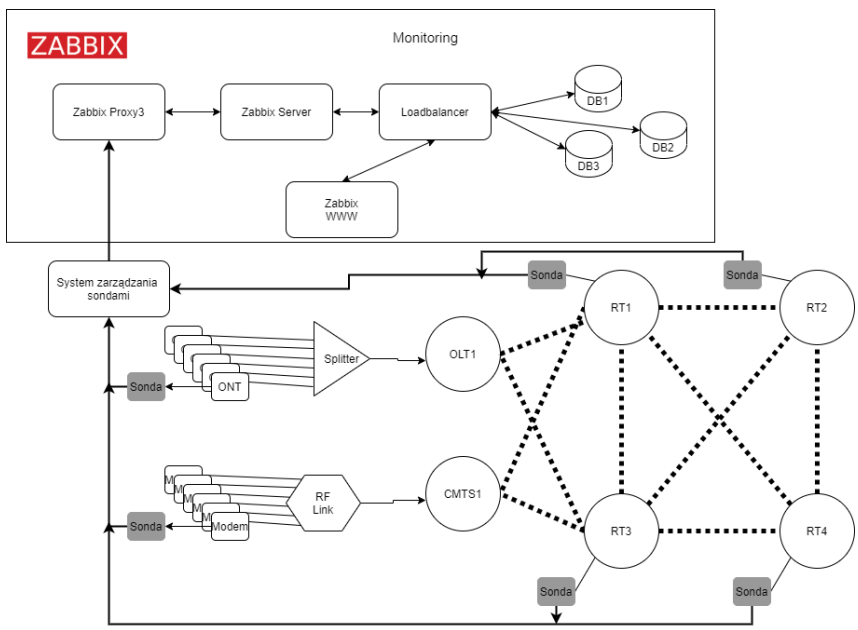
NOC podejmują decyzje operacyjne do takiego poziomu, żeby skrócić czasy podejmowania decyzji, reakcji i usuwania problemów. W wyniku analizy wytypowano trzy zależne od siebie obszary, w których zastosowanie uczenia maszynowego jest zasadne.

Bardzo ważnym obszarem monitoringu sieci ISP jest analiza jakości usług dostarczanych klientom. Zadania takie realizują zewnętrze systemy zbierające dane z sond rozmieszczonych w sieci. Dane przekazywane są do systemu monitoringu Zabbix. Schemat poglądowy rozwiązania zawiera rysunku 2.

W wyniku wykonywanych testów symulujących zachowania klientów dostarczane są duże ilości danych do analizy. Wynikiem oceny parametrów testów z sond jest indeks jakości usług.

Testy wykonywane na sondach będące składowymi indeksu to:

- ping – pomiar czasów odpowiedzi zdefiniowanego hosta;
- dns – pomiar czasu odpowiedzi serwera i weryfikacja poprawności rozwiązania zdefiniowanej listy adresów domenowych na serwerach DNS ISP;
- www – pomiar czasu pobrania testowej strony i najpopularniejszych serwisów (określone na podstawie statystyk z serwerów DNS);
- ott – weryfikacja poprawności pobierania chunk[16] i ocena parametrów;
- sip – weryfikacja poprawności logowania na centrali i ocena parametrów wykonanego połączenia.



Rysunek 2. Schemat poglądowy monitoringu jakości usług [opracowanie własne]

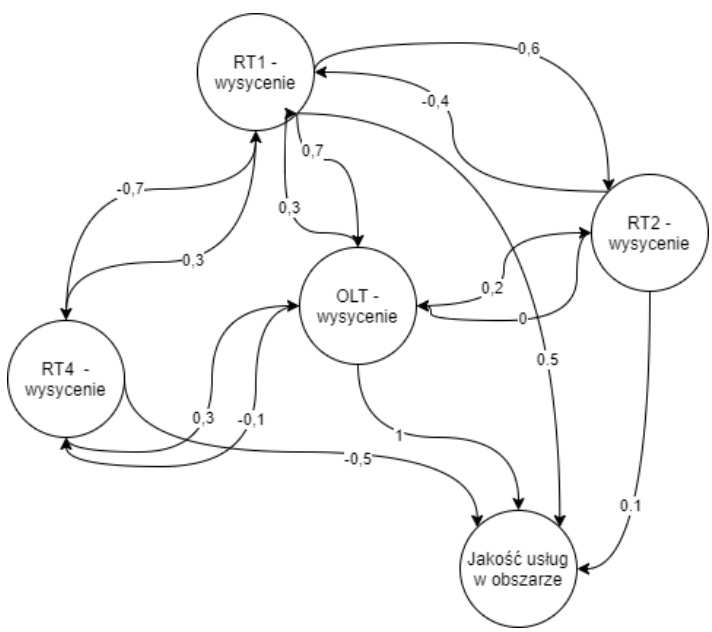
Zakres parametrów i sposób ich oceny jest określony między innymi przez Europejski Instytut Norm Telekomunikacyjnych [18]. Na podstawie standardu opracowana zostanie ocena zbioru uczącego dla algorytmów regresyjnych. W pierwszym kroku wykonane zostaną analizy i testy wykorzystania regresji liniowej z uwagi na szybkość uczenia

[16] Strunień wideo pobierany jest w częściach (chunk), które są odtwarzane w określonej kolejności.

i szybkie działanie. Jeżeli jakość predykcji oceniona za pomocą wskaźników, takich jak błąd średniokwadratowy czy średni błąd bezwzględny to badania prowadzone będą nad wykorzystaniem nadzorowanego uczenia sieci neuronowych [19]. Indeks szacowany będzie dla pojedynczych sond i dla całych obszarów sieci, w których sondy są zainstalowane. W przyszłości podjęta zostanie próba uzupełnienia indeksu o informacje z warstwy trzeciej modelu ISO OSI, takich jak retransmisje TCP z wszystkich urządzeń terminujących usługi w sieciach dostępowych.

Opracowanie indeksu pozwoli uzyskać ocenę usług w węźle sieci dostępowej, względem której podjęta zostanie próba określenia relacji z parametrami pracy urządzeń sieciowych. Rozważane jest wykorzystanie parametrów, takich jak QoS[17] Tail Drops[18], wysycenia i brak komunikacji na interfejsach sieciowych. Do określenia relacji wybrano proces mapowania kognitywnego.

Mapa kognitywna jest grafem skierowanym, którego wierzchołki oznaczają czynniki, a intensywność wzajemnego wpływu czynników określają wagi. Teoretyczny przykład mapy określającej wpływ wysyceń na routerach sieciowych i OLT, a jakością usług w danym obszarze zawiera rysunek 3. Wartości określają natężenie oddziaływania i kierunek oddziaływań (tzn. czy wzrost danego czynnika powoduje spadek czy wzrost czynnika, na który wpływa).



Rysunek 3. Poglądowy przykład relacyjnej mapy kognitywnej [opracowanie własne]

Z uwagi na działanie na liczbach właściwe w ocenie autorów wydaje się być użycie mapy relacyjnej (tzw. wariant ostry). Określenie urządzenia niepełnym zestawem parametrów wybranych ekspercko jako najistotniejsze wprowadza elementy rozmytości (nieprecyzyjne określenie czynnika). W pierwszym etapie tworzenia mapy czynniki

[17] Protokół priorytetowania transmisji pakietów w warstwie 2 modelu OS.
[18] Algorytm kolejkowania odrzucajcy pakiety w momencie napełnienia kolejki bez rozróżniania klas QoS.

i ich wzajemny wpływ muszą zostać określone przez zespół ekspertów. Parametry czynników podlegają normalizacji do przedziałów liczbowych [0,1] lub [-1,1]. Konieczne jest również określenie charakteru relacji modelowanej struktury – czy jest on statyczny (wartość czynnika w kolejnym przedziale czasowym jest zależna od pozostałych czynników i ich wzajemnego wpływu) czy dynamiczny (wartość czynnika w kolejnym przedziale czasowym jest zależna od zmiany wartości innych czynników i ich wzajemnego wpływu). Następnie należy przeanalizować i wybrać typ modelu relacyjnej mapy kognitywnej spośród następujących:

- model liniowy z nieliniowością dodatnią;
- model nieliniowy;
- model liniowy z nieliniowością szybkości zmian;
- model nieliniowy z nieliniowością szybkości zmian;
- model nieliniowy z nieliniowością dodatnią.

Ostatnim etapem przygotowania mapy w podejściu eksperckim jest określenie funkcji progowej, która może przybrać jedną z następujących postaci:

- liniowa z ograniczeniami;
- sigmoidalna;
- tangensoidalna.

Tak przygotowana mapa określa zależności określone na podstawie wiedzy eksperckiej i analizy danych historycznych. Mapa taka spełnia podstawowe kryteria jednak przy dużej złożoności istnieje duże prawdopodobieństwo, że mapa w pewnym przybliżeniu określna stan faktyczny. Istnieje konieczność adaptacji parametrów mapy przy użyciu technik uczenia. Uczenie mapy jest to proces modyfikacji mocy relacji między czynnikami. Uczenie map może być zrealizowane w oparciu o uczenie nadzorowane i metody populacyjne, takie jak algorytmy genetyczne lub optymalizacja rojem. Znane są też przykłady wykorzystania uczenia nienadzorowanego w oparciu o metodę Hebba [20-22]. Z uwagi na skalarny charakter relacyjnych map kognitywnych możliwe jest wykorzystanie wielu metod uczenia maszynowego, które będą analizowane w kolejnych etapach badań.

Ocena jakości usług i mapy określające wpływ parametrów urządzeń sieciowych na tę ocenę mogą być źródłem danych dla algorytmów klasyfikacji, które z dużym prawdopodobieństwem będą w stanie określić wpływ zmiany parametru przekraczającego próg alarmowy na jakość usług w poszczególnych obszarach sieci. Mapy kognitywne pozwolą zredukować liczbę czynników do tych, które mają istotny wpływ na usługę w danym obszarze. W pierwszym etapie badana będzie możliwość wykorzystania klasyfikacji binarnej do określenia czy zmiana ma wpływ na usługi, a następnie zostaną rozszerzone o modele klasyfikacji wieloklasowej w celu określenia rodzajów usług dotkniętych degradacją. Podstawowym założeniem jest analiza parametrów z perspektywy węzłów sieci dostępowych, takich jak OLT i CMTS.

Wszystkie zagadnienia wymienione w punktach 5.1.1 i 5.1.2 podlegają badaniom z uwagi na konieczność zmniejszenia obciążenia działów NOC oraz czasu reakcji na występujące awarie, co może mieć wpływ na podniesienie jakości usług świadczonych przez ISP.

5.2. Rozwój systemów monitoringu – etap 2

Etap drugi obejmuje badania związane z opracowaniem mechanizmu automatycznego budowania map sieci i usług. Niezbędnym elementem do rozpoczęcia badań jest opracowanie mechanizmu automatycznego i ustandaryzowanego tagowania interfejsów sieciowych, aby możliwe było określenie wzajemnych relacji między urządzeniami w sposób jednoznaczny. W obecnie dostępnych systemach monitoringu możliwe jest automatyczne tworzenie map zaledwie na poziomie sieci warstwy drugiej. Automatyczne generowanie map w sieci rozległych uzupełnione o parametry działania pozwalałoby je w szybki sposób modyfikować i odzwierciedlać zmiany struktury sieci.

6. Dalszy rozwój badań

W kolejnych etapach głównym celem prac będzie zbudowanie mechanizmów do normalizacji i wydajnego udostępniania danych z systemów monitoringu. Wybrane zostaną parametry, na podstawie których będą budowane modele uczenia maszynowego. Proces obejmuje ustandaryzowanie jednostek dla tych samych grup parametrów i odfiltrowanie danych nieprawidłowych.

Dla każdego z zastosowań opisanych w punkcie 5.1.2 prowadzone będą badania i testy w celu wybrania optymalnego modelu uczenia maszynowego. Konieczne jest opracowanie metody oceny i optymalizacji testowanych rozwiązań.

7. Podsumowanie

W niniejszym opracowaniu wskazano te elementy monitoringu ISP, których nie obejmują analizowane rozwiązania i wymagają rozwoju z uwagi na istotny wpływ na czas podejmowania decyzji w procesie diagnozowania awarii. Z tego zbioru elementów wybrano te, dla których przewidziano zastosowanie metod uczenia maszynowego. Opracowane mechanizmy w ocenie autorów pozwolą określić wpływ zmiany parametrów na jakość świadczonych usług i z dużym prawdopodobieństwem pozwolą zapobiegać degradacji i/lub awarii. Przedstawione analizy stanowią wstęp do badań, których celem jest zastosowanie metod uczenia maszynowego w monitoringu sieci i usług ISP.

Literatura

1. Uytterhoeven P., Olups R., *Zabbix 4 Network Monitoring – Third Edition*, Packt Publishing, Wielka Brytania 2019.
2. *What is MariaDB Galera Cluster?*, https://mariadb.com/kb/en/what-is-mariadb-galera-cluster/, 11.07.2021.
3. *Speech and multimedia Transmission Quality (STQ);QoS aspects for popular services in mobile networks; Part 2: Definition of Quality of Service parameters and their computation*, European Telecommunications Standards Institute, Francja 2015.
4. *AR500, AR510, AR531, AR550, and AR2500 V200R008 CLI-based Configuration Guide - Network Management and Monitoring*, https://support.huawei.com/enterprise/en/doc/EDOC1000154773/919d42c6/cwmp-configuration, 12.07.2021.
5. TR-069 CPE WAN Management Protocol, Broadband Forum, 2020.
6. *Trigger dependencies*, https://www.zabbix.com/documentation/current/manual/config/triggers/dependencies, 12.07.2021.
7. 5520 Access Management System,Release 9.6.0, Nokia, 2018.
8. Network Cloud Engine V100R019C00 Product Description, HUAWEI TECHNOLOGIES CO., LTD., Shenzhen 2019.

9. *Seliro Managed Mesh Solution*, https://seliro.com/wp-content/uploads/2020/12/Managed-Mesh-Solution-Description_Dec-2020.pdf, 15.07.2021.
10. *eSight V300R007C00 Product Documentation 16(chm)*, https://support.huawei.com/enterprise/en/doc/EDOC1000132233, 10.07.2021.
11. *What's new in Zabbix 5.4.0*, https://www.zabbix.com/documentation/current/manual/introduction/whatsnew540, 15.07.2021.
12. *Nagios Core*, https://assets.nagios.com/downloads/nagioscore/docs/nagioscore/4/en/toc.html, 13.07.2021.
13. *Cacti (tm) Documentation*, https://docs.cacti.net/, 14.07.2021.
14. *Prometheus*, https://prometheus.io/docs/prometheus/latest/, 20.07.2021.
15. *ManageEngine OpManager Documents*, https://www.manageengine.com/network-monitoring/documents.html#datasheet, 20.07.2021.
16. *Podręcznik NetCrunch*, https://pl.adremsoft.com/adoc/view/netcrunch/391486581027/podrcznik-netcrunch, 20.07.2021.
17. *Network Performance Monitor Administrator Guide*, https://documentation.solarwinds.com/en/success_center/npm/content/npm_administrator_guide.htm, 20.07.2021.
18. *Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in mobile networks; Part 2: Definition of Quality of Service parameters and their computation*, ETSI, Francja 2015.
19. Géron A., *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, O'Reilly Media, Stany Zjednoczone Ameryki 2017.
20. Słoń G., *Relacyjne rozmyte mapy kongitywne w modelowanie złożonych systemów*, Wydawnictwo Politechniki Świętokrzyskiej, Kielce 2013.
21. Elpinki I., *Fuzzy Cognitive Maps for Applied Sciences and Engineering From Fundamentals to Extensions and Learning Algorithms*, Springer-Verlag, Niemcy 2014.
22. Stach W., Kurgan Ł., Pedrycz W., *A Framework for a Novel Scalable FCM Learning Method*, Symposium on Human-Centric Computing and Data Processing, Kanada 2007.

Analiza możliwości zastosowania mechanizmów uczenia maszynowego w interpretacji danych zbieranych przez systemy monitoringu sieci i jakości usług ISP na przykładzie INEA

Streszczenie

Operatorzy telekomunikacyjni (np. INEA) stale zwiększają zasięg usług i ciągle poszerzają ich zakres. Dlatego ich infrastruktura podlega zmianom i nieustannie rośnie liczba monitorowanych parametrów. W tak dynamicznym środowisku standardowe podejście do monitoringu sieci i usług polegające na określaniu progów alarmów przez administratorów na podstawie analizy danych historycznych i ich doświadczeniu jest niewystarczające. Utrzymanie i analiza zależności między elementami infrastruktury oraz konieczność budowy coraz bardziej złożonych reguł alarmów to dla administratorów niezwykle skomplikowane i czasochłonne zadanie. Działy utrzymania sieci analizują dużą liczbę występujących często w jednym czasie alarmów, które wymagają oceny – a obecnie dostępne systemy monitoringu sieci oferują wykorzystanie w bardzo ograniczonym zakresie tylko prostych modeli predykcyjnych. Z tego względu podjęto badania, których celem jest zwiększenie zastosowania metod ucznia maszynowego w ocenie zbieranych parametrów. Celem pracy jest analiza aktualnie używanych rozwiązań w zakresie monitoringu, diagnostyki CPE (ang. *Customer premises equipment*) i badania jakości usług ISP (ang. *Internet service provider*), na przykładzie INEA oraz wskazanie potencjalnych możliwości wykorzystania uczenia maszynowego. Zastosowanie uczenia maszynowego pozwoli na usprawnienie i przyśpieszenie procesu identyfikacji obszarów sieci, w których występują problemy prowadzące do pogorszenia jakości świadczonych usług (ang. *quality of experience*). W pracy przeanalizowane zostały dostępne narzędzia i rozwiązania w zakresie monitoringu sieci i usług. Wskazano na te elementy monitoringu ISP, których nie obejmują analizowane rozwiązania, a które wymagają rozwoju. Z tego zbioru elementów wybrano te, dla których przewidziano zastosowanie metod uczenia maszynowego. Przedstawione analizy stanowią pierwszy etap prac, których celem jest zastosowanie metod uczenia maszynowego w monitoringu sieci i usług ISP.

Słowa kluczowe: monitoring, ISP, uczenie maszynowe, mapy kognitywne

Analysis of the possibility of using machine learning mechanisms in the interpretation of data collected by network monitoring systems and the services quality monitoring systems of ISP on the INEA example

Abstract

Telecommunications operators (e.g. INEA) are constantly increasing the range of services and constantly expanding their scope. Therefore, their infrastructure is constantly changing and the number of monitored parameters is constantly growing. In such a dynamic environment, the standard approach to network and service monitoring, which consists in determining alarm thresholds by administrators based on historical data analysis and their experience, is insufficient. Maintaining and analyzing the dependencies between infrastructure elements and the need to build more and more complex alarm rules is an extremely complicated and time-consuming task for administrators. Network maintenance departments analyze a large number of frequently occurring alerts that need to be assessed – and currently available network monitoring systems offer a very limited use of only simple predictive models. For this reason, research has been undertaken to increase the use of machine learner methods in the evaluation of collected parameters. The goal of the work is to analyze the currently used solutions in the field of monitoring, CPE diagnostics (Customer premises equipment) and testing the quality of ISP (Internet service provider) services on the INEA example and to indicate the potential possibilities of using machine learning. The use of machine learning will allow for the improvement and acceleration of the process of identifying network areas in which there are problems leading to a deterioration of the quality of experience. The paper analyzes available tools and solutions in the field of network and service monitoring. The elements of ISP monitoring which are not covered by the analyzed solutions and which require development were indicated. From this set of elements, those for which the use of machine learning methods was planned, were selected. The presented analyzes is the first stage of works goals at the application of machine learning methods in the monitoring of ISP networks and services.

Keywords: monitoring, ISP, machine learning, cognitive maps

Magdalena Piwowarczyk[1]

Przegląd algorytmów sterowania zawieszeniem pojazdu

1. Wstęp

Idea sterowanego zawieszenia funkcjonuje już od wielu lat w literaturze, jak i w zastosowaniach komercyjnych. W ciągu ostatnich trzydziestu lat można zaobserwować znaczny wzrost zainteresowania tematem redukcji drgań w pojazdach kołowych. Duży rozwój technologii wytwarzania sterowanych tłumików, które stały się tanie, trwałe i łatwe w implementacji spowodował, że zastosowanie zawieszenia semi-aktywnego może być konkurencyjne w stosunku do klasycznych rozwiązań. Połączenie zaawansowanych sterowników mikroprocesorowych oraz możliwość szybkiego monitorowania stanu pojazdu w trakcie ruchu pozwala na coraz szerszą implementację metod sterowanych (adaptacyjnych). Zawieszenie pasywne wraz ze sztywno przyjętymi parametrami jest w stanie działać efektywnie w wąskim zakresie warunków drogowych. Implementując semi-aktywne sterowanie możliwe jest znaczne poszerzenie zakresu wymuszeń, przy których układ redukcji drgań pracuje efektywnie.

Jednym z najpowszechniej stosowanych regulatorów do zawieszeń sterowanych jest ten oparty o metodę sky-hook. Polega ona na zastosowaniu dodatkowego tłumika łączącego masę resorowaną z wirtualnym punktem znajdującym się w nieruchomym zewnętrznym układzie odniesienia (np. w „niebie"). Takie rozwiązanie ma na celu ustawienie obiektu w stałej odległości od punktu zaczepienia tłumika. W praktyce implementacja tej metody wiążę się z użyciem dodatkowego tłumika pomiędzy masą wibroizolowaną a masą niewibroizlowaną, ponieważ przytwierdzenie tłumika do umownego punktu na „niebie" jest niewykonalne. Działanie dodatkowego elementu, którego zadanie polega jedynie na poprawieniu właściwości dynamicznych masy resorowanej wpływa negatywnie na masę nieresorowaną. Generowana przez tłumik siła oddziałuje w podobnym stopniu na obydwie masy. Dlatego odpowiedź dynamiczna praktycznego układu opierającego się o filozofie sky-hook jest gorsza od jego idealnej realizacji.

W celu poprawienia przyczepności pojazdu najczęściej stosowany jest algorytm ground-hook. Jego zadaniem jest zapewnienie ciągłego docisku masy niewibroizolowanej do nawierzchni [10]. Jest to realizowane poprzez generowanie siły dociskającej tę masę do podłoża. Analogicznie jak w przypadku sterowania typu sky-hook elementem wykonawczym jest tłumik przymocowany do wirtualnego nieruchomego punktu. Współczynnik tłumienia elementu łączącego masę niewibroizlowaną z nieruchomym punktem jest parametrem charakterystycznym przyjętego modelu.

Można również rozważyć algorytm mieszany. Ma on na celu pogodzenie dwóch strategii sterowania: sky-hook i ground-hook [10]. Zadaniem takiego algorytmu jest zapewnienie zarówno komfortu, jak i bezpieczeństwa jazdy. Dwa tłumiki mają za zadanie utrzymywać masę wibroizolowaną na stałym poziomie oraz utrzymywać stały kontakt z podłożem masy niewibroizolowanej.

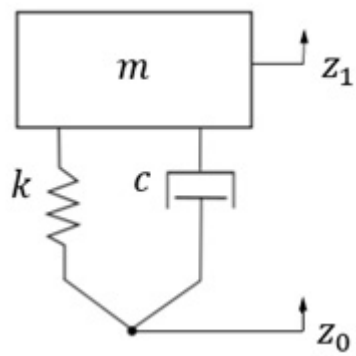
[1] piwowarczyk@agh.edu.pl KAP, WIMiR, AGH w Krakowie.

2. Kryteria oceny zawieszeń

Układ wibroizolacji pełni znaczącą funkcję w projektowaniu samochodów. Rozważmy dobrze znany i szeroko stosowany model pasywnego układu zawieszenia o jednym stopniu swobody (1 DOF), sztywności k i współczynniku tłumienia c (rys. 1).

Projektant takiego układu stoi przed następującymi zadaniami:

- Ruch względny $(z_1 - z_0)$ pomiędzy masą wibroizolowaną m a podłożem musi się odbywać w sposób kontrolowany ze względu na wielkość prześwitu oraz konieczność zapewnienia kontaktu opony z podłożem.
- Zawieszenie powinno redukować drgania masy m w stosunku do wymuszenia np. wartość przemieszczeniowej $(z_1 - z_0)$ funkcji przenoszenia drgań masy wibroizolowanej powinna być mniejsza od 1 w całym zakresie częstotliwości.



Rysunek 1. Pasywny układ zawieszenia pojazdu (1 DOF), opracowanie własne na podstawie [8]

W celu sformułowania wskaźników oceny dla układu redukcji drgań zdefiniowano pomocniczą funkcję γ: $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ równaniem (1).

$$\gamma(X, \underline{t}, \overline{t}) \int_{\underline{t}}^{\overline{t}} |x(t)|^2 dt \tag{1}$$

gdzie:
$x(t)$ to rozpatrywany sygnał,
$\underline{t}, \overline{t}$ określa granice czasu trwania pomiaru.

Zdefiniowano dwa następujące kryteria J_{rh} *i* J_c:

- J_{rh} (*'road-holding'*) – wiążę się ono bezpośrednio z bezpieczeństwem jazdy (2). Na wykresie (Rysunek 2) oś rzędnych.

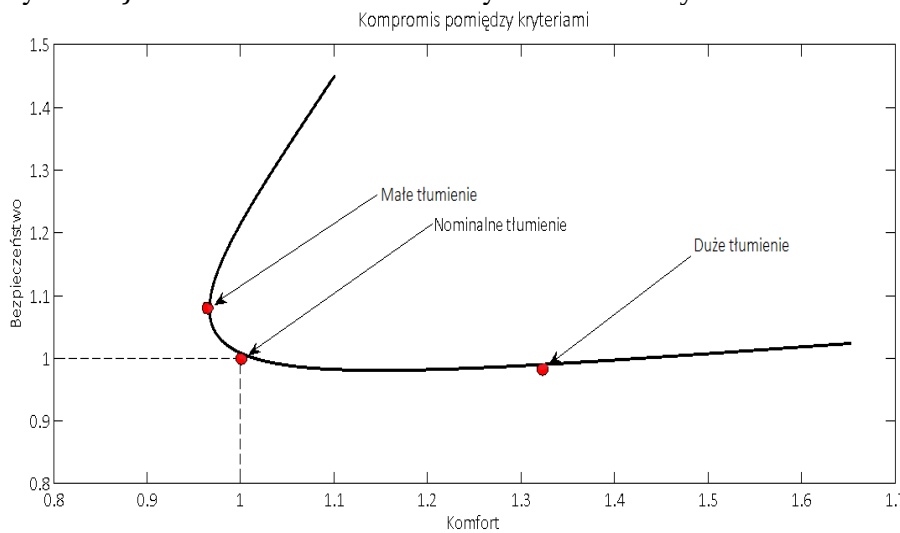
$$J_{rh} = \frac{\gamma(z_{defl}, \underline{t}, \overline{t})}{\gamma(z_{defl}^{nom}, \underline{t}, \overline{t})} \tag{2}$$

- J_c (*'comfort'*) – wskaźnik odpowiadający za komfort pasażerów (3). Na wykresie (Rysunek 2) oś odciętych.

$$J_c = \frac{\gamma(\ddot{z}_2, \underline{t}, \overline{t})}{\gamma(\ddot{z}_2{}^{nom}, \underline{t}, \overline{t})} \tag{3}$$

gdzie:
$z_{defl} = z_2 - z_1, i\ \ddot{z}_2$ to odpowiedzi czasowe obiektu,
$z_{defl}^{nom}\ i\ \ddot{z}_2^{nom}$ to odpowiedzi czasowe pasywnego układu redukcji drgań. Wszystkie pomiary były porównywane z odpowiedzią tego układu.

Na rysunku (rys. 2) przedstawiono realizację dwóch kryteriów będących miarą efektywności J zawieszenia. Można nadać kryteriom mierzalny charakter.



Rysunek 2. Kompromis pomiędzy bezpieczeństwem a komfortem jazdy, opracowanie własne na podstawie [8]

Pierwszy problem można rozwiązać poprzez zastosowanie sztywnego zawieszenia. Drugi poprzez zaprojektowanie komfortowego – miękkiego zawieszenia. Te dwa postulaty reprezentują klasyczny konflikt pomiędzy bezpieczeństwem a komfortem jazdy. Jak widać na rysunku (Rysunek 2) zastosowanie sztywnego zawieszenia nie idzie w parze z komfortem jazdy. Podobna sytuacja ma miejsce również w drugim przypadku.

Dla pasywnego układu zawieszenia przy ustalonych wartościach $c\ i\ k$ możliwe jest uzyskanie kompromisowej konfiguracji tych parametrów. W odniesieniu do współczesnych wymagań stawianych zawieszeniom pojazdu kompromis ten nie zapewnia zarówno wymaganego komfortu, jak i bezpieczeństwa. Semi-aktywne i aktywne metody redukcji drgań mają za zadanie pogodzić te dwa sprzeczne kryteria, a w szczególności:

- zapewnić pasażerowi odpowiedni poziom komfortu;
- utrzymać stabilność pojazdu przy różnych warunkach;
- kontrolować położenie masy wibroizolowanej względem koła;
- minimalizować przemieszczeniową funkcję przenoszenia drgań masy resorowanej.

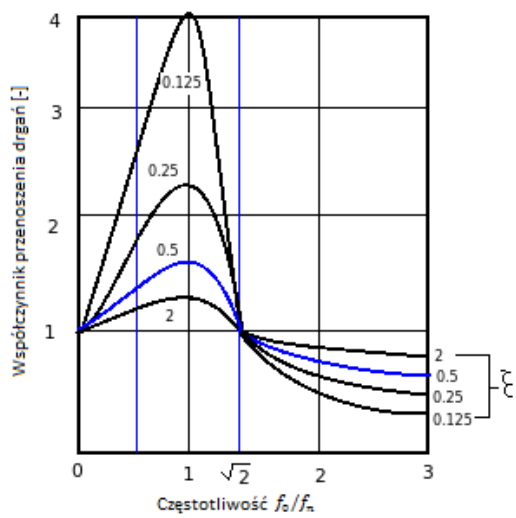
Dobór parametru c wiążę się bezpośrednio z komfortem jazdy. W układach drgających wartość bezwymiarowego współczynnika tłumienia ζ (4) determinuje zachowanie systemu.

$$\zeta = \frac{c}{2\sqrt{mk}} \tag{4}$$

W zależności od wartości parametru ζ można powiedzieć, że układ redukcji drgań:

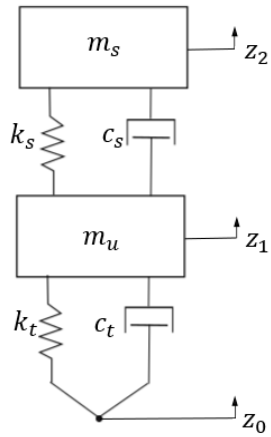
- jest silnie tłumiony, dla $\zeta > 1$,
- osiąga tłumienie krytyczne, dla $\zeta = 1$,
- jest słabo tłumiony, dla $0 < \zeta < 1$,
- nie jest tłumiony, dla $\zeta = 0$.

Zwiększając współczynnik tłumienia c, który jest bezpośrednio związany z bezwymiarowym współczynnikiem tłumienia ζ, obserwujemy zmniejszenie wartości współczynnika przenoszenia drgań dla częstotliwości wymuszenia f_0 mniejszych od częstotliwości wynikających z teoretycznego warunku wibroizolacji $f_0 < f_n\sqrt{2}$ (rys. 3). Dla wyższych częstotliwości wymuszenia f_0 zwiększenie tłumienia c, wiążę się ze wzrostem wartości współczynnika przenoszenia drgań. Ogólnie rzecz biorąc zwiększenie wartości tłumienia poprawia efektywność układu wibroizolacji dla częstotliwości zbliżonych do częstotliwości drgań własnych układu, pogarszając równocześnie dla częstotliwości wyższych.



Rysunek 3. Funkcja przenoszenia drgań dla różnych wartości bezwymiarowego współczynnika tłumienia ζ, opracowanie własne na podstawie [7]

W rozważaniach praktycznych konieczne jest uwzględnienie dynamiki opony. Powoduje to konieczność stosowania modelu o dwóch stopniach swobody (rys. 4).



Rysunek 4. Model pasywnego zawieszenia pojazdu (2 DOF), opracowanie własne na podstawie [7, 8, 10]

Podstawowym wskaźnikiem stosowanym w ocenie układów wibroizolacji jest współczynnik przenoszenia drgań. Współczynnik ten często rozpatrywany jest w funkcji częstotliwości i w zawieszeniach definiowany jako przemieszczeniowa funkcja przenoszenia drgań [8]. Poniżej przedstawiono przyjęte kryteria oceny rozważanego układu wibroizolacji (Rysunek 4) w postaci transmitancji operatorowej.

Przemieszczeniowa funkcja przenoszenia drgań związana z przemieszczeniem masy wibroizolowanej m_s i wyznaczana jako moduł z transmitancji (5). Służy ona do oceny komfortu jazdy.

$$F_{z_2}(s) = \frac{z_2(s)}{z_0(s)} \tag{5}$$

Minimalizując ten wskaźnik zmniejsza się amplitudę drgań masy wibroizolowanej, co znacząco poprawia wygodę.

Przyspieszeniowa funkcja przenoszenia drgań jest związana z przyspieszeniem masy resorowanej m_s i wyznaczana jako moduł transmitancji (6). Pozwala ona na ocenę komfortu jazdy pasażerów poddanych ekspozycji drgań zgodnie z normą. Ponadto pomiar przyspieszenia bezwzględnego masy wibroizolowanej jest znacznie łatwiejszy do realizacji niż pomiar przemieszczenia bezwzględnego tej masy.

$$F_{\ddot{z}_2}(s) = \frac{z_2(s)s^2}{z_0(s)} \tag{6}$$

Przemieszczeniowa funkcja przenoszenia drgań związana z przemieszczeniem masy niewibroizolowanej m_u i wyznaczana jest jako moduł transmitancji (7). Odpowiada za ruch pionowy, w który wprawiane jest koło podczas pokonywania przeszkód.

$$F_{z_1}(s) = \frac{z_1(s)}{z_0(s)} \tag{7}$$

Maksymalizacja tego wskaźnika powoduje zwiększenie komfortu natomiast w przypadku układów aktywnych może ona zwiększyć pobór energii.

Kryterium ugięcia opony *tire deflection* odpowiada za oddziaływania dynamiczne między pojazdem a drogą i jest definiowane jako moduł z transmitancji opisanej wzorem (8).

$$F_{z_1-z_0}(s) = \frac{z_1(s) - z_0(s)}{z_0(s)} \tag{8}$$

Manewry, takie jak: skręcanie, hamowanie wymagają stałego kontaktu z nawierzchnią (z ang. *road holding*). Poprawienie przyczepności jest możliwe, jeżeli zmiany wartości normalnych obciążeń opony są minimalne. Dzieje się tak dlatego, że siły poprzeczne i wzdłużne generowane na oponie zależą właśnie od normalnego obciążenia opony [7]. Dzięki temu wskaźnikowi można także ocenić docisk koła do nawierzchni drogi. Jest to istotne dla współczesnych systemów bezpieczeństwa pojazdu, takich jak ABS[2], ESP[3], EBD[4], EDS[5] itp., które wymagają stałego kontaktu koła z nawierzchnią.

Kryterium *rattle space* jest wskaźnikiem będącym miarą zmienności prześwitu i wyznaczana jest jako moduł transmitancji (9).

$$F_{z_2-z_1}(s) = \frac{z_2(s) - z_1(s)}{z_0(s)} \tag{9}$$

Różnica $z_2(s) - z_1(s)$ nie może być mniejsza niż skok zawieszenia. Wartość średnia tej różnicy związana jest prześwitem, który często w pojazdach terenowych zadaje kierowca.

Generalnie możemy dokonać podziału wskaźników ze względu na pełnione przez nie funkcje. Pierwsze dwa: przemieszczeniowe i przyspieszeniowe masy wibroizolowanej, odpowiedzialne są za komfort jazdy pasażera. Kolejne trzy: przemieszczeniowe masy niewibroizolowanej, prześwit oraz docisk opony do podłoża, odpowiadają stricte za bezpieczeństwo jazdy samochodem. W przypadku pasywnego zawieszenia poprawa któregokolwiek z wymienionych kryteriów może skutkować pogorszeniem pozostałych. W przypadku zawieszenia sterowanego możliwe jest uzyskanie lepszych wyników dla poszczególnych kryteriów bez negatywnego wpływu na pozostałe.

3. Sterowanie typu sky-hook oraz ground-hook

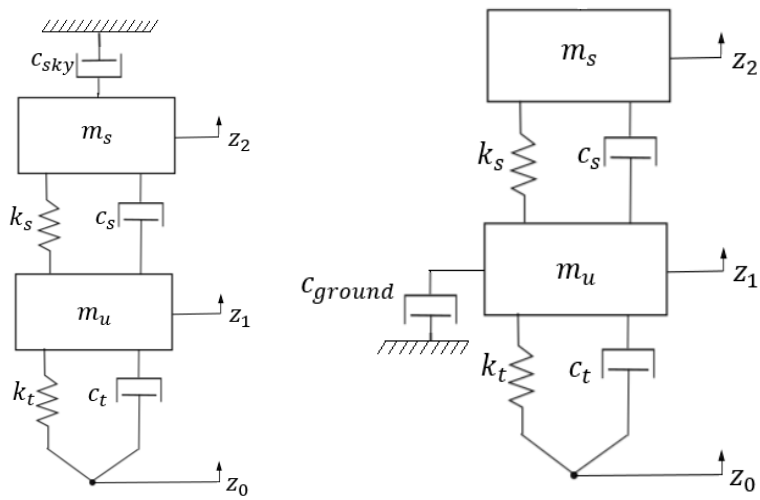
W literaturze z zakresu semi-aktywnych układów redukcji drgań można znaleźć rozwiązania, które mają na celu zapewnienie bezpieczeństwa, komfortu jazdy lub spełniają obydwa te kryteria [1]. Najpopularniejszymi rozwiązaniami są układy sterowania typu: ground-hook, sky-hook lub rozwiązania godzące ze sobą te dwa podejścia.

[2] ABS z ang. *Anti-Lock Braking System.*
[3] ESP z ang. *Electronic stability control*.
[4] EBD z ang. *Electronic Brakeforce Distribution*.
[5] EDS z niem. *Elektronische Differentialsperre*.

3.1. Algorytm sky-hook

Algorytm sterowania typu sky-hook jest najczęściej rozważanym w literaturze algorytmem redukcji drgań w układach semi-aktywnych. Celem sterowania jest utrzymanie masy resorowanej (zahaczenie) w stałej odległości od wirtualnego punktu odniesienia (sky) niezwiązanego z nawierzchnią [3]. Działanie układu jest realizowane poprzez przymocowanie masy wibroizolowanej do pewnego punktu w niebie.

a) b)



Rysunek 5. Struktura elementów w poszczególnych algorytmach sterowania: a) typu sky-hook; b) typu ground-hook, opracowanie własne na postawie [7, 8, 10]

W praktyce sprowadza się to do wprowadzenia siły proporcjonalnej do prędkości bezwzględnej masy wibroizolowanej m_s, a współczynnikiem proporcjonalności jest założony współczynnik c_{Sky}. Dzięki takiej budowie tłumienie pionowych oscylacji nadwozia może być realizowane w sposób niezależny od drgań koła. Schemat obliczeniowy przedstawiony na (rys. 5 a) opisano równaniami (10).

$$\begin{aligned} &m_s\ddot{z}_2 + k_s(z_2 - z_1) + c_s(\dot{z}_2 - \dot{z}_1) + c_{sky}\dot{z}_2 = 0 \\ &m_u\ddot{z}_1 + k_t(z_1 - z_0) + c_t(\dot{z}_1 - \dot{z}_0) - k_s(z_2 - z_1) - c_s(\dot{z}_2 - \dot{z}_1) \\ &\qquad = 0 \end{aligned} \tag{10}$$

Parametr c_{sky} reprezentuje współczynnik tłumienia elementu łączącego nadwozie z zewnętrznym punktem odniesienia (w niebie). Współczynnik tłumienia c_s reprezentuje tłumienie występujące w zawieszeniu. W rzeczywistych aplikacjach semi-aktywnych nie jest możliwe zbudowanie takiego układu redukcji drgań. Największym problemem jest zaczepienie tłumika we wspomnianym punkcie ze względu na to, że sterowane tłumiki generują siłę proporcjonalną do prędkości względnej (tłoka względem cylindra) i skierowaną przeciwnie do niej. Implementacja tego algorytmu w aplikacjach rzeczy-

wistych sprowadza się do zastosowania tłumienia opisanego wzorem (11). W przypadku gdy $\dot{z}_2(\dot{z}_2 - \dot{z}_1) < 0$ należałoby przyjąć wartość współczynnika tłumienia równą 0. Nie jest to możliwe ze względów konstrukcyjnych, przyjmuje się wówczas wartość minimalną możliwą do realizacji c_{min}.

$$c_{vs} = \begin{cases} c_{sky} \dfrac{\dot{z}_2}{\dot{z}_2 - \dot{z}_1} \; dla \; \dot{z}_2(\dot{z}_2 - \dot{z}_1) > 0 \\ c_{min} \; dla \; \dot{z}_2(\dot{z}_2 - \dot{z}_1) \leq 0 \end{cases} \tag{11}$$

Z kolei w przypadku gdy $\dot{z}_2(\dot{z}_2 - \dot{z}_1) \geq 0$ – przyjmuje się wartość współczynnika tłumienia równą założonej wartości c_{sky}. Zwiększanie wartości tłumienia c_{sky} dla idealnego sterowania sky-hook poprawia efektywność układu redukcji drgań. W aplikacjach rzeczywistych wybór zbyt dużych wartości tłumienia c_{sky} może spowodować pogorszenie efektywności redukcji drgań. Wartość tłumienia związana jest z doborem elementu wykonawczego.

Podsumowując zastosowanie algorytmu sterowania typu sky-hook reprezentuje podejście zorientowane na komfort jazdy. Największą wadą wynikającą z zastosowania algorytmu typu sky-hook jest możliwość ograniczenia przyczepności zawieszenia. Wady tej pozbawiony jest algorytm ground-hook.

3.2. Algorytm ground-hook

W celu poprawienia przyczepności pojazdu najczęściej stosowany jest algorytm ground-hook. Jego zadaniem jest zapewnienie ciągłego docisku masy niewibroizolowanej do nawierzchni [10]. Jest to realizowane poprzez generowanie siły dociskającej tę masę do podłoża (rys. 5 b). Analogicznie jak w przypadku sterowania typu sky-hook elementem wykonawczym jest tłumik przymocowany do wirtualnego nieruchomego punktu. Dynamikę rozważanego układu redukcji drgań opisuje układ równań (12), gdzie parametr c_{grd} reprezentuje współczynnik tłumienia elementu łączącego masę niewibroizlowaną z nieruchomym punktem.

$$\begin{aligned} & m_s\ddot{z}_2 + k_s(z_2 - z_1) + c_s(\dot{z}_2 - \dot{z}_1) = 0 \\ & m_u\ddot{z}_1 + k_t(z_1 - z_0) + c_t(\dot{z}_1 - \dot{z}_0) + c_{grd}\dot{z}_1 - k_s(z_2 - z_1) \\ & \qquad - c_s(\dot{z}_2 - \dot{z}_1) = 0 \end{aligned} \tag{12}$$

Tak jak w poprzednim przypadku nie jest możliwe zbudowanie takiego układu redukcji drgań. W rzeczywistych aplikacjach nie sposób przytwierdzić tłumik poruszającego się samochodu na stałe do umownego punktu. Podobnie jak w przypadku sterowania sky-hook implementacja algorytmu w aplikacjach rzeczywistych sprowadza się do zastosowania tłumienia opisanego wzorem (13).

$$c_{vg} = \begin{cases} c_{grd} \dfrac{\dot{z}_1}{\dot{z}_2 - \dot{z}_1} dla \; -\dot{z}_1(\dot{z}_2 - \dot{z}_1) > 0 \\ c_{min} \; dla \; -\dot{z}_1(\dot{z}_2 - \dot{z}_1) \leq 0 \end{cases} \tag{13}$$

Zastosowanie algorytmu typu ground-hook umożliwia efektywną redukcję pionowych oscylacji osi koła, natomiast jedno z głównych zadań, jakim jest skuteczna redukcja drgań masy wibroizolowanej nie zostaje zrealizowane. Zatem algorytm ten poprawia bezpieczeństwo jednak nie gwarantuje komfortu jazdy.

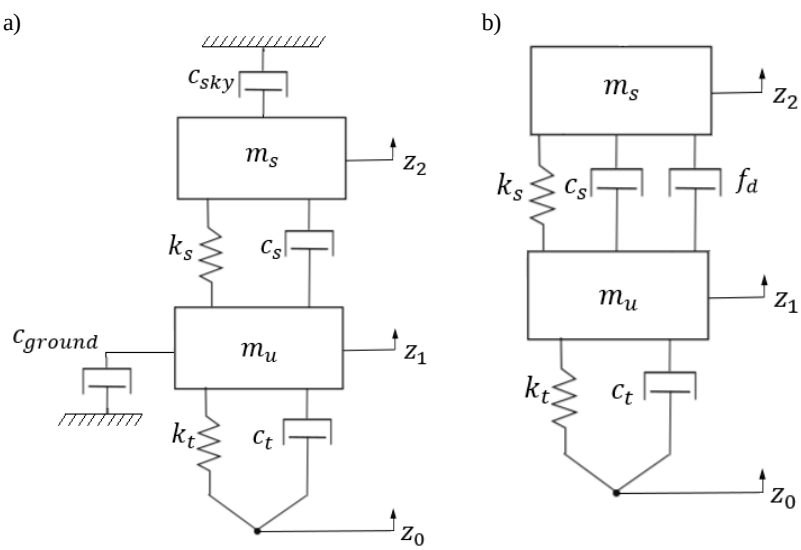
3.3. Algorytm mieszany

Jak już zostało wykazane algorytm sky-hook zapewnia komfort pogarszając przyczepność. Natomiast algorytm ground-hook zapewnia ciągły kontakt opony z podłożem kosztem wygody pasażera. Rozważany algorytm mieszany ma na celu pogodzenie tych dwóch strategii sterowania [10]. Zadaniem takiego algorytmu jest zapewnienie zarówno komfortu, jak i bezpieczeństwa jazdy. Dwa tłumiki mają za zadanie utrzymywać masę wibroizolowaną na stałym poziomie oraz stały kontakt z podłożem masy niewibroizolowanej (rys. 6 a). Jak wspomniano wcześniej w aplikacjach rzeczywistych nie jest możliwe zastosowanie idealnego sterowania zarówno typu sky-hook, jak i ground-hook. Z tego względu stosuje się jeden regulowany tłumik (rys. 6 b). Prawo sterowania tłumikiem sprowadza się do przełączania wartości współczynnika tłumienia c_{vsg} zgodnie ze wzorem (14).

$$f_d = (\alpha c_{sky}\dot{z}_2 + (1-\alpha)c_{grd}\dot{z}_1 \tag{14}$$

Dodatkowo zastosowanie sterowania mieszanego daje możliwość balansowania pomiędzy komfortem a bezpieczeństwem poprzez dobór wagi α:

$$c_{vsg} = \begin{cases} \dfrac{f_d}{\dot{z}_2 - \dot{z}_1} \; dla \; f_d(\dot{z}_2 - \dot{z}_1) > 0 \\ c_{min} \; dla \; f_d(\dot{z}_2 - \dot{z}_1) \leq 0 \end{cases} \tag{15}$$



Rysunek 6. Model układu redukcji drgań sky-ground-hook a) idealny, b) rzeczywisty, opracowanie własne na podstawie [7, 8, 10]

4. Badania porównawcze

Przeprowadzono badania symulacyjne przedstawionych powyżej strategii sterowania. W praktyce implementacja przedstawionych na rysunku (rys. 5 a i b) modeli sprowadza się do zastosowania tłumika sterowanego (rys. 6 b). Do przeprowadzeniu symulacji numerycznych przyjęto model ćwiartkowy pojazdu o dwóch stopniach swobody (2DOF). Uwzględnia on oddziaływania pomiędzy masą resorowaną, nieresorowaną, jak i podłożem. Jako element wykonawczy został wybrany tłumik o zmiennej (nastawialnej) sile tłumienia. Po zamianie zmiennych, zapisano równania równowagi dynamicznej dla takiego układu (16).

$$\begin{gathered} m_s\ddot{x}_3 + k_s(x_3 - x_1) + c_s(\dot{x}_3 - \dot{x}_1) - f_d = 0 \\ m_u\ddot{x}_1 + k_t(x_1 - u) + c_t(\dot{u} - \dot{x}_1) + k_s(x_1 - x_3) + c_s(\dot{x}_1 - \dot{x}_3) \\ +f_d = 0 \end{gathered} \tag{6}$$

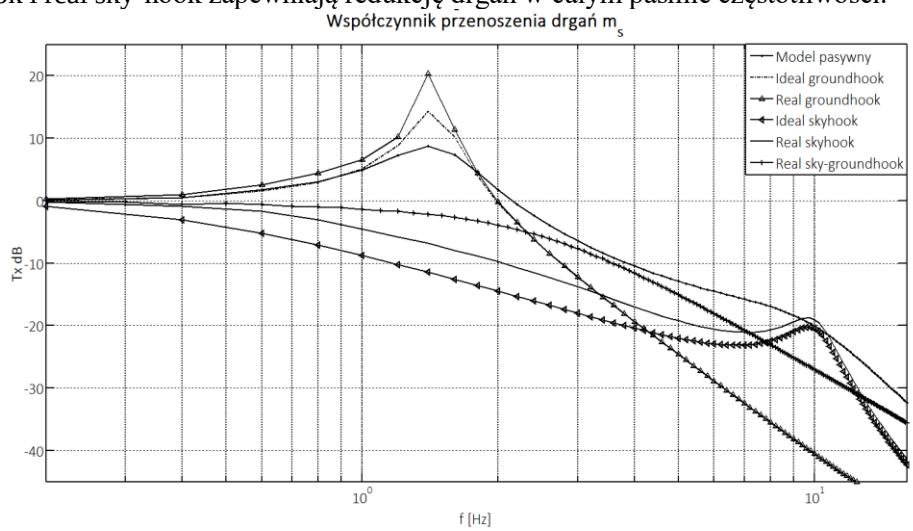
W poniższym rozdziale przyjęto, że minimalna wartość tłumienia dla tego elementu wykonawczego wynosi $c_{min} = 0$. Oprócz sterowanego tłumika uwzględniono równolegle do niego położone tłumienie o stałej wartości wynikające z konstrukcji mechanicznej układu. Obydwa elementy odzwierciedlają rzeczywisty charakter działania modelu. Dla tak przyjętego układu zaimplementowano algorytmy sterowania typu ground-hook, sky-hook lub mieszany.

Przeprowadzono badania symulacyjne działania poszczególnych algorytmów. Jako model odniesienia przyjęto zawieszenie pasywne. Wybór wartości współczynnika tłumienia c_{pas} był kompromisem pomiędzy zawieszeniem sztywny (pojazdy robocze) a komfortowym (luksusowe limuzyny). Ostatecznie przyjęto $c_{pas} = 500\ \left[\frac{Ns}{m}\right]$. W przeprowadzonych badaniach numerycznych narzucono ograniczenia na sygnał sterujący wynikające z własności fizycznych elementu wykonawczego (tłumik) realizując w ten sposób sterowanie semi-aktywne. Do symulacji numerycznych przyjęto parametry jak w tabeli 1.

Tabela 1. Parametry obiektu [opracowanie własne]

m_u	11,5 $[kg]$
m_s	86 $[kg]$
k_s	8,65 $[\frac{kN}{m}]$
k_t	37,4 $[\frac{kN}{m}]$
c_t	72 $[\frac{Ns}{m}]$
c_s	100 $[\frac{Ns}{m}]$
c_{sky}	3000 $\left[\frac{Ns}{m}\right]$
c_{ground}	2000 $[\frac{Ns}{m}]$

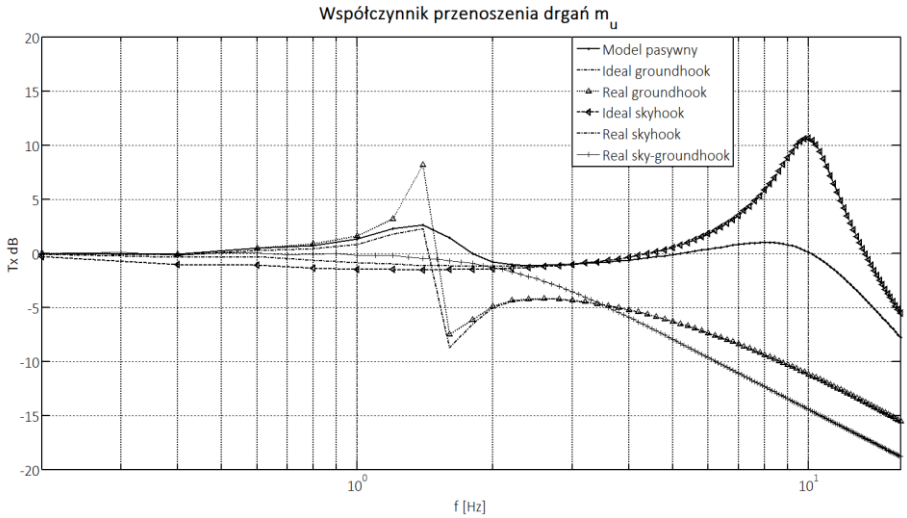
Jako wymuszenie przyjęto sygnał sinusoidalny o amplitudzie $0{,}03\ [m]$ oraz częstotliwości z zakresu $< 0{,}25, 16 > [Hz]$. Na wykresie (rys. 7) możemy zaobserwować pierwsze przyjęte kryterium, które jest miarą komfortu jazdy. Przemieszczeniowa funkcja przenoszenia drgań została wybrana jako sposób weryfikacji systemu ze względu na jego nieliniowy charakter [9]. Na osi rzędnych wykresu (rys. 7) przedstawiono wzmocnienie w dB. Model pasywny w rozpatrywanym przedziale częstotliwości osiąga najmniejsze wartości współczynnika przenoszenia drgań. W okolicach pierwszej częstotliwości rezonansowej największe wzmocnienie osiąga układ sterowany regulatorem rzeczywisty ground-hook $(20{,}32)$. Następnie plasuje się idealny model ground-hook $(14{,}35)$. Należy w tym miejscu podkreślić, że dwa wymienione algorytmy w okolicy drugiej częstotliwości rezonansowej nie powodują zwiększenia wartości współczynnika przenoszenia drgań. Charakterystyka przyjmuje podobny kształt dla algorytmu idealnego i rzeczywistego typu sky-hook. Ponadto dla układu sterowanego algorytmem idealnym typu sky-hook współczynnik przenoszenia drgań przyjmuje wartości mniejsze w całym przedziale częstotliwości. Zastosowanie algorytmu realizującego sterowanie mieszane sky-ground-hook w okolicy pierwszego, jak i drugiego rezonansu nie powoduje wzmocnienia drgań. Sterowania: sky-ground-hook, ideal sky-hook i real sky-hook zapewniają redukcję drgań w całym paśmie częstotliwości.



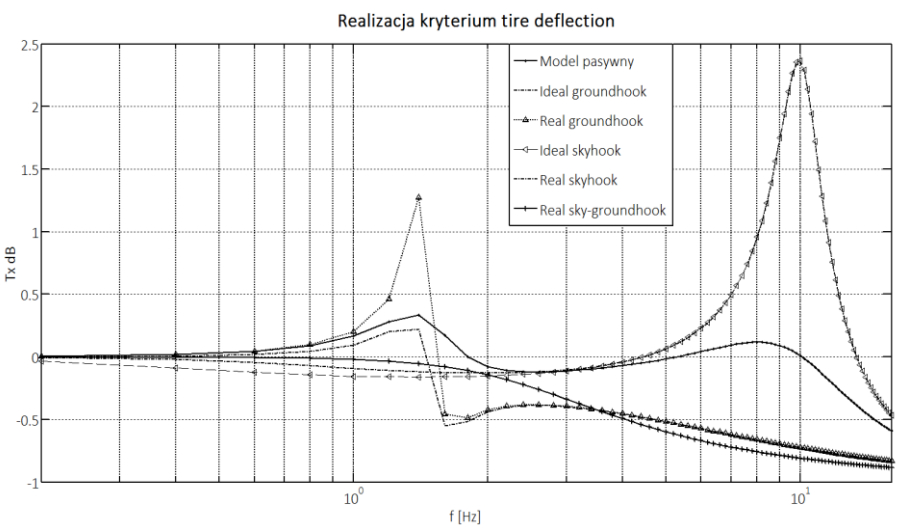
Rysunek 7. Przemieszczeniowa funkcja przenoszenia drgań – porównanie realizacji poszczególnych algorytmów [opracowanie własne]

Przemieszczeniową funkcję przenoszenia się drgań z wymuszenia z_0 na masę niewibroizolowaną m_u odpowiadającą za bezpieczeństwo w trakcie jazdy przedstawiono na rysunku 8. Funkcje te przedstawiono dla wszystkich wymienionych wcześniej algorytmów sterowania. W okolicy pierwszego rezonansu obserwujemy wzmocnienie dla układu pasywnego $(2{,}64)$ oraz układów sterowanych regulatorami rzeczywisty ground-hook $(8{,}16)$ i idealny ground-hook $(2{,}28)$. Układy sterowane algorytmami idealny sky-hook i rzeczywisty sky-hook $(10{,}36)$ w okolicy drugiego rezonansu osiągają największą wartość wzmocnienia. Największą redukcję drgań w okolicy drugiej częstotliwości drgań własnych uzyskano dla układu sterowanego algorytmem mieszanym $(-14{,}35)$. Układy sterowane algorytmami idealny sky-hook oraz mieszany w okolicy pierwszej

częstotliwości drgań własnych osiągają wartości funkcji przenoszenia drgań mniejsze od 0, co oznacza zmniejszenie drgań na wyjściu w stosunku do wymuszenia. Układ z regulatorem mieszanym redukuje drgania masy niewibroizolowanej w całym zakresie częstotliwości.



Rysunek 8. Przemieszczeniowa funkcja przenoszenia drgań dla masy nieresorowanej [opracowanie własne]



Rysunek 9. Realizacja kryterium tire deflection [opracowanie własne]

Ze względu na fakt, że rozpatrywana jest składowa dynamiczna sygnału ugięcia opony podobnie jak w poprzednich przypadkach należy minimalizować wskaźnik funkcji przenoszenia drgań. W przypadku docisku koła do podłoża (rys. 9) podobnie jak w poprzednich kryteriach sterowanie mieszane w całym rozważanym paśmie częstotliwości osiąga ujemne wartości bezwymiarowego współczynnika tłumienia. Następnie najlepszy docisk koła do nawierzchni uzyskują kolejno układy sterowane

regulatorami ground-hook: idealny i rzeczywisty. Dla tych układów w okolicy pierwszej częstotliwości drgań własnych obserwujemy (odpowiednio $0{,}22\ i\ 1{,}4$) wzmocnienie.

Niezależnie od wybranego kryterium można zaobserwować, że regulatory, których głównym zadaniem jest uzyskanie największego komfortu, uzyskują najniższe wartości funkcji przenoszenia drgań dla niskich częstotliwości. Natomiast dla wysokich częstotliwości uzyskują najwyższe wartości funkcji przenoszenia drgań z wszystkich badanych przypadków. Realizacja metod nastawionych na bezpieczeństwo jazdy wykazuje odwrotne właściwości: dla niskich częstotliwości wykazują najwyższe wartości funkcji przenoszenia drgań, a dla wysokich częstotliwości uzyskują najmniejsze wartości tej funkcji. Regulator bazujący na modelu mieszanym sky-ground-hook łączy najlepsze cechy wymienionych wcześniej sterowań. Sterowanie mieszane w całym paśmie częstotliwości osiąga ujemne wartości przyjętych wskaźników komfortu i bezpieczeństwa.

5. Podsumowanie

Charakterystyczną cechą układów o dwóch stopniach swobody są dwie częstotliwości własne. Dla zawieszeń pojazdów kołowych zawierają się one w paśmie $< 0.1{,}16 > [Hz]$. W celu poprawienia komfortu oraz bezpieczeństwa jazdy należy dążyć do jak największej redukcji drgań w tym paśmie częstotliwości przy jednoczesnym zachowaniu jak największego ugięcia opony (docisku koła do nawierzchni). Jak zostało pokazane zawieszenie sterowane algorytmem sky-hook dobrze redukuje drgania w okolicach pierwszej częstotliwości własnej. W okolicach drugiej częstotliwości można zauważyć nieznaczne wzmocnienie. Tam zdecydowanie lepiej redukuje drgania algorytm sterowania ground-hook. Jeżeli chcemy uzyskać tłumienie drgań w całym paśmie przenoszenia najlepiej zastosować sterowanie mieszane. Dla tego algorytmu obserwujemy przebieg przemieszczeniowej funkcji przenoszenia drgań praktycznie bez żadnego wzmocnienia ($T_x dB < 0$).

Planuje się przeprowadzenie implementacji algorytmu sterowania ślizgowego z modelem referencyjnym typu sky-hook i ground-hook.

Podziękowania

Praca została sfinansowana ze środków z Grantu dziekańskiego (grant nr 15.11130.183).

Literatura

1. Dupont P., Stokes A., Street C., *Semi-active Control*, Journal of Sound and Vibration, 1995, s. 203-218.
2. Edwards C., Spurgeon S.K., *Sliding mode control: theory and applications*, Taylor & Francis, 1998.
3. Karnopp D., Crosby M.J., Harwood R.A., *Vibration Control Using Semi-Active Force Generators*, ASME. J. Eng. Ind., 96 (May 1974), s. 619-626, http://doi.org/10.1115/1.3438373.
4. Karnopp D., *Vehicle dynamics, stability, and control*, CRC Press, 2013.
5. Kowal J., *Sterowanie drganiami*, „Gutenberg", 1996.
6. Porumamilla H., *Modeling , analysis and non-linear control of a novel pneumatic semi-active vibration isolator: a concept validation study*, Retrospective Theses and Dissertations, Paper 15765, 2007.
7. Rajamani R., *Vehicle Dynamics and Control*, Springer, 2006, http://doi.org/10.1007/978-1-4614-1433-9_2.

8. Savaresi S., Poussot-Vassal C., Spelta C., *Semi-Active Suspension Control Design for Vehicles*, Butterworth-Heinemann, 2010.
9. Sibielak M., Raczka W., Konieczny J., Kowal J., *Optimal control based on a modified quadratic performance index for systems disturbed by sinusoidal signals*, Mechanical Systems and Signal Processing 64-65, 2015, s. 498-519.
10. Valášek M., Kortüm W., Šika Z., Magdolen L., Vaculín O., *Development of semi-active road-friendly truck suspensions*, Control Engineering Practice, 6(6), 1998, s. 735-744.
11. Yokoyama M., Hedrick J.K., Toyama S., *A model following sliding mode controller for semi-active suspension systems with MR dampers*, Proceedings of the 2001 American Control Conference, (Cat. No.01CH37148), 4, 2001, s. 2652-2657.

Przegląd algorytmów sterowania zawieszeniem pojazdu

Streszczenie
W poniższym rozdziale opisano syntezę algorytmów sterowania semi-aktywnym zawieszeniem pojazdu. Rozważono ćwiartkowy model zawieszenia pojazdu kołowego. Rozważono algorytmy sterowania: sky-hook, ground-hook, algorytm mieszany oraz dla porównania uwzględniono pasywne zawieszeni pojazdu. Zbadano wpływ zmiany rodzaju sterowania na efektywność układu wibroizolacji modelu ćwiartkowego zawieszenia. Każdy algorytm był testowany pod kątem trzech kryteriów: przemieszczeniowa funkcja przenoszenia drgań, kryterium ugięcia opony (*'tire deflection'*) oraz kryterium (*'rattle space'*).
Słowa kluczowe: zawieszenie semi-aktywne, sky-hook, ground-hook, algorytm mieszny

Survey of vehicle suspension control algorithms

Abstract
The synthesis of control algorithms for the semi-active suspension of the vehicle is described in the following section. A quadrant suspension model of a wheeled vehicle was considered. The control algorithms: sky-hook, ground-hook, mixed algorithm, and comparison were considered passive vehicle suspensions. The impact of change of control type on the effectiveness of the vibration isolation system of the quarter suspension model was studied. Each algorithm was tested for three criteria: displacement vibration, tire deflection, and rattle space.
Keywords: semi-active suspension, sky-hook, ground-hook, mixed algorithm

Andrzej Tutaj[1], Paweł Król[2]

Dobór suboptymalnych czasów realizacji elementarnych ruchów manipulatora w urządzeniu przygotowującym leki cytostatyczne

1. Wprowadzenie

Jednym z klasycznych przykładów optymalizacji pracy systemów dynamicznych jest poszukiwanie sterowania czasooptymalnego. Znajduje ono zastosowanie wszędzie tam, gdzie minimalizacja czasu realizacji procesu jest celem zasadniczym, zaś czynniki, takie jak wydatek energii czy zużycie sprzętu mają znaczenie drugorzędne. W przypadku przemysłowych procesów produkcyjnych, optymalizacja czasu jest stosowana przede wszystkim dla linii montażowych i podobnych, realizujących operacje o charakterze dyskretnym.

Interesującym przykładem pozaprzemysłowej instalacji realizującej proces dyskretny, w którym pożądana jest optymalizacja czasu operacji, jest urządzenie przeznaczone do przygotowania roztworów leków cytostatycznych. Wyręcza ono człowieka z uciążliwego i niebezpiecznego dla zdrowia zadania (cytostatyki są wysoce toksyczne), jednocześnie ograniczając ryzyko groźnych w skutkach błędów i pomyłek. W ramach pojedynczego cyklu produkcyjnego urządzenie pobiera z fiolki określoną przez lekarza porcję skondensowanego leku i rozprowadza ją w worku z płynem fizjologicznym do infuzji. Realizowany jest przy tym szereg dodatkowych czynności o charakterze pomocniczym i kontrolnym [1-3].

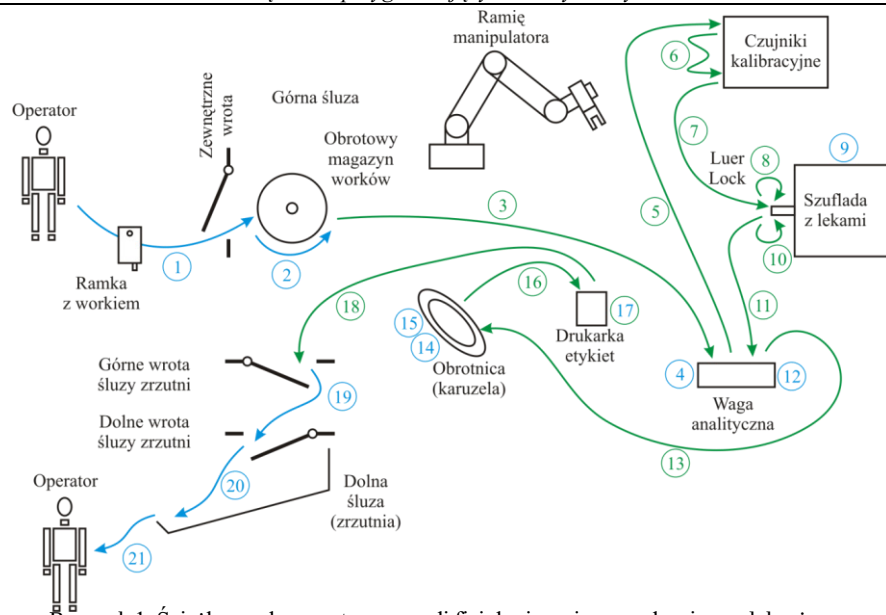
W ramach projektu badawczego CYTOROBOT, zrealizowanego w Akademii Górniczo-Hutniczej, opracowano i zbudowano stanowisko demonstracyjne o funkcjonalności urządzenia do przygotowania leków cytostatycznych [4]. Choć z założenia ma ono charakter przedprototypu (*proof of concept*), to jednak w większości oparte jest o w pełni funkcjonalne, profesjonalne podzespoły i rozwiązania. Jednym z głównych jego elementów jest uniwersalny manipulator przemysłowy Universal Robot 5 (UR5) [5]. Jego zadaniem jest przenoszenie worka z roztworem soli fizjologicznej, umocowanego w ramce usztywniającej, między kolejnymi stanowiskami (gniazdami obróbczymi) wchodzącymi w skład instalacji. Ścieżkę, którą pokonuje worek, przedstawiono schematycznie na rysunku 1., który pokazuje również główne elementy urządzenia. Sumaryczny czas transportu międzyoperacyjnego, realizowanego przez manipulator, stanowi istotny składnik czasu wykonania zadania produkcyjnego. Uproszczony harmonogram takiego pojedynczego zadania, nieuwzględniający szeregu czynności przygotowawczych i pomocniczych, jest następujący:

1. Umieszczenie ramek z workami i zainstalowanymi złączami Luer lock w magazynie górnej śluzy powietrznej. Czynność tę realizuje operator urządzenia.

[1] tutaj@agh.edu.pl, AGH Akademia Górniczo-Hutnicza, al. A. Mickiewicza 30, 30-059 Kraków, https://www.agh.edu.pl.
[2] pakrol@agh.edu.pl, AGH Akademia Górniczo-Hutnicza, al. A. Mickiewicza 30, 30-059 Kraków, https://www.agh.edu.pl.

2. Obrót magazynka, pełniącego zarazem rolę wewnętrznych wrót śluzy. Śluza zapobiega wydostawaniu się par leku z urządzenia produkcyjnego do pomieszczenia, w którym się ono znajduje.
3. Pobranie ramki z workiem z magazynu i umieszczenie jej na precyzyjnej wadze elektronicznej. Czynność tę wykonuje manipulator UR5.
4. Ważenie ramki z workiem.
5. Pobranie ramki z workiem z wagi i przeniesienie jej w okolicę zasięgu laserowych czujników optycznych używanych do kalibracji; jest ona niezbędna ze względu na nieuchronny rozrzut parametrów geometrycznych chwytu ramki i montażu worka.
6. Przeprowadzenie procedury kalibracji, służącej określeniu i kompensacji mimośrodu – przesunięcia osi symetrii złącza Luer lock względem osi 0z układu współrzędnych związanego z kołnierzem ostatniego członu kiści manipulatora.
7. Przeniesienie ramki z workiem od czujników kalibracyjnych w okolicę wybranego złącza Luer lock wybranej szuflady z fiolkami zawierającymi skondensowane leki. Urządzenie posiada wiele szuflad, z których każda mieści wiele fiolek.
8. Połączenie złącza Luer lock worka i szuflady, realizowane przez dociśnięcie i obrót.
9. Dozowanie leku za pomocą wielokanałowej pompy perystaltycznej zainstalowanej w szufladzie.
10. Rozłączenie złącza Luer lock.
11. Przeniesienie worka z okolicy szuflady na wagę.
12. Powtórne ważenie.
13. Przeniesienie worka z wagi na obrotnicę (karuzelę), gdzie realizowane jest mieszanie roztworu bądź zawiesiny i wykrywanie obecności zanieczyszczeń stałych.
14. Mieszanie zawartości worka.
15. Rozpoznawanie zanieczyszczeń przez komputerowy system wizyjny.
16. Przeniesienie worka z obrotnicy w pobliże drukarki etykiet.
17. Wydruk etykiety i przylepienie jej do powierzchni worka.
18. Przeniesienie worka z okolicy drukarki na stolik zamocowany na górnej powierzchni górnych wrót dolnej śluzy (zrzutni). Śluza powietrzna zapobiega wydostaniu się par leku z wnętrza gabloty urządzenia do pomieszczenia, w którym przebywa personel medyczny (farmaceuta, lekarz, pielęgniarka).
19. Otwarcie i zamknięcie górnych wrót śluzy w celu wprowadzenia worka do śluzy.
20. Otwarcie i zamknięcie dolnych wrót dolnej śluzy w celu wprowadzenia worka do końcowego odcinka zrzutni.
21. Pobranie worka przez operatora.

Manipulator zaangażowany jest w realizację dziesięciu spośród dwudziestu jeden punktów harmonogramu (3, 5, 6, 7, 8, 10, 11, 13, 16, 18). Czas realizacji poszczególnych zadań transportowych wykonywanych przez manipulator ma zatem istotny udział w ogólnym czasie realizacji kompletnego zadania produkcji. Optymalizacja czasowa poszczególnych etapów przenoszenia worka jest więc istotna dla zapewnienia efektywności działania urządzenia.



Rysunek 1. Ścieżka worka z roztworem soli fizjologicznej w urządzeniu produkcyjnym

W początkowej części niniejszego opracowania skrótowo opisana zostanie konstrukcja użytego manipulatora UR5, sprzężonego z nim sterownika oraz metody programowania tego zestawu. Następnie przedstawione zostanie zadanie optymalizacji czasu ruchu robota, rozważane w ramach realizacji projektu badawczego. Dalej objaśnione zostaną narzędzia i metody użyte do rozwiązania tego zadania. Następnie zaprezentowane będą uzyskane wyniki symulacji komputerowych oraz eksperymentów przeprowadzonych na rzeczywistym urządzeniu. Całość zamknie podsumowanie.

2. Budowa i metody programowania robota

Zestaw robotyczny Universal Robots o udźwigu 5 kg (UR5) składa się z przegubowego ramienia z serwomechanizmami elektrycznymi, sterownika (kontrolera) oraz panelu operatorskiego (teach pendant). Ramię UR5 (rys. 2), którego konstrukcja nośna wykonana jest z rur aluminiowych, posiada sześć stopni swobody, realizowanych przez sześć przegubów walcowych. Kolejnym przegubom nadano następujące nazwy: podstawa (base), ramię (shoulder), łokieć (elbow), nadgarstek (wrist) 1, nadgarstek 2, nadgarstek 3.

Rysunek 2. Ramię manipulatora UR5.
Źródło: Service Manual Robot: UR5 with CB2-controller [6]

We wnętrzu ramienia zainstalowane są przekładnie harmoniczne pośredniczące w przekazaniu napędu, silniki PMSM, pełniące funkcję urządzeń wykonawczych (aktuator), podwójne zestawy enkoderów, wykorzystywanych w roli sensorów położenia kątowego i układy automatyki sterowania bezpośredniego wraz z końcówkami mocy, realizujące algorytm sterowania wektorowego (*field oriented control* – FOC). W skrzynce sterownika CB2 (Control Box 2) znajdują się między innymi: zasilacz, obciążenie aktywne, moduł układów bezpieczeństwa oraz wejść i wyjść analogowych i cyfrowych, a także komputer klasy PC wykonany w standardzie Mini ITX. Ze sterownikiem współpracuje kaseta sterownicza (*teach pendant*), wyposażona w monitor z nakładką dotykową, przycisk bezpieczeństwa (grzybek), przycisk nauki (*teach button*), przycisk włączania zasilania zestawu oraz gniazdo typu A magistrali USB.

Zestaw robotyczny może pracować całkowicie autonomicznie i jest w stanie samodzielnie realizować złożone i rozbudowane zadania, po wcześniejszym odpowiednim zaprogramowaniu. Z drugiej strony, przewidziane są również mechanizmy zdalnego sterowania jego pracą, w trakcie której robot realizuje na bieżąco przesyłane zdalnie polecenia. Wykorzystuje się do tego komunikację przez sieć TPC/IP i protokół umożliwiający dwukierunkową wymianę danych. Robot odbiera i realizuje wysokopoziomowe rozkazy oraz odsyła w czasie rzeczywistym wyczerpujące informacje o stanie procesu. System robotyczny jest więc elastyczny i dopuszcza różne przypadki użycia. Kontroler robota może zaś pełnić w układzie automatyki bądź rolę sterownika samodzielnego lub nadrzędnego, bądź też być umieszczony na niższym poziomie hierarchii całościowego systemu sterowania.

Algorytmy generowania trajektorii robota oraz rozwiązywania odwrotnego i prostego zadania kinematyki i dynamiki są realizowane przez oprogramowanie URControl, pracujące pod kontrolą systemu operacyjnego Linux, na komputerze klasy PC, wchodzącym w skład sterownika CB2. Z aplikacją URControl współpracuje program PolyScope, pełniący funkcję graficznego interfejsu użytkownika (GUI). PolyScope umożliwia tworzenie programu z poziomu edytora graficznego i pozwala na wykorzystanie trybu uczenia. Operator może programować kolejne punkty trajektorii bądź wprowadzając ich współrzędne, bądź doprowadzając do nich ramię robota. Ta ostatnia czynność może być realizowana przy pomocy przycisków kierunkowych na klawiaturze ekranowej lub w trybie uczenia, przez ręczne wywieranie siły na manipulator, przy jednoczesnym przytrzymaniu przycisku uczenia na kasecie sterowniczej.

Zgodnie z dokumentacją robota, istnieją trzy sposoby bądź poziomy sterowania czy programowania robota: poziom interfejsu graficznego (GUI level) [7], poziom języka skryptowego (script level) [8] i poziom interfejsu programistycznego w języku C (C API level). W czasie realizacji opisywanego projektu wykorzystywano dwa pierwsze.

Programowanie na poziomie GUI jest realizowane w środowisku PolyScope. Program tak utworzony jest wyświetlany w postaci graficznego drzewa. Przygotowanie programu jest stosunkowo proste, a programista może korzystać z trybu uczenia. Zbiór dostępnych konstrukcji programistycznych jest jednak nieco ograniczony, przez co podejście to nie jest elastyczne.

Programowanie na poziomie skryptowym jest realizowane przez przygotowanie programu w postaci pliku tekstowego i przesłanie go do aplikacji URControl. Ten sposób sterowania robota jest o wiele elastyczniejszy od programowania z poziomu GUI i daje programiście większą swobodę tworzenia aplikacji.

Program tworzony jest w języku URScript [8], oferującym zbiór specjalistycznych funkcji dostosowanych do specyfiki zadania sterowania robotem. W języku tym dostępne są typowe elementarne operacje arytmetyczne i logiczne oraz popularne konstrukcje sterujące. Oprócz podstawowych typów danych, znanych z innych języków programowania, występują też typy specyficzne, jak choćby poza określająca pozycję (*position*) i orientację (*attitude*) końcówki roboczej narzędzia. Język URScript umożliwia definiowanie wątków i realizację obliczeń równoległych.

Kompletny program (skrypt) lub pojedyncze polecenia mogą być przesłane do aplikacji URControl z zewnętrznego komputera za pośrednictwem sieci TCP/IP. URControl pełni bowiem rolę serwera, nasłuchującego na porcie TCP 3002, gotowego do przyjęcia poleceń w języku URScript. Ten mechanizm komunikacji nosi nazwę Secondary Client. Inny mechanizm, zwany Real Time Client, umożliwia otrzymanie ze sterownika strumienia komunikatów płynących z częstotliwością 125 Hz i niosących próbki wielkości fizycznych opisujących ruch i stan ramienia, takich jak położenia, prędkości, przyspieszenia, momenty siły, prądy czy temperatury serwomechanizmów.

Warto wspomnieć, że istnieje możliwość mieszania dwóch sposobów programowania: graficznego i skryptowego. W programie tworzonym w środowisku PolyScope można umieszczać bądź jednolinijkowe wstawki w języku skryptowym, bądź odwołania do rozbudowanych skryptów zapisanych w odrębnych plikach dyskowych w komputerze sterownika. To podejście pozwala połączyć łatwość programowania w środowisku graficznym z elastycznością i dużymi możliwościami kodowania w języku skryptowym.

W opisywanym projekcie skorzystano z możliwości łączenia dwóch sposobów programowania. Szkielet aplikacji zrealizowano w środowisku graficznym PolyScope, zaś poszczególne fragmenty programu robota zapisano w skryptach, do których odwołuje się program główny. Ze względu na konieczność współdziałania manipulatora z innymi, współpracującymi urządzeniami cytorobota, wykorzystano możliwości komunikacji sieciowej między sterownikiem CB2 a nadrzędnym komputerem zarządzającym urządzenia produkcyjnego. Komunikację tę realizują stosowne funkcje języka skryptowego, działające w trybie blokującym z przeterminowaniem (*timeout*). Dzięki temu uzyskano wymaganą synchronizację zadań i możliwość zawieszania programu robota na czas realizacji czynności obsługowych na stanowiskach obróbczych.

3. Programowanie ruchów manipulatora

Jedną z podstawowych operacji elementarnych, realizowanych przez manipulator, jest tak zwany *ruch* (move), czyli zachodząca w czasie zmiana konfiguracji (ułożenia) ramienia. Ruch jest charakteryzowany przede wszystkim przez punkt początkowy, punkt końcowy, rodzaj czy typ ruchu (omówione nieco dalej), a także przyspieszenia i prędkości maksymalne – kątowe bądź liniowe. Większość poleceń (funkcji) realizujących ruchy jest dostępna zarówno na poziomie programowania graficznego w PolyScope, jak i w języku skryptowym URScript. W środowisku graficznym brak jest jednak funkcji *movec*, odpowiedzialnej za ruch po łuku.

Przy definiowaniu ruchów w języku skryptowym, punkty: początkowy i końcowy mogą być podawane na jeden z dwóch sposobów: jako poza (pozycja i orientacja) w kartezjańskim układzie współrzędnych związanych z podstawą robota lub jako zestaw współrzędnych przegubowych. Poza jest wektorem sześcioelementowym. Trzy pierwsze elementy określają pozycję (położenie liniowe – współrzędne x, y i z) centralnego

punktu narzędzia (TCP – *tool centre point*) i są wyrażone w metrach. Domyślnie, punkt TCP jest ulokowany w początku układu współrzędnych związanego z kołnierzem mocowania narzędzia na ostatnim członie robota. Trzy pozostałe elementy pozy definiują orientację narzędzia (a tym samym – ostatniego członu robota) z użyciem konwencji oś-kąt: kierunek i zwrot opisanego nimi wektora określa oś i kierunek obrotu, zaś długość wektora – kąt obrotu wyrażony w radianach. Domyślnie, układy współrzędnych narzędzia (a dokładniej – TCP) i kołnierza ostatniego członu robota pokrywają się. Ich wzajemne położenie liniowe i kątowe (pozycja i orientacja) może być jednak zmieniane programowo, w trakcie wykonania aplikacji (na przykład funkcją *set_tcp* języka URScript) lub zdefiniowane w parametrach wstępnej konfiguracji robota, zapisanych w pliku określającym sposób jego *instalacji*.

Każdemu punktowi w przestrzeni przegubowej odpowiada dokładnie jedna poza. Z drugiej strony, danej pozie może odpowiadać więcej niż jedna konfiguracja ramienia (i tym samym – więcej niż jeden punkt w przestrzeni przegubowej). Możliwa jest również sytuacja, gdy konfiguracja realizująca daną pozę nie istnieje. Wynika to wprost z faktu istnienia i jednoznaczności rozwiązania prostego zadania kinematyki oraz niejednoznaczności i możliwości nieistnienia rozwiązania zadania odwrotnego.

W języku skryptowym dostępne są różne funkcje realizujące różne typy ruchów, na przykład: *movej*, *movel*, *movec*, *movep*. W niniejszym projekcie wykorzystano dwie z nich: *movej* oraz *movel*. Obie realizują liniową interpolację trajektorii między punktem początkowym i końcowym. W przypadku *movel*, interpolacji podlegają współrzędne TCP w przestrzeni kartezjańskiej związanej z podstawą robota. Oznacza to, że punkt TCP porusza się po linii prostej. Jednak konfiguracja ramienia może zmieniać się podczas takiego ruchu w sposób trudny do przewidzenia i nieregularny. W szczególności, zakresy i prędkości ruchu poszczególnych punktów ramienia robota mogą być znacznie większe niż końcówki TCP, a przebiegi czasowe prędkości ruchu obrotowego poszczególnych przegubów mogą wykazywać znaczną zmienność w czasie. Z kolei funkcja *movej* dokonuje interpolacji liniowej w przestrzeni współrzędnych przegubowych. Z punktu widzenia algorytmu sterowania zaimplementowanego w sterowniku, zadanie to jest prostsze w realizacji, a ramię porusza się harmonijniej. Programista ma też większą kontrolę nad ewolucją konfiguracji w czasie. Jednakże tor TCP podczas takiego ruchu nie jest zwykle prostoliniowy.

Zgodnie z zaleceniami producenta robota, w opisywanym projekcie, funkcję *movel* wykorzystano, ilekroć manipulator wykonuje ruch, w którym występuje kontakt elementów przenoszonej ramki lub worka z elementami umieszczonymi w gablocie lub regale bądź istotne zbliżenie do tych elementów. Pełna kontrola nad trajektorią ramki w przestrzeni kartezjańskiej, którą daje interpolacja liniowa ruchu funkcją *movel*, pozwala wyeliminować ryzyko kolizji chwytaka z elementami wewnątrz gabloty urządzenia. Z kolei z funkcji *movej* skorzystano przy realizacji znacznych przemieszczeń pomiędzy kolejnymi stanowiskami pracy. Pełniejsza kontrola nad konfiguracją ramienia i jednoznaczność jego trajektorii, którą daje ten typ ruchu, pozwoliły oddalić ryzyko kolizji poszczególnych członów i przegubów robota z elementami obudowy i wyposażenia urządzenia produkcyjnego. Przegubem robota, który z największym prawdopodobieństwem może uczestniczyć w kolizji, jest łokieć, a ryzyko zderzenia występuje zwłaszcza w sytuacjach, gdy robot operuje końcówką w pobliżu swojej podstawy.

4. Zadanie sterowania czasooptymalnego

Jednym z podstawowych wskaźników jakości rozważanego procesu produkcyjnego jest czas realizacji zadania produkcji leku. Pożądane jest uzyskanie czasu jak najkrótszego. Można to osiągnąć między innymi przez optymalizację czasów realizacji poszczególnych etapów harmonogramu. Istotny udział w ogólnym czasie produkcji mają czynności międzyoperacyjnego transportu ramki z workiem. Są one realizowane przez manipulator, a każda z nich opisana jest ścieżką składającą się z ciągu elementarnych ruchów ramienia. W niniejszej pracy skupimy się na minimalizacji czasu pokonania takiej złożonej ścieżki.

Zadanie poszukiwania sterowania optymalnego dla manipulatorów nie jest zagadnieniem nowym. Prace naukowe poświęcone temu tematowi prowadzono już w latach osiemdziesiątych [9-12]. Badania kontynuowano w latach późniejszych. Przykładowo, w pracy [13] sterownik został zaprojektowany w taki sposób, aby system podążał za pożądaną odpowiedzią określoną przez model docelowy w miarę przetwarzania ruchu. Obliczenia zrealizowano dla robota SEIKO TT3000 typu SCARA. Gasparetto i Zanotto w pracy [14] poszukują optymalnej trajektorii dla robota (manipulatora) sześcioosiowego. Autorzy minimalizują funkcję celu złożoną z dwóch składników: pierwszego składnika proporcjonalnego do całkowitego czasu wykonania i drugiego proporcjonalnego do całki kwadratu szarpnięcia definiowanego jako pochodna przyspieszenia wzdłuż trajektorii (zapewnia to uzyskanie wystarczająco gładkiej trajektorii). Proponowana technika umożliwia uwzględnienie ograniczeń kinematycznych na ruch robota, wyrażonych jako górne ograniczenia bezwzględnych wartości prędkości, przyspieszenia i szarpnięcia. Ghasemi, Kashiri i Dartel w pracy [15] omawiają optymalne w czasie planowanie trajektorii manipulatorów robotów w ruchu punkt-punkt metodą pośrednią. Ta metoda wykorzystuje pośrednie rozwiązanie problemu optymalnej regulacji w otwartej pętli, tak, że problem optymalnej regulacji jest tłumaczony na nieliniowy problem dwupunktowej wartości granicznej. Menon, Prakash i Behera [16] przedstawiają optymalizację sterowania ruchu manipulatora. Stopnie swobody robota sterowane są profilem prędkości, który został zrealizowany za pomocą krytycznych sieci adaptacyjnych (*single network adaptive critic* – SNAC). Zastosowane rozwiązanie zapewnia zbieżność do pożądanego optymalnego kosztu, jednocześnie gwarantując stabilność sterowania kinematycznego w pętli zamkniętej. Badania zrealizowano z wykorzystaniem robota UR10. Ferrentino i inni w pracy [17] analizują również kwestie minimalizacji czasu śledzenia trajektorii przy ograniczonej ścieżce i ograniczeniach momentu obrotowego siłownika. Praca ma charakter raczej teoretyczny – przeprowadzono symulacyjne badanie dla robota z dwoma stopniami swobody (2R). W artykule [18] Kim i inni proponują algorytm planowania ruchu dla manipulatorów robotów przy użyciu podwójnego opóźnionego głębokiego deterministycznego gradientu polityki (TD3), który jest algorytmem uczenia się wzmacniającego, dostosowanym do MDP z ciągłym działaniem. Zaproponowany algorytm wygładził ścieżki robotów dla manipulatorów o dwóch i trzech stopniach swobody.

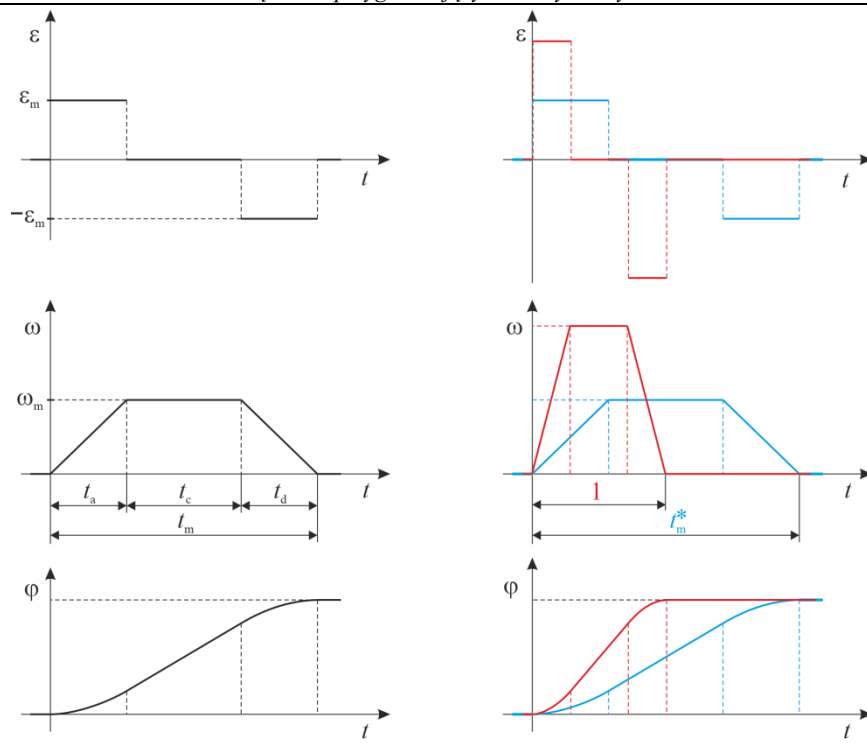
Dla rozważanego w niniejszej pracy urządzenia, jedną z metod minimalizacji czasu pokonania ścieżki złożonej z ciągu ruchów elementarnych jest niezależna optymalizacja poszczególnych ruchów. W ogólnym przypadku, uzyskane tak rozwiązanie nie jest optymalne – byłoby nim na przykład, gdybyśmy żądali, by trajektoria przechodziła precyzyjnie przez wszystkie zadane węzły ścieżki i by ramię zatrzymywało się w każdym węźle. Jednak nawet jeśli warunki te nie są spełnione, uzyskane rozwiązanie

można traktować jako suboptymalne, mogące przynieść istotne skrócenie czasu produkcji leku. Metoda ta zastosowana została w niniejszej pracy do optymalizacji czasu ruchów elementarnych manipulatora UR5.

Ruch przegubów robota podlega szeregowi ograniczeń określonych przez szczytowe dopuszczalne wartości położeń kątowych przegubów, prędkości kątowych, przyspieszeń kątowych oraz momentów siły rozwijanych przez zainstalowany w przegubach serwomechanizmy. Przekroczenie któregokolwiek z limitów jest wykrywane przez sterownik i skutkuje natychmiastowym, samoczynnym zatrzymaniem manipulatora. Realizowany program jest wówczas przerywany i wymagana jest interwencja operatora. Ponowne uruchomienie programu, bez wcześniejszej modyfikacji (poprawienia) kodu, niesie ze sobą ryzyko ponownego przekroczenia ograniczeń. Nie stanowi wiec skutecznego i trwałego rozwiązania problemu. Zatrzymanie robota podczas realizacji zadania produkcji leku może wymusić przerwanie procesu produkcji, doprowadzić do zmarnowania leku i spowodować przestój urządzenia. Jest to więc zdarzenie wysoce niepożądane, którego należy za wszelką cenę uniknąć. Można to osiągnąć przez takie zaplanowanie parametrów poszczególnych ruchów, by nie dopuścić do niebezpiecznego zbliżenia się położeń kątowych, prędkości, przyspieszeń i momentów w poszczególnych przegubach do limitów określonych przez producenta. Limity te, po zacieśnieniu w stopniu wynikającym z przyjętego współczynnika bezpieczeństwa, powinny wiec spełniać rolę ograniczeń w matematycznym sformułowaniu zadania optymalizacji czasu ruchu.

Postać zmiennych decyzyjnych w zadaniu optymalizacji jest narzucona przez składnię wywołania funkcji języka skryptowego realizujących ruch robota. Funkcje te przyjmują jako argumenty między innymi maksymalne przyspieszenie ε_m lub a_m i maksymalną prędkość ω_m lub v_m, które opisują trapezoidalny profil prędkości (rys. 3). Dla funkcji *movej* profil ten określa prędkość kątową przegubu zakreślającego największy kąt (przegubu wiodącego), zaś w przypadku *movel* opisuje on prędkość liniową punktu TCP. W ramach rozwiązania zadania optymalizacji nie są modyfikowane punkty: początkowy ani końcowy, charakteryzujące ruch. Niezmieniony pozostaje też kształt trajektorii w przestrzeni przegubowej czy kartezjańskiej. Od zmiennych decyzyjnych, którymi są parametry trapezoidalnego profilu, uzależniony jest natomiast czas trwania ruchu, który podlega minimalizacji.

W przypadku funkcji *movej*, przez przyjęcie za zmienne decyzyjne wielkości ε_m i ω_m, uzyskuje się bezpośrednią kontrolę nad prędkościami i przyspieszeniami kątowym w przegubach (wiodącym i pozostałych). Łatwo więc zapewnić nieprzekroczenie ograniczeń dla ω i ε. Pozostaje jednak nierozwiązany problem momentu siły, gdyż samo przyspieszenie kątowe w przegubie nie przekłada się bezpośrednio na jego wartość. Sytuacja komplikuje się jeszcze bardziej w przypadku funkcji *movel*, która przyjmuje jako argumenty szczytowe przyspieszenie a_m i szczytową prędkość liniową v_m punktu TCP. Tutaj nawet stosunkowo niewielkie prędkości i przyspieszenia liniowe centralnego punktu narzędzia mogą skutkować znacznymi prędkościami i przyspieszeniami kątowymi przegubów lub wielkimi wartościami działających w nich momentów siły. Efekt ten jest najlepiej widoczny, gdy robot znajduje się w pobliżu konfiguracji osobliwej. Dzieje się tak na przykład wówczas, gdy osie przegubów nadgarstek 1 i nadgarstek 3 stają się równoległe lub gdy ramię sięga do punktu na granicy obszaru roboczego (zasięgu).



Rysunek 3. Lewa kolumna: profile przyspieszenia (schodkowy), prędkości (trapezoidalny) i położenia kątowego przegubu (S-kształtny). Prawa kolumna: kolor czerwony odpowiada jednostkowemu czasowi ruchu, zaś kolor niebieski – otrzymanemu czasowi suboptymalnemu

Dobranie parametrów i przyspieszeń szczytowych, przekazanych jako argumenty funkcjom *movel* lub *movej*, gwarantujących nieprzekroczenie ograniczeń, wymaga więc rozwiązania prostego zadania dynamiki, a w przypadku funkcji *movel* – także odwrotnego zadania kinematyki dla manipulatora. Rozwiązań tych zadań można poszukiwać bądź analitycznie, bądź metodami numerycznymi i symulacjami komputerowymi. Znajomość uniwersalnych, analitycznych, zwartych formuł umożliwiłaby łatwe i szybkie testowanie dowolnych ruchów manipulatora. Jednak znalezienie wzorów analitycznych dla ramienia o sześciu stopniach swobody jest przedsięwzięciem niebanalnym. Z kolei wadą podejścia numerycznego jest konieczność powtarzania czasochłonnej procedury numerycznej dla każdego ruchu z osobna.

W rozważanym projekcie wystarcza, by obliczenia przeprowadzić dla skończonej liczby z góry znanych ruchów elementarnych. Co więcej, czas obliczeń nie jest specjalnie istotny, bo są one przeprowadzane jednorazowo, w trybie *off-line*. Dlatego zdecydowano wykorzystać metody numeryczne i symulacje komputerowe.

Dla uproszczenia zadania optymalizacji zdecydowano o zmniejszeniu liczby zmiennych decyzyjnych, związanych z pojedynczym ruchem elementarnym, z dwóch do jednej. Założono w tym celu arbitralny stosunek r czasów przyspieszania t_a i hamowania t_d, charakteryzujących trapezoidalny profil prędkości pokazany na rysunku 12., do całkowitego czasu ruchu elementarnego t_m.

$$\frac{t_a}{t_m} = \frac{t_d}{t_m} = r = \text{const} \rightarrow t_a = t_d = r\, t_m \qquad (1)$$

Czas poruszania się ze stałą prędkością t_c jest określony zależnością:

$$t_m = t_a + t_c + t_d \rightarrow t_c = t_m - t_a - t_d = (1 - 2\,r)\,t_m \qquad (2)$$

Znaczenie użytych indeksów jest następujące: a – *acceleration*, d – *deceleration*, c – *cruise*, m – *move*. Przyjęcie stałego stosunku r pozwala na operowanie tylko jedną, zastępczą zmienną decyzyjną – całkowitym czasem ruchu – zamiast dwiema oryginalnymi – szczytowymi wartościami prędkości i przyspieszenia. W rozważanym zadaniu taka zmienna decyzyjna jest zarazem wskaźnikiem jakości sterowania. Przyjęte uproszczenie ułatwia znalezienie rozwiązania, należy jednak pamiętać, że zawęża też przestrzeń poszukiwań, więc w ogólnym przypadku otrzymane rozwiązanie będzie jedynie suboptymalne. Relacje między czasem t_m, dystansem kątowym ruchu w przegubie Δq oraz szczytowymi wartościami przyspieszenia ε_m i prędkości ω_m są następujące:

$$|\Delta q| = \left(\frac{1}{2}t_a + t_c + \frac{1}{2}t_d\right)\omega_m = (1 - r)\,t_m\,\omega_m$$

$$\omega_m = \frac{|\Delta q|}{(1-r)\,t_m} \qquad (3)$$

$$\omega_m = \varepsilon_m\,t_a = \varepsilon_m\,r\,t_m \rightarrow \varepsilon_m = \frac{\omega_m}{r\,t_m} \qquad (4)$$

Wstawiając (3) do (4) otrzymujemy:

$$\varepsilon_m = \frac{|\Delta q|}{(1-r)\,r\,t_m^2} \qquad (5)$$

Wartość Δq dla każdego z przegubów jest wyznaczana w oparciu o znajomość wartości współrzędnych przegubowych dla początkowego i końcowego punktu ruchu elementarnego. Te zaś wyznaczane są doświadczalnie, w procesie uczenia robota, podczas planowania ścieżki jego ruchu.

Momenty siły, które muszą być przyłożone w przegubach, by poprowadzić ramię po zadanej trajektorii, można rozłożyć na kilka składowych: bezwładnościową, Coriolisa, odśrodkową i grawitacyjną (składową tarciową zaniedbano). Składowe bezwładnościowe są proporcjonalne do przyspieszeń. Stąd oraz ze wzoru (5) wynika, że są odwrotnie proporcjonalne do kwadratu całkowitego czasu ruchu elementarnego t_m. Składowe: Coriolisa i odśrodkowa są proporcjonalne do kwadratów lub dwuelementowych mieszanych iloczynów prędkości. Stąd oraz z (3) wynika, że one również są odwrotnie proporcjonalne do t_m^2. Dlatego te trzy składowe będziemy dalej traktować łącznie i oznaczać symbolem M_{icc} (icc – inertial, Coriolis, centrifugal). Załóżmy, że znany jest profil momentu M_{icc} odpowiadający jednostkowemu czasowi ruchu ($t_m = 1$) i niezmienionemu stosunkowi r. Oznaczmy go przez M_{ICC}. Wówczas dla dowolnego czasu t_m prawdziwa jest zależność:

$$M_{icc}(t) = \frac{1}{t_m^2} M_{ICC}\left(\frac{t}{t_m}\right), t \in [0, t_m\,] \qquad (6)$$

Ostatnia składowa momentu siły – grawitacyjna – nie zależy ani od prędkości, ani od przyspieszeń. Oznaczmy ją przez M_g. Relacja między profilem M_g dla dowolnego

czasu t_m a profilem dla czasu jednostkowego ($t_\mathrm{m} = 1$), który oznaczymy przez M_G, jest następująca:

$$M_\mathrm{g}(t) = M_\mathrm{G}\left(\frac{t}{t_\mathrm{m}}\right), t \in [0, t_m\ ] \tag{7}$$

Wypadkowy moment dany jest wzorem:

$$M(t) = M_\mathrm{icc}(t) + M_\mathrm{g}(t) \tag{8}$$

Na każdym etapie ruchu moment M rozwijany przez dany serwomechanizm powinien się mieścić w przedziale wyznaczonym przez maksymalną dopuszczalną wartość określoną przez producenta, pomniejszoną o założony margines bezpieczeństwa, na przykład 20%. Oznaczmy tę pomniejszoną wartość przez M_a:

$$-M_\mathrm{a} \le M(t) \le M_\mathrm{a}, t \in [0, t_m\ ]$$
$$-M_\mathrm{a} \le M_\mathrm{icc}(t) + M_\mathrm{g}(t) \le M_\mathrm{a} \tag{9}$$

Po podstawieniu (6) i (7) do (9) dostajemy:

$$-M_\mathrm{a} \le \frac{M_\mathrm{ICC}(t/t_\mathrm{m})}{t_\mathrm{m}^2} + M_\mathrm{G}(t/t_\mathrm{m}) \le M_\mathrm{a} \tag{10}$$

Zastosowanie w nierówności (10) podstawienia $\tau = t/t_\mathrm{m}$ i jej nieznaczne przekształcenie daje:

$$-M_\mathrm{a} - M_\mathrm{G}(\tau) \le \frac{1}{t_\mathrm{m}^2} M_\mathrm{ICC}(\tau) \le M_\mathrm{a} - M_\mathrm{G}(\tau), \tau \in [0, 1] \tag{11}$$

Po pomnożeniu stron nierówności przez kwadrat czasu ruchu elementarnego, który ma zawsze wartość dodatnią, otrzymujemy:

$$\left(-M_\mathrm{a} - M_\mathrm{G}(\tau)\right)t_\mathrm{m}^2 \le M_\mathrm{ICC}(\tau) \le \left(M_\mathrm{a} - M_\mathrm{G}(\tau)\right)t_\mathrm{m}^2 \tag{12}$$

Załóżmy, że składowa M_G pozostaje zawsze we wnętrzu przedziału o końcach wyznaczonych przez graniczną wartość dopuszczalną M_a:

$$-M_\mathrm{a} < \mathrm{M_G}(\tau) < M_\mathrm{a}, \tau \in [0, 1]$$

Stąd:

$$-M_\mathrm{a} - M_\mathrm{G}(\tau) < 0, M_\mathrm{a} - M_\mathrm{G}(\tau) > 0 \tag{13}$$

Rozbijmy nierówność podwójną (12) na dwie pojedyncze. Następnie, uwzględniając relacje (13), przekształćmy je do następującej postaci:

$$t_\mathrm{m}^2 \ge \frac{M_\mathrm{ICC}(\tau)}{-M_\mathrm{a} - M_\mathrm{G}(\tau)}, t_\mathrm{m}^2 \ge \frac{M_\mathrm{ICC}(\tau)}{M_\mathrm{a} - M_\mathrm{G}(\tau)}, \tau \in [0, 1] \tag{14}$$

Nierówności (14) można zastąpić jedną, wspólną zależnością:

$$t_\mathrm{m}^2 \ge \max_{\tau \in [0,1]} \max\left\{\frac{M_\mathrm{ICC}(\tau)}{-M_\mathrm{a} - M_\mathrm{G}(\tau)}, \frac{M_\mathrm{ICC}(\tau)}{M_\mathrm{a} - M_{\mathrm{G}(\tau)}}\right\} \tag{15}$$

Relacja (15) stanowi jedno z ograniczeń zadania optymalizacji. Pozostałe ograniczenia wynikają z (3) i (5) i mają postać:

$$t_{\mathrm{m}} \geq \frac{|\Delta q|}{(1-r)\,\omega_{\mathrm{a}}} \tag{16}$$

$$t_{\mathrm{m}}^{2} \geq \frac{|\Delta q|}{(1-r)\,r\,\varepsilon_{\mathrm{a}}} \tag{17}$$

Tutaj ω_{a} oraz ε_{a} to szczytowe dopuszczalne wartości odpowiednio przegubowej prędkości kątowej i przegubowego przyspieszenia kątowego, podane przez producenta robota, pomniejszone o 20%. Arbitralnie przyjęty dwudziestoprocentowy margines bezpieczeństwa ma zapewnić niezawodne działanie i jednocześnie uodpornić wynik na nieuniknione uproszczenia, które przyjęto podczas modelowania układu.

Optymalizacja polega na poszukiwaniu minimalnego czasu t_{m}, który spełnia ograniczenia (15)-(17). Ostateczny wynik dany jest wzorem:

$$t_{\mathrm{m}}^{*} = \max_{i\in\{1,2,3,4,5,6\}} \max\left\{\sqrt{\max_{\tau\in[0,1]} \max\left\{\frac{M_{\mathrm{ICC}_i}(\tau)}{-M_{\mathrm{a}_i}-M_{\mathrm{G}_i}(\tau)}, \frac{M_{\mathrm{ICC}_i}(\tau)}{M_{\mathrm{a}_i}-M_{\mathrm{G}_i(\tau)}}\right\}}, \sqrt{\frac{|\Delta q_i|}{(1-r)r\varepsilon_{\mathrm{a}_i}}}, \frac{|\Delta q_i|}{(1-r)\omega_{\mathrm{a}_i}}\right\} \tag{18}$$

Indeks *i* to numer przegubu ramienia robota. Poszukiwany jest czas, który zapewni jednoczesne spełnienie ograniczeń dla wszystkich przegubów. Oczekujemy, że dla rozwiązania optymalnego, w przedziale $[0, t_{\mathrm{m}}]$ znajdzie się przynajmniej jedna chwila czasu, w której aktywne będzie przynajmniej jedno z opisanych wyżej ograniczeń dla co najmniej jednego przegubu.

Numeryczne rozwiązanie zadania optymalizacji przebiega według następującego scenariusza. Doświadczalnie, na stanowisku robotycznym, określane są konfiguracje i pozy odpowiadające początkowym i końcowym punktom wszystkich segmentów zaprogramowanej ścieżki. Konfiguracje, opisane zestawem współrzędnych przegubowych, stanowią naturalny opis w przypadku ruchu typu *movej*, zaś pozy, określone pozycją i orientacją końcówki złącza Luer lock, są typowe dla ruchu typu *movel*. Dla każdego ruchu elementarnego, w oparciu o znajomość współrzędnych przegubowych punku początkowego i końcowego, przeprowadzane są dwie symulacje komputerowe. Wykorzystywane są do tego celu modele zbudowane w środowisku MATLAB/Simulink/ Simscape/SimMechanics [19], opisane w następnym rozdziale. Pierwsza symulacja, przeprowadzana na bardzo długim horyzoncie czasowym i ze stałymi wartościami prędkości przegubowych, ma za zadanie wyznaczenie przebiegów momentów siły rozwijanych przez silniki serwomechanizmów w celu pokonania sił ciężkości działających na ramię. Po odpowiedniej interpolacji, daje ona profile M_{G}. Druga symulacja, przeprowadzana na przedziale czasu o jednostkowej długości dla trapezoidalnych profili prędkości (czerwony wykres na rysunku 3), daje informacje o sumarycznych przebiegach momentów, obejmujących cztery składniki: grawitacyjny, bezwładnościowy, Coriolisa i odśrodkowy. Wspomniana wcześniej interpolacja wyników pierwszej symulacji jest niezbędna ze względu na niezgodność czasu ruchu i kształtu profilów prędkości dla obu symulacji. W interpolacji tej za zmienne niezależne przyjmowane są kątowe współrzędne przegubowe. Odejmując wyniki symulacji – pierwszej od drugiej – uzyskujemy profile M_{ICC} momentów siły obejmujące składowe: bezwładnościowe, Coriolisa i odśrodkowe. Uzyskane wyniki, podstawione do wzoru (18), pozwalają

obliczyć suboptymalny czas trwania ruchu elementarnego. Na jego podstawie, stosując wzory (3) i (5), wyznaczamy argumenty ω_a i ε_a, przekazywane do funkcji *movej*. Wynikowe profile przyspieszenia, prędkości i położenia zaznaczone są na rysunku 3. niebieską linią. Przeprowadzenie tej procedury dla pozostałych ruchów elementarnych daje rozwiązanie suboptymalne dla całej rozważanej ścieżki. Warto zauważyć, że w opisanej metodzie nie jest konieczne stosowanie metod iteracyjnych. Zapewnia to stosunkowo szybkie uzyskanie wyniku w przewidywalnym czasie (uzależnionym głównie od czasu trwania symulacji numerycznych ruchu ramienia).

W przypadku segmentu ścieżki realizowanego funkcją *movel*, powyższa procedura wymaga pewnej modyfikacji. Jak poprzednio, ograniczeniom podlegają wielkości przegubowe: prędkości, przyspieszenia, momenty siły. Jednak funkcja *movel* oczekuje podania parametrów opisujących ruch liniowy końcówki roboczej: maksymalnej prędkość v_m i maksymalnego przyspieszenia a_m. Dlatego, zamiast wzorów (3) i (5), należy zastosować odmienne formuły:

$$v_\mathrm{m} = \frac{|\Delta s|}{(1-r)\,t_\mathrm{m}}, a_\mathrm{m} = \frac{|\Delta s|}{(1-r)\,r\,t_\mathrm{m}^2} \tag{19}$$

Wielkość $|\Delta s|$ oznacza tu dystans w przestrzeni kartezjańskiej, jaki pokonuje końcówka robocza. Profile kątowych przyspieszeń i prędkości przegubowych nie są już trapezoidalne. Dlatego muszą być wyznaczane symulacyjnie (podobnie jak przebiegi momentów siły) i przetwarzane podobnie jak rozważana wyżej wielkość M_ICC. W konsekwencji, zmianie ulegnie postać wzoru (18) – drugi i trzeci element w nawiasach klamrowych przybiorą postać podobną do pierwszego.

W niniejszej pracy, procedura optymalizacji numerycznej została zaimplementowana w języku MATLAB w postaci M-pliku skryptowego, zaś do przeprowadzenia symulacji ruchu ramienia wykorzystano przybornik SimMechanics.

5. Numeryczne rozwiązanie zadań kinematyki i dynamiki

Istotnym elementem opisanej procedury optymalizacji są symulacje komputerowe ruchu ramienia robota. Ich zadaniem jest rozwiązanie prostego zadania dynamiki, a w przypadku ruchu typu *movel* – także odwrotnego zadania kinematyki [20, 21]. Do symulacji ruchu manipulatora UR5 wykorzystano środowisko obliczeń numerycznych MATLAB, wraz z przybornikami: Simulink, Simscape oraz SimMechanics (*Simscape Multibody*). W zestawie tym MATLAB zapewnia interaktywną linię poleceń, interpreter języka skryptowego, silnik obliczeń numerycznych oraz narzędzia do wizualizacji i prezentacji wyników obliczeń. Simulink jest środowiskiem do graficznego modelowania systemów dynamicznych i numerycznej symulacji ich zachowania. W swej podstawowej formie umożliwia modelowanie zwane *matematycznym*, gdyż budowane w nim schematy blokowe stanowią prosty odpowiednik formuł matematycznych i równań różniczkowych zwyczajnych. Przybornik Simscape rozszerza możliwości Simulinka o modelowanie zwane *fizycznym* bądź *fizykalnym*, gdzie elementy graficzne, składające się na budowany model, reprezentują rzeczywiste obiekty fizyczne, jak elementy RLC obwodów elektrycznych czy różnorakie podzespoły mechanizmów. Wreszcie SimMechanics jest przybornikiem przeznaczonym do modelowania fizycznego łańcuchów kinematycznych, realizujących ruch w przestrzeni trójwymiarowej. Przybornik służy do rozwiązywania zagadnienia układów mechanicznych złożonych z wielu oddziałujących ze sobą ciał (multibody dynamics). Szczególnie dobrze nadaje

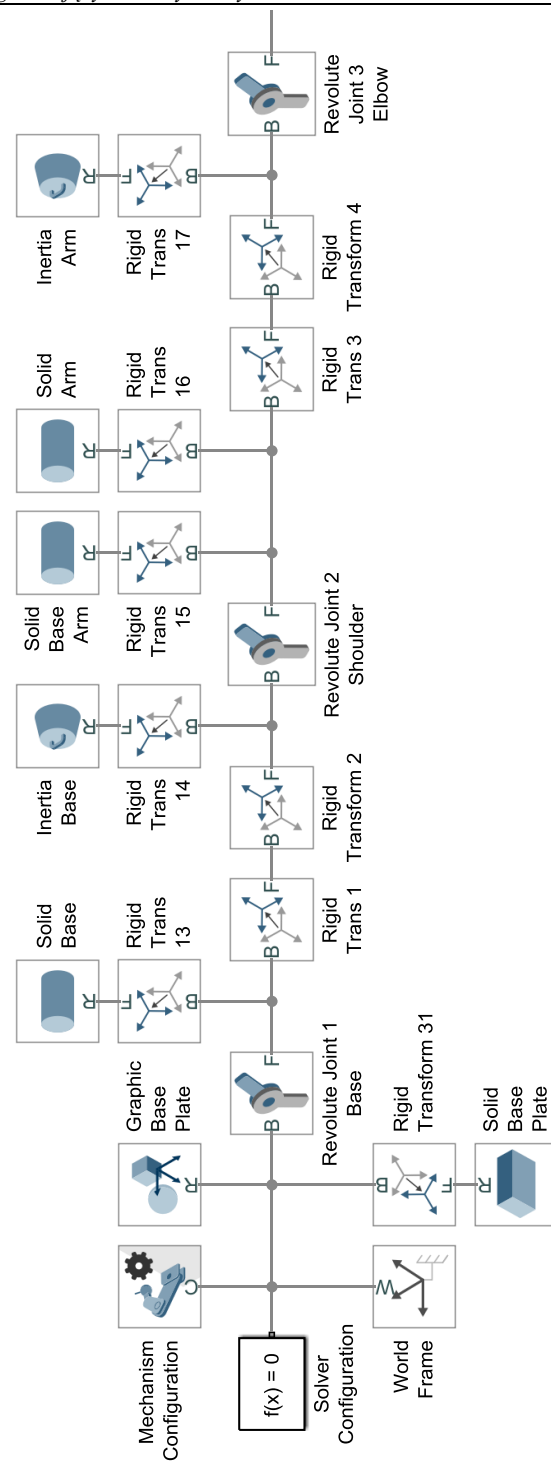
się do modelowania robotów przemysłowych czy manipulatorów. Może być użyty do rozwiązywania zadań kinematyki i dynamiki, zarówno prostych, jak i odwrotnych.

Model manipulatora UR5, zbudowany w środowisku SimMechanics (Simscape Multibody), pokazany jest na rysunku 4. Ma on postać schematu blokowego utworzonego w oknie edytora graficznego Simulink. Trójwymiarowa wizualizacja ramienia, uzyskana w oknie narzędzia Mechanics Explorers, pokazana jest na rysunku 5.

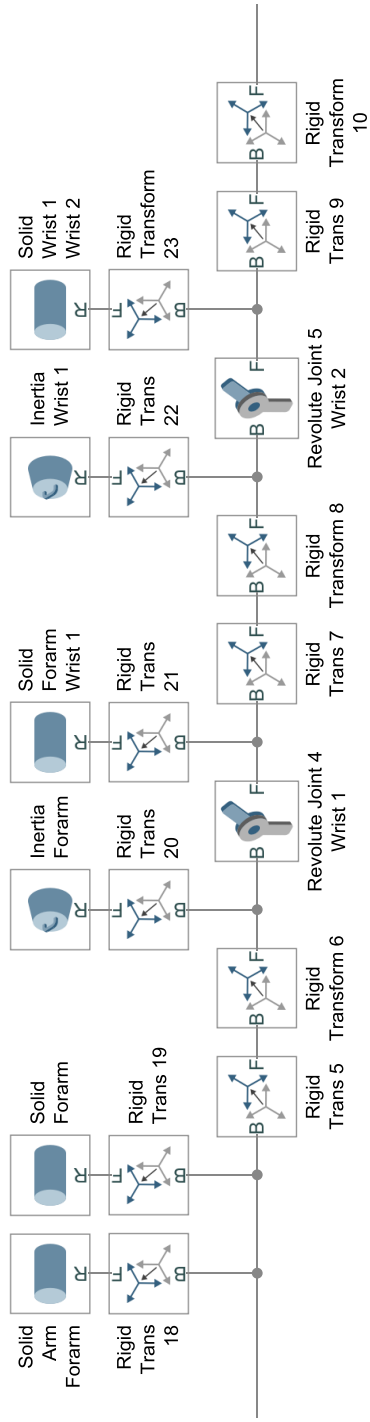
Dla budowy modelu i prowadzenia symulacji komputerowych niezbędna jest znajomość szeregu parametrów opisujących ramię robota, takich jak: wymiary geometryczne poszczególnych członów i przegubów, masy członów, położenia ich środków ciężkości, tensory momentu bezwładności wszystkich członów. Dla rozwiązania postawionego wcześniej zadania optymalizacji wymagana jest dodatkowa wiedza o wartościach ograniczeń narzuconych przez producenta i konstrukcję układu na położenia, prędkości i przyspieszenia kątowe poszczególnych przegubów oraz momenty siły rozwijane przez serwonapędy. Wymienione dane pozyskano z różnych źródeł, przedstawionych poniżej.

Wymiary ramienia odczytano z dokumentacji technicznej mechanicznej umieszczonej w witrynie producenta. Pozyskane z nich dane są zobrazowane na rysunku 5. Masy i lokalizacje środków ciężkości członów ramienia, a także parametry Denavita-Hartenberga, podane są w witrynie producenta, w dziale *FAQ* [22] oraz w podręczniku *Calibration Manual* [23]. Wymieniono je w tabeli 1. Położenia kątowe wszystkich przegubów są ograniczone do przedziału [−360°, +360°], zaś maksymalne prędkości ruchu przegubów – do 180°/s. Dane te można znaleźć w nocie katalogowej [24]. Maksymalne dopuszczalne momenty siły, jakie mogą być rozwijane przez zainstalowane w przegubach serwonapędy, zebrane są w tabeli 2. Pochodzą one z notatki *Max. joint torques – 17260* [25], dostępnej w witrynie producenta, na podstronie *Support* w dziale *FAQ*.

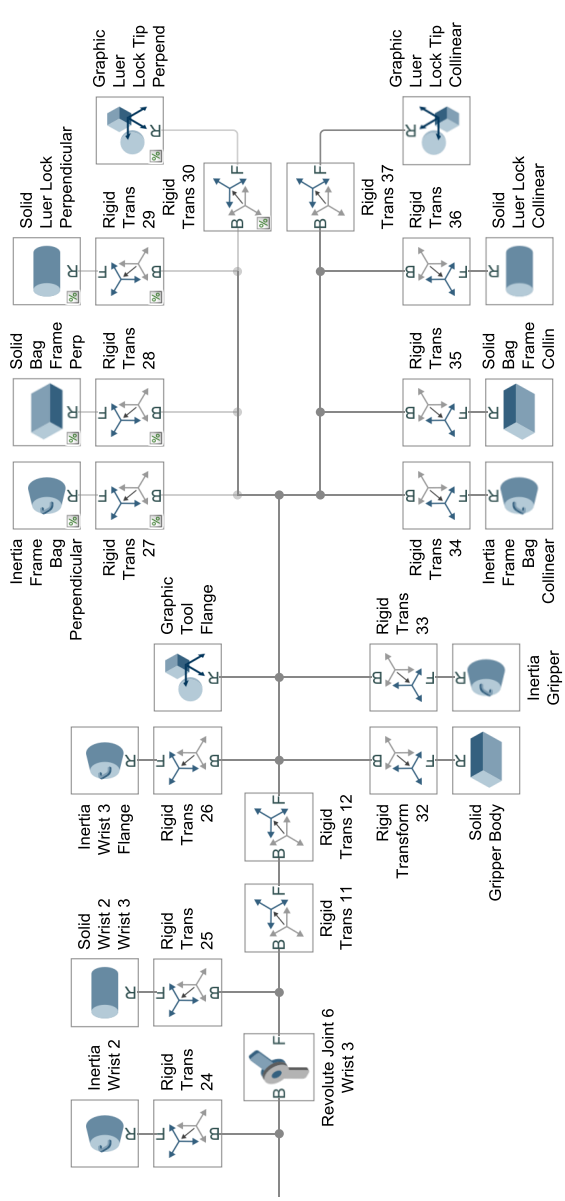
Większość danych zebranych w tabelach 1 i 2 można również odczytać z pliku konfiguracyjnego `urcontrol.conf` aplikacji URControl, który znajduje się w katalogu `/root/.urcontrol` na nośniku pamięci masowej komputer PC wbudowanego w sterownik robota. Wartości maksymalnych przyspieszeń kątowych dla poszczególnych przegubów znaleziono w plikach `joint_size3_rev13.conf` oraz `joint_size1_rev13.conf`. Niestety ani w internetowej witrynie producenta, ani w plikach konfiguracyjnych nie udało się odnaleźć danych o tensorach bezwładności członów ramienia (w pliku `urcontrol.conf` wszystkie elementy macierzy inercji są zdefiniowane jako zerowe). Zaczerpnięto je wobec tego z pracy [26], gdzie zostały wyznaczone w sposób przybliżony, przy założeniu, że masa każdego segmentu ramienia skupia się głównie na jego końcach, gdzie zainstalowano serwomechanizmy. Są one podane w tabeli 3.



Rysunek 4a. Model manipulatora UR5 zbudowany w środowisku SimMechanics (Simscape Multibody) – ciąg dalszy rysunku na następnej stronie



Rysunek 4b. Model manipulatora UR5 zbudowany w środowisku SimMechanics (Simscape Multibody) – kontynuacja rysunku z poprzedniej strony



Rysunek 4c. Model manipulatora UR5 zbudowany w środowisku SimMechanics (Simscape Multibody) – kontynuacja rysunku z poprzedniej strony

Tabela 1. Parametry Denavita-Hartenberga oraz masy i środki ciężkości segmentów ramienia

Przegub	Nazwa przegubu	Parametry Denavita--Hartenberga	Masa ramienia m_i [kg]	Współrzędne środka ciężkości (CoG) [m]
1 A	podstawa base	$a = 0{,}00000$ $d = 0{,}089159$ $\alpha = 1{,}570796327$ $s = -1$	3,7000	[0, −0,02561, 0,00193]

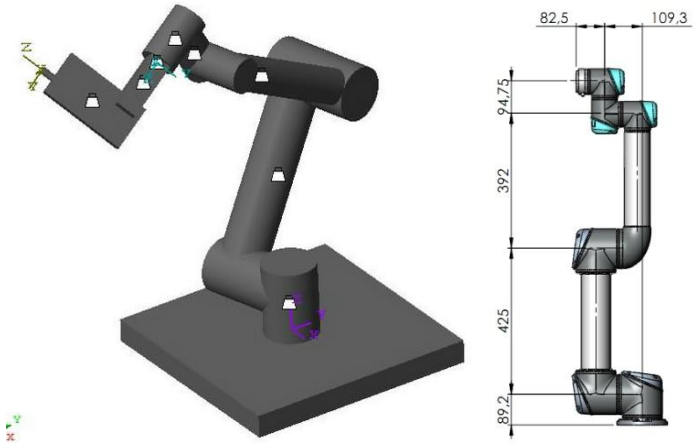
2 B	bark shoulder	$a = -0{,}42500$ $d = 0{,}00000$ $\alpha = 0$ $s = -1$	8,3930	[0,2125, 0, 0,11336]
3 C	łokieć elbow	$a = -0{,}39225$ $d = 0{,}00000$ $\alpha = 0$ $s = 1$	2,2750	[0,11993, 0, 0,0265]
4 D	nadgarstek1 wrist1	$a = 0{,}00000$ $d = 0{,}10915$ $\alpha = 1{,}570796327$ $s = 1$	1,2190	[0, −0,0018, 0,01634]
5 E	nadgarstek2 wrist2	$a = 0{,}00000$ $d = 0{,}09465$ $\alpha = -1{,}5707963$ $s = 1$	1,2190	[0, 0,0018, 0,01634]
6 F	nadgarstek3 wrist3	$a = 0{,}0000$ $d = 0{,}0823$ $\alpha = 0$ $s = 1$	0,1879	[0, 0, −0,001159]

Tabela 2. Ograniczenia kinematyczne i dynamiczne dla przegubów manipulatora

Numer przegubu	Nazwa przegubu	Maksymalna prędkość kątowa $dq/dt = \omega$ [rad/s] [°/s]	Maksymalne przyspieszenie kątowe $d^2q/dt^2 = \varepsilon$ [rad/s^2]	Maksymalny moment napędowy $M = T$ [N·m]
1 A	podstawa base	π rad/s 180°/s	15	150
2 B	bark shoulder			
3 C	łokieć elbow			
4 D	nadgarstek1 wrist1		25	28
5 E	nadgarstek2 wrist2			
6 F	nadgarstek3 wrist3			

Tabela 3. Tensory momentów bezwładności członów ramienia manipulatora

Przegub	Nazwa przegubu	Elementy tensora momentu bezwładności [kg·m^2]
1 A	podstawa base	$I_{xx} = 0{,}0067$, $I_{yy} = 0{,}0064$, $I_{zz} = 0{,}0067$, $I_{xy} = I_{xz} = I_{yz} = 0$
2 B	bark shoulder	$I_{xx} = 0{,}0149$, $I_{yy} = 0{,}3564$, $I_{zz} = 0{,}3553$, $I_{xy} = I_{xz} = I_{yz} = 0$
3 C	łokieć elbow	$I_{xx} = 0{,}0025$, $I_{yy} = 0{,}0551$, $I_{zz} = 0{,}0546$, $I_{xz} = 0{,}0034$, $I_{xy} = I_{yz} = 0$
4 D	nadgarstek1 wrist1	$I_{xx} = 0{,}0012$, $I_{yy} = 0{,}0012$, $I_{zz} = 0{,}0009$, $I_{xy} = I_{xz} = I_{yz} = 0$
5 E	nadgarstek2 wrist2	$I_{xx} = 0{,}0012$, $I_{yy} = 0{,}0012$, $I_{zz} = 0{,}0009$, $I_{xy} = I_{xz} = I_{yz} = 0$
6 F	nadgarstek3 wrist3	$I_{xx} = 0{,}0001$, $I_{yy} = 0{,}0001$, $I_{zz} = 0{,}0001$, $I_{xy} = I_{xz} = I_{yz} = 0$

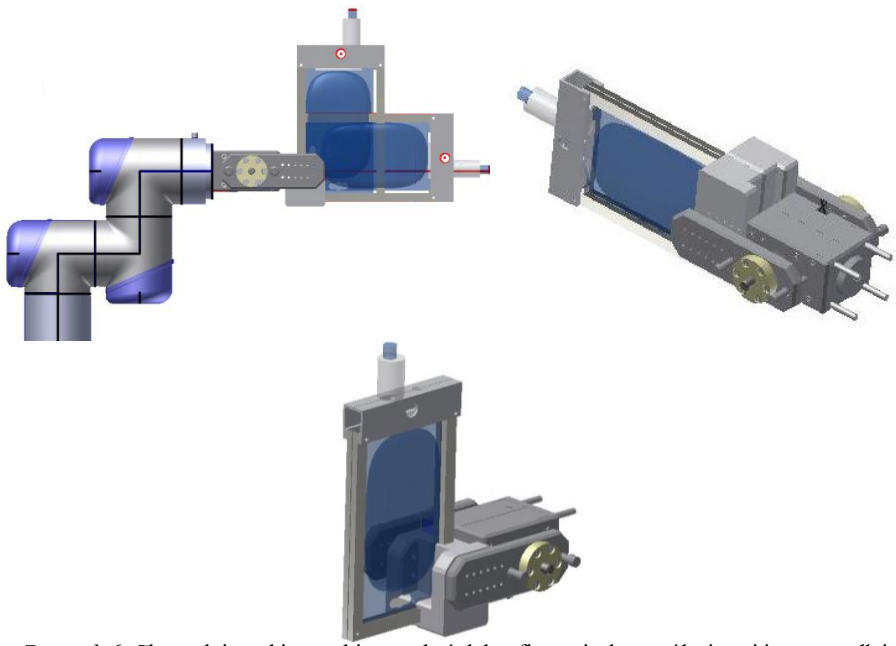


Rysunek 5. Wizualizacja modelu ramienia w oknie Mechanics Explorers. Wymiary ramienia manipulatora. Na podstawie dokumentacji robota UR5

Na końcu ramienia, na kołnierzu kiści, umocowany jest chwytak, którego zadaniem jest ujmowanie ramki z workiem do transfuzji, zakończonym złączem Luer lock. Kształtowe nakładki szczęk chwytaka uformowano tak, by możliwe było dzierżenie ramki na dwa różne sposoby, pokazane na rysunku 6. W pierwszym przypadku osie symetrii kołnierza kiści robota i złącza Luer lock pokrywają się, zaś w drugim – są prostopadłe. W tabeli 4. podano masy, położenia środków ciężkości, tensory momentów bezwładności samego chwytaka oraz zestawu złożonego z chwytaka oraz ramki z workiem i złączem Luer lock, dla dwóch przypadków uchwycenia ramki (współosiowego i prostopadłego). Umieszczono tu też parametry samej ramki, w obu ułożeniach. Parametry te obliczono, korzystając ze stosownych funkcji programu Autodesk Inventor, na podstawie zbudowanego w tym programie modelu mechanicznego 3D. Położenie środka ciężkości wyrażone jest w kartezjańskim układzie współrzędnych związanym z kołnierzem kiści robota [5], zaś tensor momentu bezwładności – w układzie centralnym powstałym przez przesunięcie poprzedniego do środka ciężkości. W obliczeniach wykorzystano dane o gęstość stali nierdzewnej (8000 kg/m^3), z której wykonany jest korpus i szczęki chwytaka, stali stanowiącej materiał śruby (7850 kg/m^3) oraz aluminium (2700 kg/m^3), użytego do budowy ramki. W przypadku silnika krokowego, założono jednorodny rozkład masy, a jej wartość (0,34 kg) odczytano z dokumentacji udostępnianej przez producenta. W obliczeniach pominięto drobne elementy, na przykład nakrętki. W tabeli 4. podano też współrzędne końcówki złącza Luer lock w układzie współrzędnych związanym z kołnierzem kiści robota.

Zasadnicza część symulacyjnego modelu ramienia chwytaka, pokazana na rysunku 4., odwzorowuje łańcuch kinematyczny ramienia o sześciu stopniach swobody. Sześć przegubów walcowych reprezentowanych jest przez bloki *Revolute Joint*. Pary bloków *Rigid Transform*, włączone pomiędzy kolejne przeguby, reprezentują względne przesunięcie i obrót układów współrzędnych z nimi związanych. W oknach dialogowych tych bloków wprowadzono parametry Denavita-Hartenberga opisujące kolejne człony manipulatora. Bloki *Inertia*, dołączone do łańcucha kinematycznego, opisują własności inercyjne poszczególnych członów: masy oraz tensory momentów bezwładności. Bloki *Rigid Transform*, pośredniczące między blokami *Inertia* a głównym torem łańcucha,

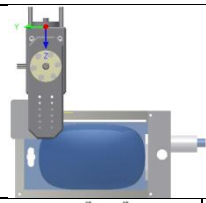
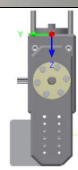
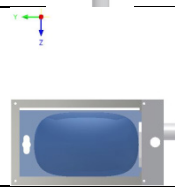
reprezentują przesunięcia środków mas członów w stosunku do początków układów współrzędnych związanych z przegubami. Bloki *Solid* są wykorzystywane wyłącznie dla potrzeb wizualizacji w oknie *Mechanics Explorers*. Przy ich pomocy zdefiniowano walce i prostopadłościany, które odwzorowują w przybliżeniu wygląd faktycznego manipulatora wraz z chwytakiem dzierżącym ramkę z workiem (patrz rys. 5). Bloki *Rigid Transform*, włączone przed blokami *Solid*, są niezbędne, gdyż geometryczne środki figur reprezentujących poszczególne elementy mechaniczne nie pokrywają się z początkami układów współrzędnych związanych z przegubami.



Rysunek 6. Chwytak i ramki z workiem w dwóch konfiguracjach: współosiowej i prostopadłej

Tabela 4. Parametry bezwładnościowe chwytaka i ramki z workiem – razem i osobno

Konfiguracja	Rysunek	Masa [kg]	Położenie środka ciężkości (masy) oraz końcówki złącza Luer lock [mm]	Tensor momentu bezwładności [kg·mm^2]
Zestaw chwytaka z workiem w ułożeniu współosiowym		$m = 2{,}318$	$x_{CoG} = -0{,}039$ $y_{CoG} = -1{,}242$ $z_{CoG} = 66{,}865$ $x_{TCP} = 0$ $y_{TCP} = 0$ $z_{TCP} = 314$	$I_{xx} = 8911{,}7$ $I_{yy} = 9564{,}8$ $I_{zz} = 2001{,}4$ $I_{xy} = 2{,}942$ $I_{xz} = -16{,}286$ $I_{yz} = 469{,}99$

Zestaw chwytaka z workiem w ułożeniu prostopadłym		$m = 2{,}318$	$x_{CoG} = -0{,}040$ $y_{CoG} = -6{,}927$ $z_{CoG} = 61{,}181$ $x_{TCP} = 0$ $y_{TCP} = -183$ $z_{TCP} = 161$	$I_{xx} = 7038{,}8$ $I_{yy} = 6151{,}2$ $I_{zz} = 3542{,}1$ $I_{xy} = 9{,}985$ $I_{xz} = -9{,}053$ $I_{yz} = 1406{,}4$
Sam chwytak		$m = 2{,}051$	$x_{CoG} = -0{,}110$ $y_{CoG} = 0{,}620$ $z_{CoG} = 50{,}043$	$I_{xx} = 2708{,}7$ $I_{yy} = 3680{,}3$ $I_{zz} = 1672{,}1$ $I_{xy} = 0{,}612$ $I_{xz} = 4{,}850$ $I_{yz} = -72{,}030$
Sama ramka z workiem w ułożeniu współosiowym		$m = 0{,}268$	$x_{CoG} = 0{,}501$ $y_{CoG} = -15{,}509$ $z_{CoG} = 195{,}744$ $x_{TCP} = 0$ $y_{TCP} = 0$ $z_{TCP} = 314$	$I_{xx} = 1115{,}5$ $I_{yy} = 858{,}44$ $I_{zz} = 267{,}64$ $I_{xy} = -0{,}006$ $I_{xz} = -0{,}044$ $I_{yz} = -14{,}339$
Sama ramka z workiem w ułożeniu prostopadłym		$m = 0{,}268$	$x_{CoG} = 0{,}499$ $y_{CoG} = -64{,}744$ $z_{CoG} = 146{,}509$ $x_{TCP} = 0$ $y_{TCP} = -183$ $z_{TCP} = 161$	$I_{xx} = 1115{,}5$ $I_{yy} = 267{,}64$ $I_{zz} = 858{,}44$ $I_{xy} = -0{,}044$ $I_{xz} = -0{,}006$ $I_{yz} = -14{,}339$

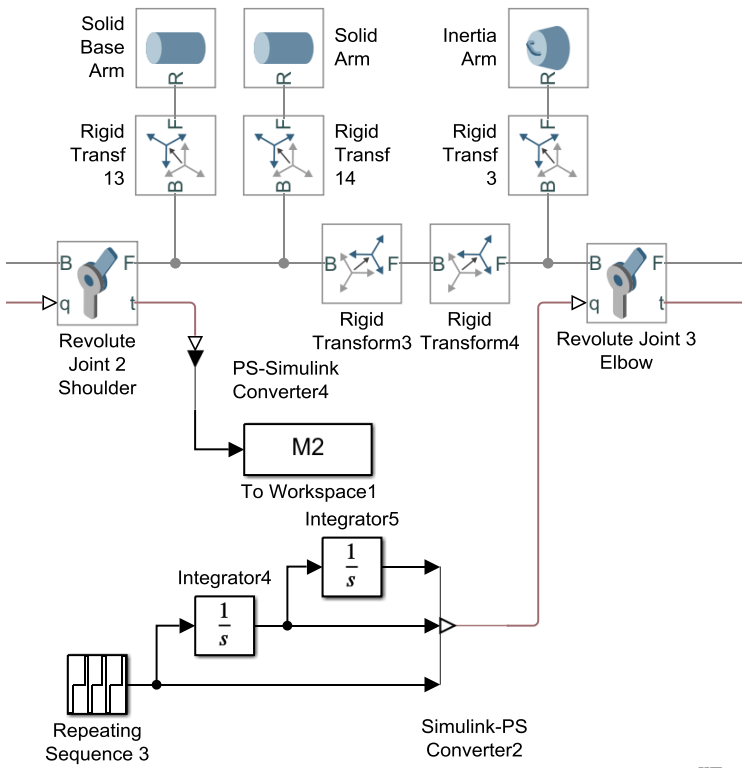
W modelu utworzonym w środowisku SimMechanics nie uwzględniono momentów bezwładności wirników silników wchodzących w skład serwomechanizmów zainstalowanych w przegubach ramienia. Tymczasem, w przypadku dużego przyspieszenia kątowego w przegubie, ze względu na wysokie przełożenie przekładni redukcyjnej serwomechanizmu, bezwładność ta stanowi istotny czynnik wpływający na wartość całkowitego momentu siły, który musi rozwijać silnik. A to właśnie dla całkowitego momentu obowiązują graniczne dopuszczalne wartości podane w tabeli 2. Dlatego konieczne jest wyznaczenie momentów siły, wynikających z bezwładności wirników i dodanie ich (po odpowiednim przeliczeniu) do przebiegów czasowych wyznaczonych numerycznie w Simulinku. Momenty te można obliczyć w oparciu o profil przyspieszenia kątowego w przegubie, wykorzystując znajomość momentu bezwładności wirnika i przełożenia przekładni. Parametry te, dla poszczególnych przegubów, zebrano w tabeli 5. Odczytano je z plików `joint_size1_rev13.conf` oraz `joint_size3_rev13.conf`. Przeliczenie momentu z wirnika silnika na przegub odbywa się przez pomnożenie go przez kwadrat przełożenia przekładni.

Tabela 5. Momenty bezwładności wirników, przełożenia przekładni i stałe mechaniczne silników serwomechanizmów poszczególnych przegubów ramienia manipulatora

Numer przegubu	Nazwa przegubu	Moment bezwładności wirnika silnika serwomechanizmu [kg·m^2]	Przełożenie przekładni redukcyjnej serwo-mechanizmu	Stała mechaniczna (*torque constant*) silnika serwo-mechanizmu [N·m/A]
1 A	podstawa base	187,740e-06	110	0,15
2 B	bark shoulder	187,740e-06	110	0,15

3 C	łokieć elbow	187,740e-06	110	0,15
4 D	nadgarstek1 wrist1	2,0767e-05	110	0,09
5 E	nadgarstek2 wrist2	2,0767e-05	110	0,09
6 F	nadgarstek3 wrist3	2,0767e-05	110	0,09

Dla znalezienia suboptymalnych czasów realizacji poszczególnych ruchów robota dla przypadku ruchu typu *movej* niezbędne jest rozwiązanie prostego zadania dynamiki. Z kolei dla ruchu typu *movel* wymagane jest dodatkowo rozwiązanie odwrotnego zadania kinematyki. W niniejszym opracowaniu oba zadania rozwiązywane są w środowisku SimMechanics. Stosowne fragmenty schematów blokowych pokazano na rysunkach 7 i 8.

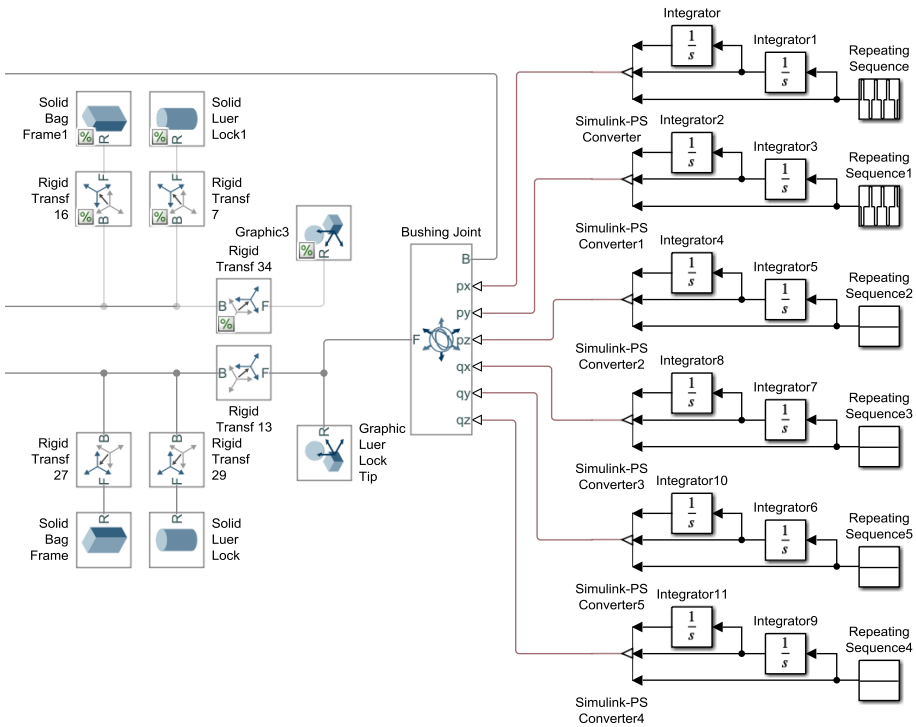


Rysunek 7. Fragment modelu w Simulinku wymuszający ruch przegubów manipulatora

6. Wyniki symulacji i eksperymentów

Poniżej zaprezentowane są wyniki zastosowania opisanej procedury optymalizacji czasu dla przykładowego ruchu typu *movej*. Jest on fragmentem ścieżki, wzdłuż której ramka z workiem jest przenoszona z magazynka worków na wagę elektroniczną. Konfiguracje oraz odpowiadające im pozy dla początkowego i końcowego punktu ruchu podane są w tabeli 6. W obliczeniach założono $r = 0{,}2$. W wyniku optymalizacji uzyskano czas $t_m = 0{,}7168$ s. Na rysunkach 9.-12. przedstawione są symulacyjne przebiegi czasowe wielkości fizycznych opisujących ruch ramienia, związanych z poszczególnymi przegubami: położeń kątowych, prędkości, przyspieszeń i momentów siły

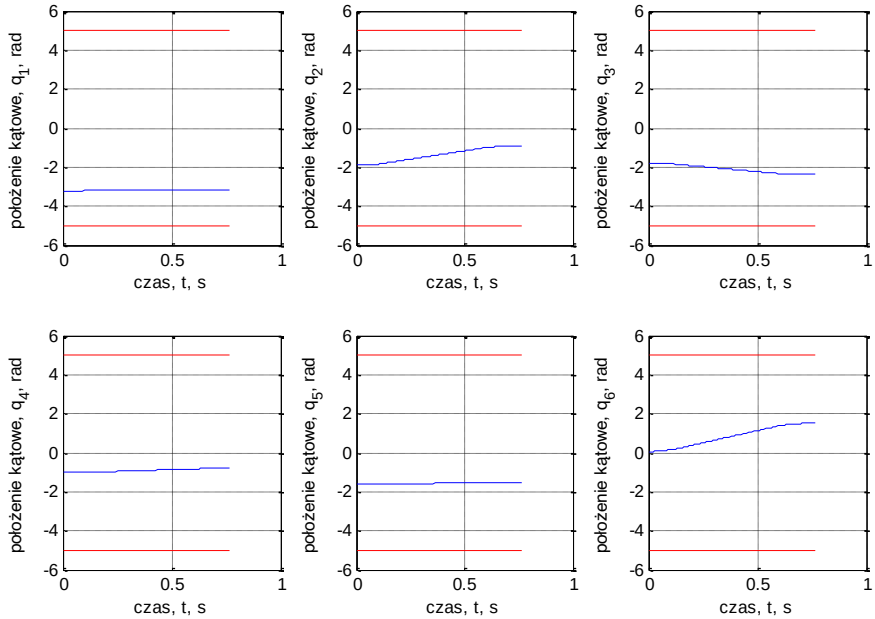
wywieranych przez serwomechanizmy. Przebiegi te odpowiadają trajektorii uzyskanej w wyniku optymalizacji. Na wykresach poziomymi liniami zaznaczono przyjęte dopuszczalne wartości graniczne, pomniejszone o 20% w stosunku do tych podanych przez producenta. Jak widać, nigdzie nie są one przekroczone, co świadczy o osiągnięciu założonego współczynnika bezpieczeństwa. Jednocześnie, aktywne jest ograniczenie dla przyspieszenia kątowego w drugim przegubie ramienia, co potwierdza uzyskanie wyniku suboptymalnego. Bliska ograniczeniu jest również prędkość kątowa w przegubie szóstym.



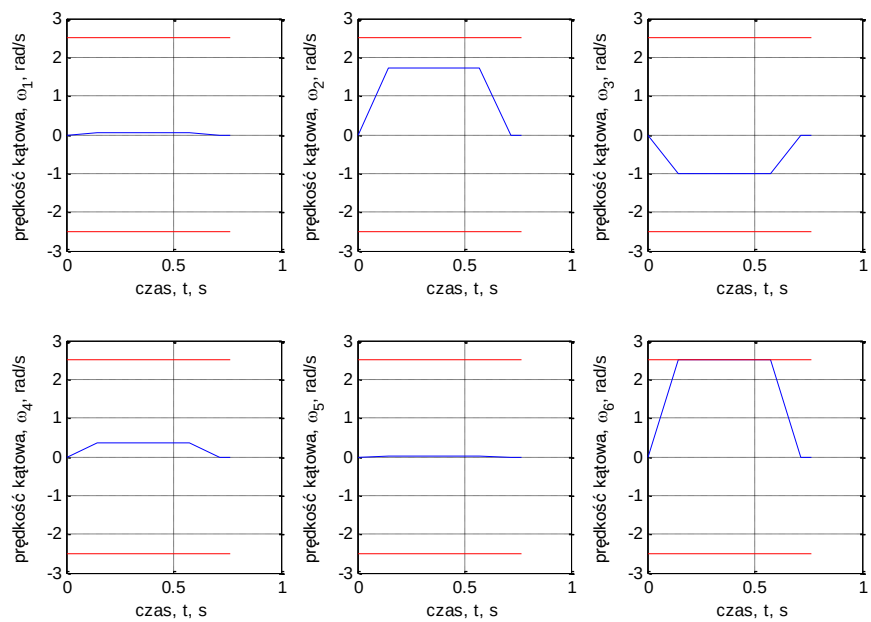
Rysunek 8. Fragment modelu w Simulinku wymuszający trajektorię końcówki Luer lock

Tabela 6. Konfiguracje i pozy (pozycje i orientacje TCP) punktów: początkowego i końcowego dla przykładowego ruchu *movej*

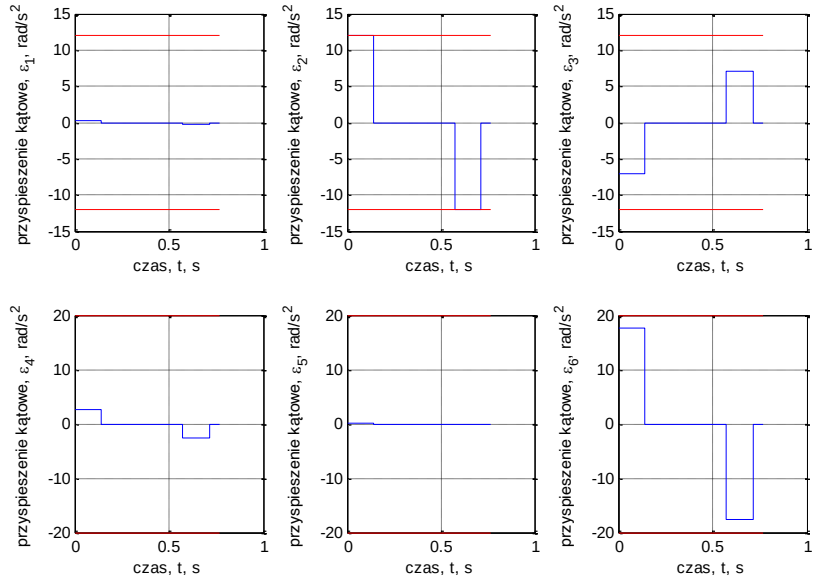
	Konfiguracja (współrzędne przegubów w osiach 1-6 [deg])	Poza (pozycja *xyz* [mm] i orientacja punktu TCP w osiach *xyz* [rad])
Punkt początkowy	−183,69 −108,90 −102,78 −58,32 −90,00 3,69	−557,90 145,39 681,53 −0,0000 0,0000 −1,5708
Punkt końcowy	−182,09, −52,38, −136,33, −46,31, −89,70, 86,16	25,43, 110,38, 745,37, −0,0144, 0,6107, −0,0988



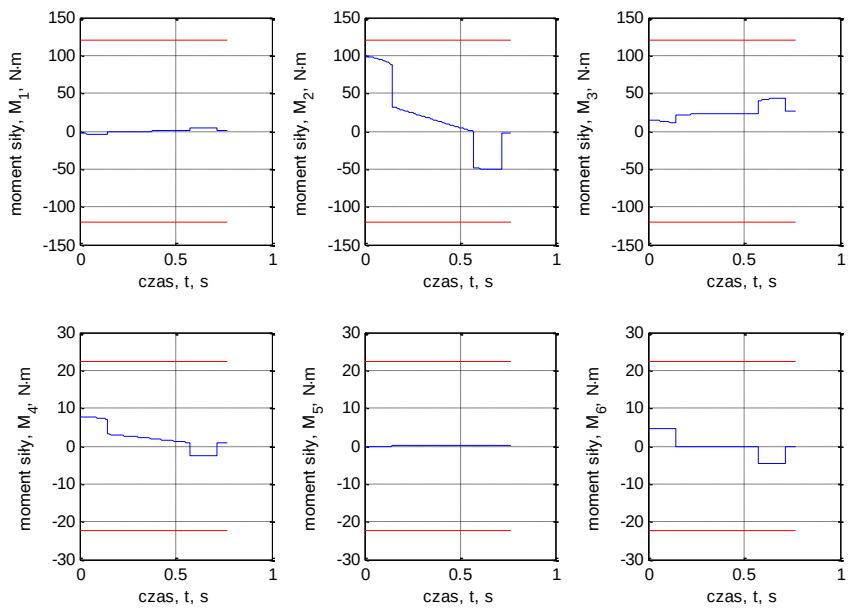
Rysunek 9. Przebiegi czasowe położeń kątowych poszczególnych przegubów manipulatora



Rysunek 10. Przebiegi czasowe prędkości kątowych poszczególnych przegubów manipulatora



Rysunek 11. Przebiegi czasowe przyspieszeń kątowych poszczególnych przegubów manipulatora



Rysunek 12. Przebiegi czasowe momentów siły w poszczególnych przegubach manipulatora

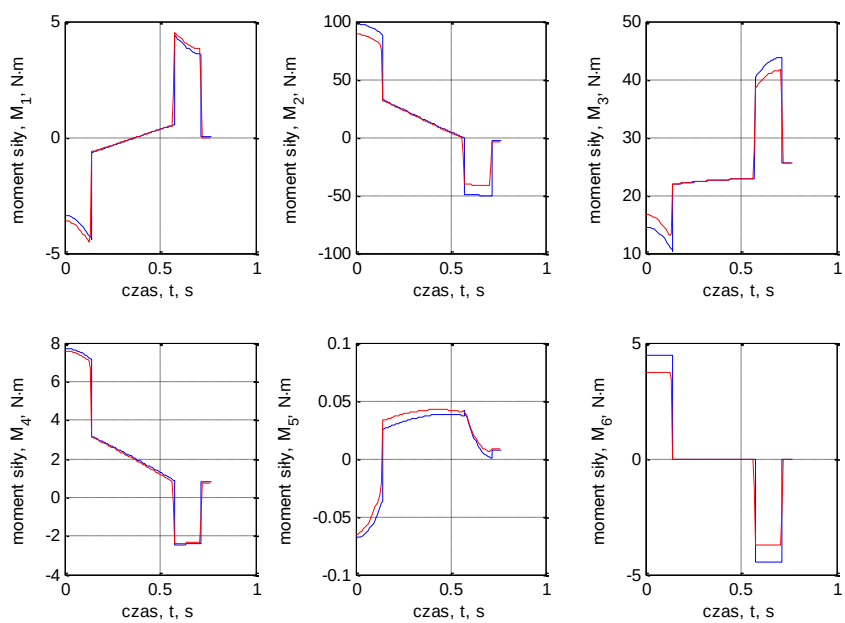
By zweryfikować poprawność wykonanych symulacji, porównano ich wyniki z przebiegami czasowymi wielkości fizycznych zarejestrowanych w rzeczywistym układzie regulacji. Do pozyskania tych danych wykorzystano mechanizm komunikacji *Real-Time Client*, udostępniany i obsługiwany przez sterownik robota. Za pośrednictwem sieci TCP/IP, z częstotliwością 125 Hz, sterownik wysyła pakiety danych niosących informacje o aktualnych wartościach położeń, prędkości, przyspieszeń,

momentów siły, prądów elektrycznych i innych wielkości fizycznych występujących w ramieniu manipulatora. Dane te zarejestrowano przy pomocy komputera klasy PC z systemem operacyjnym Linux, wydając w linii poleceń interpretera systemowego `bash` następującą, komendę złożoną z dwóch poleceń:

```
echo "play" > /dev/tcp/10.1.1.151/29999

cat < /dev/tcp/10.1.2.16 > data.dat
```

Pierwszy człon służy zdalnemu uruchomieniu programu wczytanego do środowiska PolyScope, pracującego na sterowniku robota, a wykorzystuje on mechanizm zdalnego sterowania zwany *Dashboard Server*. Drugi człon loguje dane płynące kanałem komunikacji *Real-Time Client*. Uzyskany plik binarny przekształcono do postaci CSV (*comma-separated values*), wykorzystując specjalnie w tym celu napisany program w języku C. Istotnym elementem programu była zamiana kolejności bajtów w zmiennych typu double, z konwencji sieciowej na obowiązującą w użytym komputerze klasy PC. Dane w postaci CSV zaimportowano do programu MATLAB poleceniem `csvread`. Na rysunku 13. porównano przebieg momentów siły rozwijanych przez serwonapędy zainstalowane w poszczególnych przegubach ramienia, wyznaczone symulacyjnie przy pomocy przybornika SimMechanics oraz uzyskane eksperymentalnie ze sterownika robota mechanizmem *Real-Time Client*. Warto tu zauważyć, że dane ze sterownika nie pochodzą z pomiarów (ramię nie jest wyposażone w czujniki momentów czy sił), lecz są przez sterownik obliczane, zapewne w celu użycia jako sygnały *feedforward* w pętlach regulacyjnych. Przebiegi na rysunku wykazują zadowalającą zgodność. Do pewnego stopnia potwierdza to prawidłowość wykonanych obliczeń i symulacji.



Rysuneku 13. Porównanie przebiegów czasowych momentów sił w poszczególnych przegubach manipulatora (linia czerwona – dane uzyskane ze sterownika robota mechanizmem Real-Time Client, linia niebieska – wyniki symulacji w środowisku: MATLAB, Simulink, Simscape, SimMechanics)

7. Podsumowanie

W artykule opisano uproszczoną metodę poszukiwania suboptymalnego czasu ruchu ramienia robotycznego realizującego zadanie transportu międzyoperacyjnego. Autorzy przetestowali ją w przedprototypie urządzenia przygotowującego roztwory leków cytostatycznych. Wykorzystuje ona symulacje komputerowe, a dzięki uproszczonemu sformułowaniu zadania, uniknięto konieczności stosowania iteracyjnych metod optymalizacji numerycznej. Optymalizacji podlegają poszczególne segmenty złożonej ścieżki robota, nie są przy tym modyfikowane ani punkty: początkowy i końcowy segmentów, ani charakter profili położeń, prędkości i przyspieszeń.

W opisywanej optymalizacji przyjęto założenie stałości stosunku $t_\mathrm{a}/t_\mathrm{m} = t_\mathrm{d}/t_\mathrm{m} = r$. Zamiast tego można r uczynić drugą, obok t_m, zmienną decyzyjną lub też wrócić do pierwotnych zmiennych (przyspieszenia szczytowego i prędkości szczytowej). Zadanie optymalizacji skomplikowałoby się wówczas, ale uzyskalibyśmy większą swobodę kształtowania ruchu, a co za tym idzie – możliwość dalszego poprawienia uzyskanego wyniku. Rozwiązanie tak postawionego problemu wymagałoby zapewne użycia procedur iteracyjnych.

W przedstawionych wyżej obliczeniach pominięto wpływ momentów żyroskopowych powodowanych przez wirujące rotory silników serwonapędów przegubów. Nie uwzględniono również tarcia suchego statycznego ani tarcia mokrego w przekładniach serwomechanizmów i łożyskach przegubów. Wzięcie tych zjawisk pod uwagę mogłoby poprawić dokładność otrzymanych rezultatów, skomplikowałoby jednak postać modelu i zwiększyło czas obliczeń. Zaszłaby też konieczność pozyskania informacji o wymaganych parametrach urządzenia: momentach bezwładności wirników oraz współczynnikach tarcia.

Literatura

1. Prusak D., Kobus K., Żarów J., Kańtoch E., *Automatyzacja przygotowywania leków cytostatycznych*, Projektowanie mechatroniczne. Zagadnienia wybrane, praca zbiorowa pod red. Michała Mańki, Katedra Robotyki i Mechatroniki, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków 2015, s. 181-192.
2. Geersing T.H., Franssen E.J.F., Pilesi F., Crul M., *Microbiological performance of a robotic system for aseptic compounding of cytostatic drugs*, European Journal of Pharmaceutical Sciences, 130, 2019, s. 181-185.
3. Geersing T.H., Klous M.G., Franssen E.J.F., van den Heuvel J.J.G., Crul M., *Robotic compounding versus manual compounding of chemotherapy: Comparing dosing accuracy and precision*, European Journal of Pharmaceutical Sciences, 155, 2020, 105536.
4. Rotter P., Muroń W., *Algorytmy wizyjnego systemu sterowania i kontroli dla zautomatyzowanego procesu przygotowywania leków cytostatycznych*, NOT, Rzeszów 2016.
5. Universal Robots, Universal Robots User Manual, 2018.
6. Universal Robots, Service Manual. Robot UR5 with CB2 controller.
7. Universal Robots, PolyScope Manual, 2018.
8. Universal Robots, The URScript Programming Language, 2013.
9. Vukobratovic M., Stokic D., *Contribution to suboptimal control of manipulation robots*, IEEE Transactions on Automatic Control, 10, 1983, s. 981-985.
10. Bobrow J.E., Dubowsky S., Gibson J.S., *Time-optimal control of robotic manipulators along specified paths*, International Journal of Robotics Research, 3, 1985, s. 3-17.

11. Dubowsky S., Norris M.A., Shiller Z., *Time-optimal robotic manipulator task planning*, RoManSy 6. Proceedings of the Sixth CISM-IFToMM Symposium on Theory and Practice of Robots and Manipulators, 1987, s. 615-622.
12. Sontag E.D., *Time-optimal control of manipulators*, IEEE International Conference on Robotics and Automation, USA, San Francisco 1986.
13. Cheah C.-C., Wang D., *The practical implementation of time-optimal control for robotic arm manipulators*, IEEE Transactions on Robotics, 3, 1998, s. 452-465.
14. Gasparetto A., Zanotto V., *A technique for time-jerk optimal planning of robot trajectories*, Robotics and Computer-Integrated Manufacturing, 3, 2008, s. 415-426.
15. Ghasemi M.H., Kashiri N., Dardel M., *Time-optimal trajectory planning of robot manipulators in point-to-point motion using an indirect method*, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2, 2012, s. 473-484.
16. Menon A., Prakash R., Behera L., *Adaptive critic based optimal kinematic control for a robot manipulator*, IEEE International Conference on Robotics and Automation, 2019, s. 1316-1322.
17. Ferrentino E., Della Cioppa A., Marcelli A., Chiacchio P., *An evolutionary approach to time-optimal control of robotic manipulators*, Journal of Intelligent & Robotic Systems, 2, 2020, s. 245-260.
18. Kim M.S., Han D.K., Park J.H., Kim J.S., *Motion planning of robot manipulators for a smoother path using a twin delayed deep deterministic policy gradient with hindsight experience replay*, Applied Sciences, 2, 2020, 575.
19. Mathworks, Mathworks Simscape – Manual, Dostępne: https://www.mathworks.com/help/physmod/simscape/.
20. Morecki A., Knapczyk J., Galicki M., Marszalec E., Marszalec J., Wiliński A., Zielińska T., *Podstawy robotyki. Teoria i elementy manipulatorów robotów*, Wydawnictwa Naukowo-Techniczne, Warszawa 1999.
21. Spong M.W., Vidyasagar M., *Dynamika i sterowanie robotów*, Wydawnictwa Naukowo-Techniczne, Warszawa 1997.
22. Universal Robots, DH parameters for calculation of kinematics and dynamics.
23. Universal Robots, Kinematic Calibration Manual for CB3, 2009.
24. Universal Robot, Technical Specifications UR5, 2016.
25. Universal Robots, MAX. JOINT TORQUES, Dostępne: https://www.universal-robots.com/articles/ur/robot-care-maintenance/max-joint-torques/.
26. Kufieta K., *Force Estimation in Robotic Manipulators: Modeling, Simulation and Experiments. The UR5 Manipulator as a Case Study*, Diploma Thesis, Norwegian University of Science and Technology, 2014.

Dobór suboptymalnych czasów realizacji elementarnych ruchów manipulatora w urządzeniu przygotowującym leki cytostatyczne

Streszczenie

W układach automatyki, w których roboty przemysłowe wykonują zadanie przenoszenia obrabianego elementu, podstawową wielkością podlegającą optymalizacji jest często czas ruchu. Kształt ścieżki chwytaka jest zwykle podyktowany ograniczeniami przestrzeni roboczej i nie podlega optymalizacji. Ograniczony zestaw instrukcji języka programowania robota narzuca z kolei kształt profilów prędkości. Stąd zmiennymi decyzyjnymi w zadaniu optymalizacji mogą być wyłącznie maksymalne prędkości i przyspieszenia przekazywane jako parametry do funkcji realizujących elementarne segmenty trajektorii. Ograniczeniami są z kolei maksymalne dopuszczalne konstrukcje robota prędkości i przyspieszenia przegubowe oraz momenty siły serwomechanizmów. Przekroczenie ograniczeń skutkuje zwykle awaryjnym zatrzymaniem robota. W rozdziale zaprezentowano uproszczoną metodę poszukiwania sterowania suboptymalnego. Autorzy zastosowali ją do ramienia robota stanowiącego podzespół przedprototypu urządzenia do przygotowania

roztworów leków cytostatycznych. Każdy segment złożonej trajektorii jest tu optymalizowany niezależne. Jedyną zmienną decyzyjną w każdym podzadaniu jest czas realizacji ruchu elementarnego. Definiuje on w pełni charakter ruchu, gdyż stosunek czasów rozpędzania i hamowania do czasu całkowitego jest zadany z góry. Każdy elementarny ruch jest dwukrotnie symulowany w środowisku Simulink/SimMechanics pakietu MATLAB. Pierwsza symulacja służy wyznaczeniu grawitacyjnych składowych przebiegów momentów siły serwomechanizmów. Druga daje informacje o całościowych przebiegach momentów dla jednostkowego czasu ruchu. Tak uzyskane dane, wraz z podanymi przez producenta granicznymi wartościami dopuszczalnymi, pozwalają wyznaczyć poszukiwany suboptymalny czas trwania ruchu elementarnego. Nie jest przy tym konieczne stosowanie procedur iteracyjnych. Działanie opisanej metody zilustrowano wynikami symulacji komputerowych oraz doświadczeń przeprowadzonych na stanowisku do przygotowania leków z robotem UR5.
Słowa kluczowe: sterowanie suboptymalne czasowo, optymalizacja ruchu ramienia robota, stanowisko przygotowania leku, symulacja dynamiczna wielu ciał

Determination of suboptimal durations of elementary manipulator movements in a cytostatic drug preparing station

Abstract

In an automation system, where an industrial robot performs a task of moving and positioning a workpiece, the main objective function to be optimised is often the movement duration. The shape of a gripper path is usually determined by geometrical constraints of a working space and does not undergo any optimization. Moreover, a limited instruction set of a robot programming language determines the general character of workpiece or joints velocity profiles. Hence, the only quantities that can be selected as decision variables are maximum speeds and accelerations passed as arguments to functions implementing elementary trajectory segments. The constraints are: maximum allowable joint speeds, accelerations, and servomechanisms torques. Exceeding the limits usually results in an emergency stop of the robot. The chapter presents a simplified method for suboptimal control search. The authors applied it to a robot arm being a component of a pre-prototype of a cytostatic drug solutions preparation station. Each segment of a complex trajectory is optimized independently. The only decision variable in each subproblem is the duration of the elementary move. It fully defines the character of the motion, since a ratio of acceleration and deceleration durations to the total movement duration is set in advance arbitrarily. Each elementary movement is simulated twice in a Simulink/SimMechanics environment of MATLAB packet. The first simulation is used to determine gravitational components of servomechanisms torque profiles. The second one provides overall torque profiles corresponding to unitary movement duration. Data obtained this way, together with limit values specified by the manufacturer, allow to determine sub-optimal durations of elementary movements. The method does not require any iterative calculations. Performance of the described method is illustrated with results of computer simulations and real-world experiments carried out on a drug preparation station with the UR5 robot.
Keywords: time suboptimal control, robotic arm movement optimisation, drug preparation station, dynamic multibody simulation

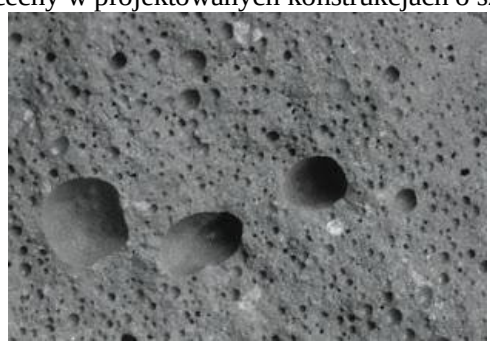
Danuta Miedzińska[1], Radosław Ciepielewski[2]

Rozwój nowych technologii materiałowych – materiały porowate i cellulary w przemyśle XXI wieku

1. Wprowadzenie

Materiały porowate, inaczej zwane pianami, to struktury zbudowane z litej matrycy, w której znajdują się pory – puste przestrzenie wypełnione cieczą lub gazem.

Materiały porowate są znane ludziom od wieków, szczególne ze względu na ich występowanie w przyrodzie czy środowisku naturalnym. Przykładami takich materiałów mogą być niektóre typy skał, np. magmowe, drewno czy gąbki – organizmy żyjące w morzach i oceanach (rys. 1). Rozwój metod wytwarzania spowodował, iż ludzie nauczyli się produkować materiały o strukturze porowatej oraz wykorzystywać ich unikalne cechy w projektowanych konstrukcjach o szerokim zakresie stosowania.

a)  b)

Rysunek 1. Naturalne materiały porowate a) struktura skały magmowej [1], b) gąbka morska [2]

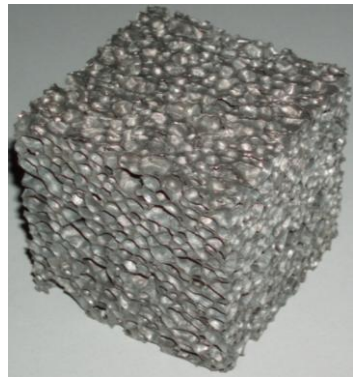
Materiały porowate mogą mieć budowę losową (rys. 2) bądź uporządkowaną, mieć matrycę wykonaną z różnych materiałów, np. metali i ich stopów, ceramiki, różnego typu polimerów, a nawet papieru. Jednym z przykładów takich materiałów, które zasługują na uwagę, są tzw. cellulary – materiały o regularnej strukturze plastra miodu, wykonane ze stopów metali, ale również wykorzystywane ostatnio w przemyśle opakowaniowym (rys. 3).

Istnieje wiele metod klasyfikacji materiałów porowatych. Główną cechą opisującą te materiały jest porowatość, czyli udział przestrzeni porowej w objętości materiału, wyrażona wzorem:

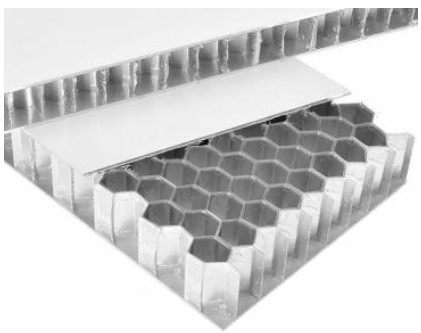
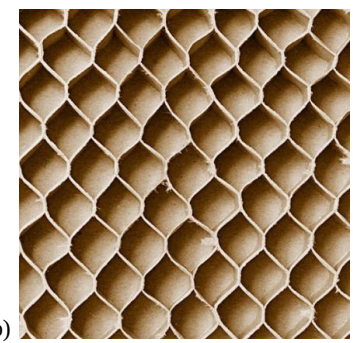
$$p = \frac{V_p}{V} \tag{1}$$

gdzie: V_p – objętość porów, V – objętość całego ciała.

[1] danuta.miedznska@wat.edu.pl, Instytut Mechaniki i Inżynierii Obliczeniowej, Wydział Inżynierii Mechanicznej, Wojskowa Akademia Techniczna, www.wat.edu.pl.
[2] radoslaw.ciepielewski@wat.edu.pl, Instytut Mechaniki i Inżynierii Obliczeniowej, Wydział Inżynierii Mechanicznej, Wojskowa Akademia Techniczna, www.wat.edu.pl.



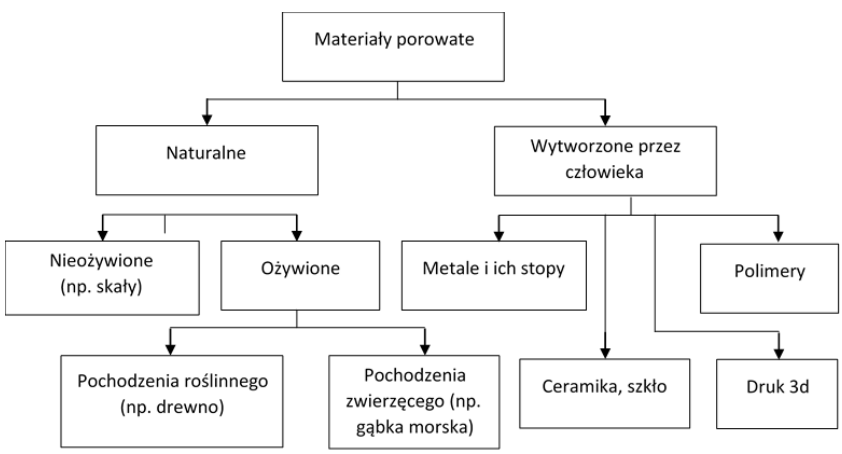
Rysunek 2. Piana aluminiowa o budowie losowej

a)  b) 

Rysunek 3. Cellulary a) aluminiowe [3], b) papierowe [4]

Ze względu na rodzaj porowatości, omawiane materiały (bez względu na materiał, z jakiego została wykonana matryca) dzieli się na te o porach otwartych (pory są ze sobą połączone, może następować między nimi przepływ cieczy lub gazu) oraz o porach zamkniętych (pory nie są ze sobą połączone).

Inny podział struktur porowatych dotyczy materiałów, z jakich może być wykonana matryca piany. Schemat takiego podziału przedstawiona na rysunku 4.



Rysunek 4. Klasyfikacja materiałów porowatych ze względu na materiał matrycy

2. Charakterystyka i metody wytwarzania materiałów porowatych z uwzględnieniem materiału matrycy

2.1. Materiały porowate o matrycy z metali i ich stopów

2.1.1. Piany metaliczne

Piany metaliczne, zwane też spienionymi metalami, należą wciąż do grupy nowoczesnych materiałów, którymi zainteresowanie rośnie w ostatnich latach. Do zalet tych materiałów można zaliczyć małą masę, dużą wytrzymałość mechaniczną i izolacyjność elektromagnetyczną, cieplną i akustyczną, odporność na oddziaływanie środowisk agresywnych chemicznie. Wadą jest trudność ich łączenia z innymi elementami istniejących konstrukcji i przejmowania od nich obciążeń [5].

Piany metaliczne są szeroko stosowane w przemyśle, np. budownictwie, budowie statków, samochodów, kontenerów, pojazdów szynowych. Są wykorzystywane jako lekkie, dźwiękochłonne ścianki działowe, płyty podłogowe lub sufitowe, ekrany elektromagnetyczne. Wybrane przykłady zastosowań pian metalicznych przedstawiono w tabeli 1.

Tabela 1. Przykłady zastosowań pian metalicznych [6]

Zastosowanie	Cecha piany metalicznej
Konstrukcje lekkie (np. panele typu sandwich)	Wysoka względna sztywność i względna granica plastyczności
Mechaniczne tłumienie drgań i absorpcja akustyczna	Zdolność tłumienia jest o rząd wielkości wyższa niż w przypadku metali litych.
Pochłanianie energii zderzenia	Wysoka energochłonność
Termoizolacja i ochrona przed płomieniem	Wysoka powierzchnia do przenoszenia ciepła w połączeniu z krawędziami komórek o wysokiej przewodności cieplnej.
Filtry	Piany o porach otwartych mogą być wytwarzane przy zachowaniu zadanej wielkości porów
Elektrody i katalizacja	Wysoka powierzchnia na jednostkę objętości w piankach o porach otwartych

Piany metaliczne są produkowane z różnych materiałów (tab. 2), takich jak magnez, tytan, austenityczne stale nierdzewne i stopy niklu. Najczęściej jednak do wytwarzania pian stosuje się aluminium i jego stopy. Reprezentatywne cechy komercyjnie dostępnych pian aluminiowych pokazano w tabeli 3.

Tabela 2. Zestawienie materiałów stosowanych na piany metaliczne [7]

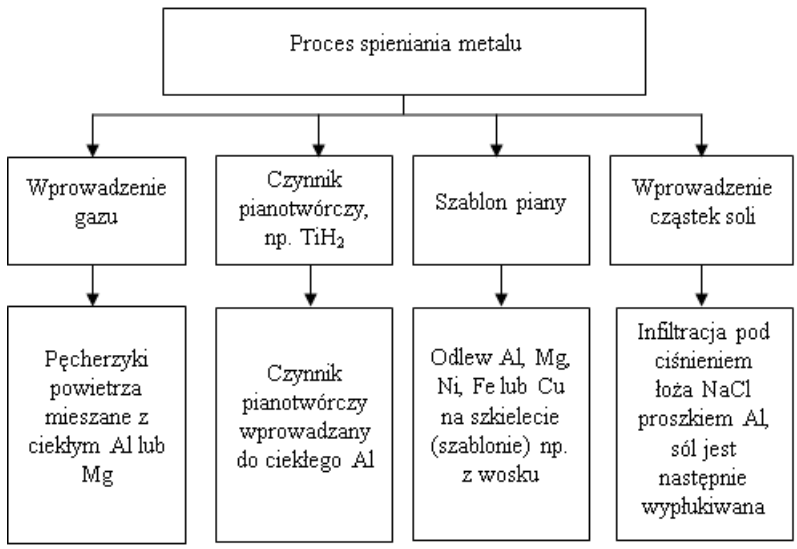
Materiał	Gęstość piany (kg m^{-3})
MCrAlY (M = Fe lub Ni-Co)	360-420
Cu/60-40 mosiądz	450/420
stal (eutektyczna i austenityczna stal nierdzewna)	316-390
nikiel	450
srebro	530
tytan	230

Tabela 3. Wybrane właściwości komercyjnie dostępnych pian aluminiowych [8]

Właściwość	Wartość
Materiał	czyste Al lub kompozyt Al-SiC
Struktura	Pory otwarte lub zamknięte

Gęstość piany (kg m^{-3})	70-1 000
Gęstość piany / gęstość równoważnej objętości materiału litego	0,02-0,35
Moduł Younga (MPa)	20-12 000
Wytrzymałość na ściskanie (kPa)	40-14 000
Wytrzymałość na rozciąganie (kPa)	50-30 000
Wytrzymałość na zginanie (kPa)	40-25 000
Odporność na kruche pękanie (kPa m)	30-1 600
Pojemność cieplna właściwa (J kg^{-1} K^{-1})	830-950
Przewodność cieplna (mWm^{-1} K^{-1})	300-35 000

Piany metaliczne są otrzymywane najczęściej metodami ciekło-fazowymi (metodą spieniania w stanie ciekłym i odlewania, metalurgii proszków) oraz – rzadziej – metodami osadzania i naparowywania (osadzanie chemiczne, elektrochemiczne, naparowywanie próżniowe). Metody wytwarzania pian metali pokazano na rysunku 5.

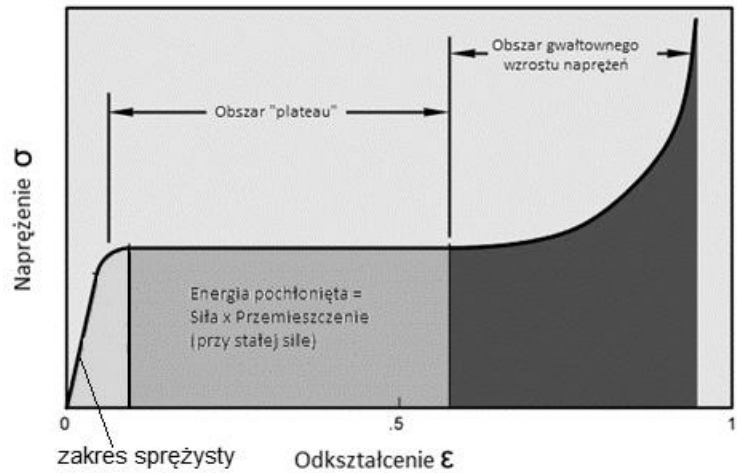


Rysunek 5. Zestawienie metod wytwarzania pian metalicznych [9]

Jak już wspomniano, jednym z istotnych zastosowań pian metalicznych są elementy energochłonne. Z tego względu często analizuje się proces ściskania tych materiałów. Typowy wykres naprężenie w funkcji odkształcenia dla tego procesu pokazano na rysunku 6.

Na wykresie tym można wyraźnie zaobserwować trzy charakterystyczne i znaczące dla energochłonności obszary:

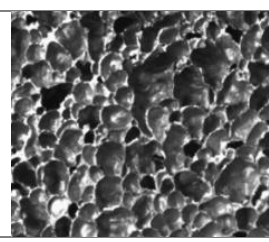
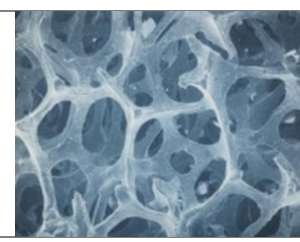
- sprężysty (linia prosta pochyła) – określa on moduł Younga piany, w tym etapie materiał bazowy piany (np. aluminium) pracuje sprężyście;
- „plateau" (linia prosta pozioma) – w tym etapie następuje zamykanie się porów, jest to główny obszar zainteresowań materiałoznawców, gdyż jest najbardziej znaczący dla energochłonności piany, dąży się do tego, aby był jak najdłuższy;
- gwałtowny wzrost naprężeń (ostatni fragment krzywej) – rozpoczyna się proces zgniatania litego materiału powstałego po zamknięciu porów, stąd gwałtowny wzrost naprężeń.



Rysunek 6. Typowa charakterystyka naprężenie – odkształcenie dla ściskania materiału pianowego

W tabeli 4 porównano cechy piany aluminiowej o porach zamkniętych ALPORAS firmy Gleich i o porach otwartych DUOCEL firmy ERG Aerospace.

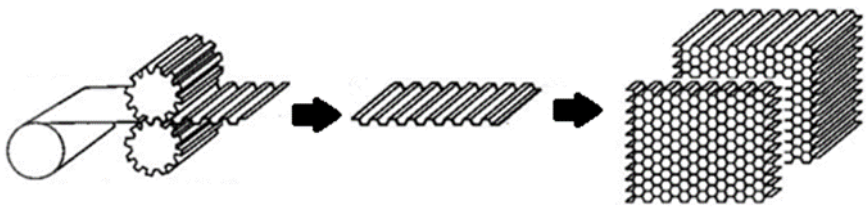
Tabela 4. Porównanie właściwości piany aluminiowej o porach zamkniętych ALPORAS firmy Gleich i o porach otwartych DUOCEL firmy ERG Aerospace [10]

Cecha	ALPORAS	DUOCEL
Moduł Younga	0,7 GPa	103,08 MPa
Wytrzymałość na rozciąganie R_m	1,6 MPa	1,24 MPa
Wytrzymałość na ściskanie	1,5 MPa	2,53 MPa
Struktura		

2.1.2. Cellulary

Jednym z rodzajów materiałów o budowie porowatej, ale uporządkowanej, są cellulary, inaczej zwane strukturami ulowymi czy materiałami o strukturze plastra miodu. Wytwarzane są najczęściej z cienko walcowanych blach – folii metalowych. Znane są również konstrukcje tego typu wykonane z kompozytów włóknistych w osnowie polimerowej (wykorzystywane najczęściej w przemyśle lotniczym na zewnętrzne segmenty poszycia kadłuba, skrzydeł czy elementy usterzenia z uwagi na wyższy stosunek sztywności do masy) czy z masy papierowej (przemysł opakowaniowy). Cellulary wykonane z metali są najczęściej stosowane i wykorzystywane jako elementy absorberów energii kinetycznej.

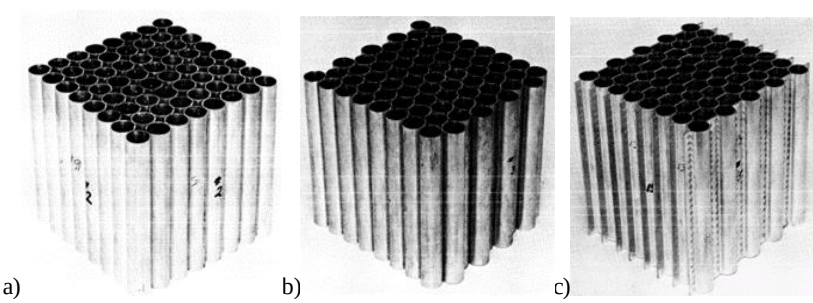
Cellulary powstają z blach kształtowanych wstępnie między zębatymi rolkami blach. Kolejne pakiety uformowanych blach łączone są ze sobą adhezyjnie. Blachy układane są na sobie i klejone w sposób pozwalający na uzyskanie powtarzalnych, sześciokątnych komórek. Następnie pakiety wstępnie ukształtowanych blach dociskane są do siebie do czasu zastygnięcia spoiwa, a później rozciągane do uzyskania pożądanego kształtu komórki. Proces ten zilustrowano na rysunku 7.



Rysunek 7. Proces tworzenia aluminiowego plastra miodu [11]

W efekcie powstaje struktura posiadająca charakterystyczne parametry geometryczne, rzutujące bezpośrednio na jej walory użytkowe i wytrzymałościowe.

Struktury o zorientowanej budowie komórkowej, m.in. w formie plastra miodu, posiadają długą historią udanych aplikacji. Zostały na przykład użyte w początkowym okresie podbojów kosmosu, w projekcie Apollo 11 do budowy elementów przedziału załogowego i technicznego. Badacze NASA [12], testując szeroką gamę tego rodzaju materiałów, których przykłady przedstawiono na rysunku 8., doszli do wniosku, że najlepsze własności energochłonne w stosunku do masy właściwej posiadają elementy o topologii plastra miodu.



a) b) c)

Rysunek 8. Przykłady elementów energochłonnych badanych w programie Apollo: a) rurki w szyku prostym; b) rurki w szyku naprzemiennym; c) plaster miodu [12]

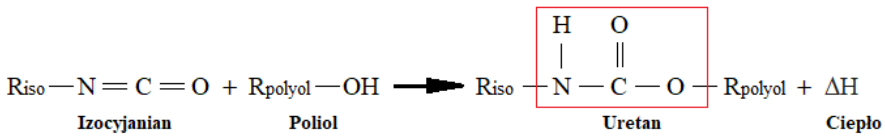
Wraz z rozwojem materiałów o budowie zorientowanej, komórkowej, zaczęto wykorzystywać ich właściwości do tworzenia konstrukcji kombinowanych – jako wypełniacz struktur przekładkowych czy sandwiczowych.

2.2. Materiały porowate o matrycy z tworzyw sztucznych

W ostatnich latach notuje się szybki rozwój w dziedzinie tworzyw sztucznych, w tym w zakresie wytwarzania i zastosowania spienionych tworzyw sztucznych, przede wszystkim pian polimerowych np. poliuretanowych.

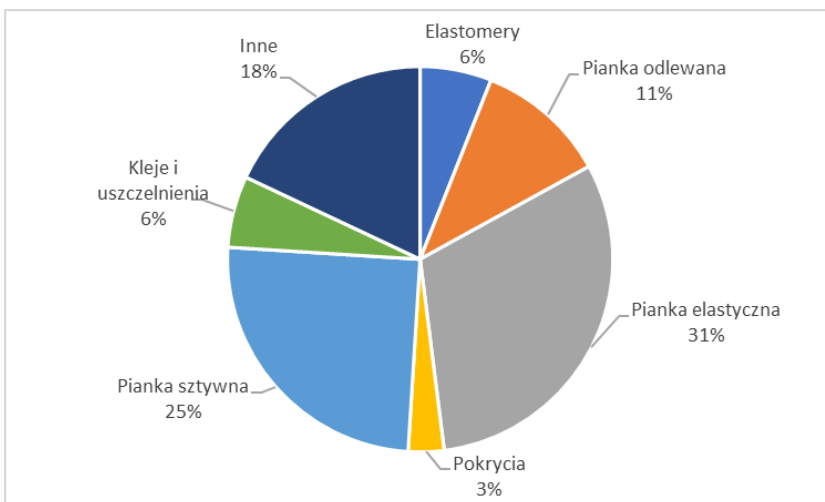
Pierwszy uretan został zsyntetyzowany w 1849 roku przez Wurtza. Następnie, w 1937 roku, Otto Bayer syntetyzował poliuretan w reakcji między diolem poliestrowym a diizocyjanianem. Początkowo polimer ten został uznany za bezużyteczny [13].

Poliuretany są polimerami, które powstają w wyniku reakcji pomiędzy grupami OH (hydroksylowych) poliolu z grupami NCO (grupa funkcjonalna izocyjanianu), a nazwa jest związana z powstałym połączeniem uretanu [13]. Reakcja ta ma charakter egzotermiczny i prowadzi do produkcji grup uretanu, co pokazano na rysunku 9 [14].



Rysunek 9. Schemat reakcji wytwarzania uretanu, gdzie Riso pochodzi z monomeru izocyjanianu, podczas gdy Rpolyol pochodzi ze składnika poliolu [14]

Obecnie poliuretany są używane powszechnie w produktach życia codziennego. Światowa konsumpcja poliuretanu została oszacowana na 60,5 mld USD w 2017 roku. Przewiduje się, że do 2021 roku wyniesie ona ponad 79 mld USD [15]. W 2016 roku stanowiła prawie 9% światowego zużycia tworzyw sztucznych [16]. Ponadto, jak pokazano na rysunku 10., poliuretan jest używany głównie w postaci pianek [17].



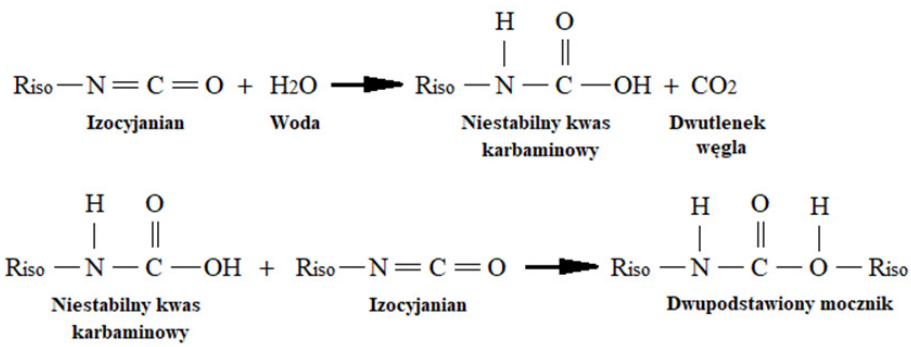
Rysunek 10. Światowa konsumpcja poliuretanu w 2016 roku [17]

Piany poliuretanowe stanowią więc 67% globalnej konsumpcji poliuretanu. Ponadto, ponieważ technologia produkcji tych pian jest powszechnie znana, stanowią one połowę rynku całej pianki polimerowej [11].

Poliole, poliestry lub polietery są podstawowym surowcem do produkcji pianek poliuretanowych. Związki te mają w cząsteczce dwie lub więcej grup wodorotlenowych. Często stosuje się także olej rycynowy i jego pochodne. Do produkcji pianek elastycznych stosuje się mieszaniny polioli zawierające więcej składników dwufunkcyjnych, a do produkcji pianek sztywnych stosuje się większe dodatki polioli trój- lub więcej

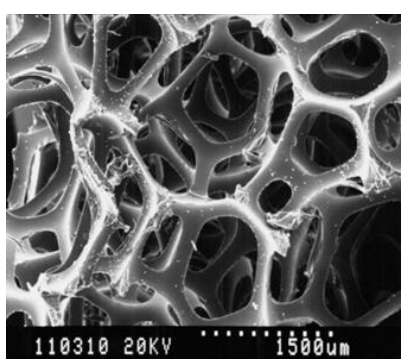
funkcyjnych. Od charakteru i funkcyjności poliolu zależą przede wszystkim właściwości pianki, takie jak wytrzymałość i odporność chemiczna [18].

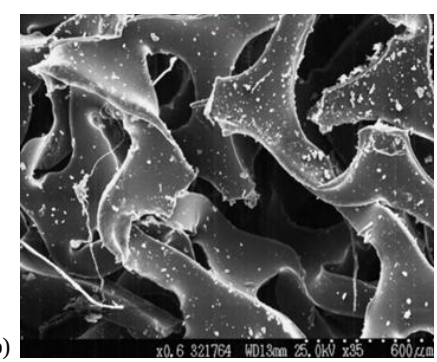
Środkami spieniającymi w poliuretanach jest woda lub niskowrzące, obojętne rozpuszczalniki zwykle chlorowcopochodne. Woda reaguje z grupami izocyjanianowymi, wydzielając CO_2, zaś niskowrzące rozpuszczalniki odparowują podczas egzotermicznej reakcji powstawania poliuretanów, tworząc pory. Reakcję izocyjanianu z wodą pokazano na rysunku 11.



Rysunek 11. Schemat reakcji izocyjanianu z wodą [14]

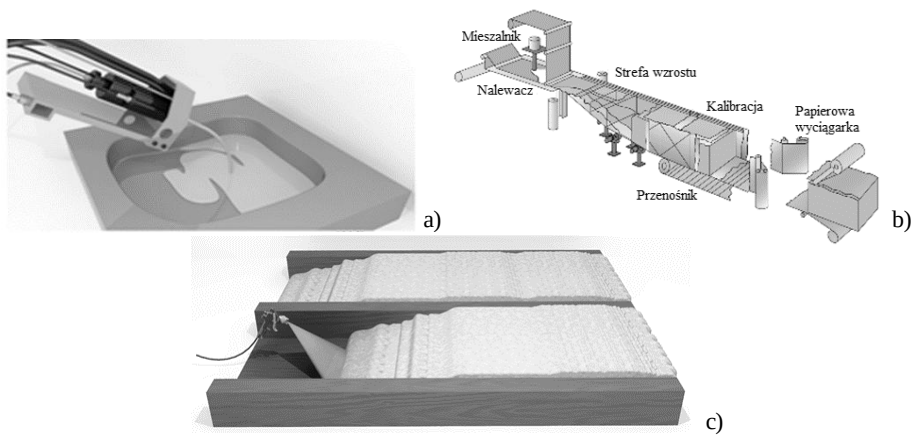
Piany polimerowe (rys. 12a) charakteryzują się mniejszą wytrzymałością niż piany metaliczne. Piany o porach otwartych mogą być natomiast specjalnie preparowane w celu modyfikacji właściwości. Preparacja polega głównie na ściśnięciu struktury, aż do uzyskania zniszczenia niektórych połączeń w strukturze (rys. 12b). Można dzięki temu otrzymać tzw. materiał auksetyczny, dla którego współczynnik Poissona przyjmuje wartości ujemne. Taki materiał zachowuje się odmiennie od typowego – pod wpływem ściskania jednoosiowego zmniejsza pole przekroju poprzecznego, a pod wpływem rozciągania – zwiększa. Materiały tego typu bardzo dobrze nadają się do produkcji siedzeń w pojazdach. Zapewniają dobre „trzymanie" ciała pasażera. Inną ciekawą właściwością materiałów pianowych jest tzw. pamięć kształtu.

a) 

b) 

Rysunek 12. a) Piana o porach otwartych, b) wskutek preparacji polegającej na ściśnięciu niektóre z łączników w pianie mogą zostać zniszczone (przerwane). Piana o zmodyfikowanej strukturze ma właściwości auksetyczne [19]

W zależności od wymaganych właściwości lub zastosowań, do produkcji pian poliuretanowych można stosować różne technologie, przy czym najczęściej stosowane w skali produkcyjnej są: odlewanie (rys. 13a), „slabstock" (rys. 13b) i wtryskiwanie (rys. 13c). W procesie odlewania mieszaninę reagentów wstrzykuje się do formy. Po utwardzeniu uformowana pianka jest usuwana z formy; jest to proces, który jest powszechnie stosowany do produkcji poduszek siedzeń dla samochodów i mebli. Pianki „slabstock" są wytwarzane przez ciągłe wylewanie mieszaniny reagentów na ruchomym przenośniku. Tą metodą wytwarzane są elementy o wysokiej jakości powierzchni i dokładności wymiarowej. Wtrysk poliuretanu polega na nanoszeniu go na danej powierzchni. Jest on zwykle używany do produkcji warstw izolacyjnych na płaskich powierzchniach, takich jak dachy, ale także takich jak sferyczne zbiorniki i rury oraz konstrukcje budowlane [20].



Rysunek 13. Procesy wytwarzania piany poliuretanowej: a) odlewanie, b) „slabstock", c) wtryskiwanie [21]

Chociaż produkcja pian poliuretanowych jest znaną technologią, ze względu na potrzebę wykorzystania nowych surowców i/lub wypełniaczy, a także opracowania produktów o nowych formach, kształtach i wymiarach, w przyszłości mogą być wymagane nowe technologie produkcyjne, takie jak opisany poniżej druk 3D.

Właściwości materiałowe wybranych, dostępnych komercyjnie pian poliuretanowych pokazano w tabeli 5.

Tabela 5. Właściwości materiałowe i rozmiary komórek pian poliuretanowych badanych w pracy [22]

Producent piany	Gęstość [kg/m^3]	Gęstość względna [-]	Naprężenia przy 20% ściśnięciu [kPa]	Średnica komórki [μm]
Royal medica	130,8	0,108	6,8	300 ±110
Dynamics systems – Sunmate soft	87,0	0,072	5,6	290 ±180
Dynamics systems – Sunmate medium	79,9	0,066	13,7	290 ±100
Dynamics systems – Sunmate firm	82,3	0,069	20,9	290 ±90
Dynamics systems – Pudgee	221,9	0,185	2,7	330 ±180

Nowością na rynku pian poliuretanowych jest wprowadzenie surowców odnawialnych do receptur piankowych w zamian klasycznego petrochemicznego poliolu [23]. Źródłem nowych surowców poliolowych są przede wszystkim oleje roślinne (rzepakowy, palmowy, sojowy) oraz glikolizaty [24]. Rośnie również zainteresowanie stosowaniem nanonapełniaczy głównie glinokrzemianów, które zaczynają być stosowane jako modyfikatory spienionych materiałów poliuretanowych [25].

Ze względu na różnorodność zastosowań, produkcja pian poliuretanowych wzrosła w ostatnich dziesięcioleciach, co prowadzi do wzrostu ilości odpadów poliuretanowych. Odpady wytwarzane w produkcji pian lub odzyskiwane po ich zużyciu mogą być unieszkodliwiane na składowiskach lub spalane. Jednak długoterminowe cele ekologiczne nakazują gospodarkę bezodpadową, jak również ochronę surowców. W rzeczywistości recykling jest nie tylko wymogiem w zapobieganiu zanieczyszczeniom i ochronie środowiska, ale także jest potrzebny w celu obniżenia kosztów produkcji i zwiększenia wydajności wykorzystania materiału. Istnieją dwie główne metody recyklingu pian poliuretanowych: recykling fizyczny i recykling chemiczny. W recyklingu fizycznym kawałki poliuretanu są bezpośrednio ponownie wykorzystywane bez obróbki chemicznej, podczas gdy recykling chemiczny polega na depolimeryzacji poliuretanu na oligomery, które mogą być później wykorzystywane jako reagenty do produkcji nowych materiałów poliuretanowych [26].

2.3. Materiały porowate o matrycy z ceramiki

Jednym z najnowszych, najbardziej zaawansowanych technologicznie materiałów jest porowata ceramika, która zaczyna znajdować zastosowanie w wielu różnych gałęziach przemysłu, od filtracji, absorpcji, katalizatorów i podpór katalizatora po lekkie elementy konstrukcyjne. W ciągu ostatniej dekady, ogromne zasoby zostały poświęcone na badania nad innowacyjnymi technologiami przetwarzania porowatej ceramiki, co w efekcie dało lepszą kontrolę jej struktury i znaczną poprawę właściwości użytkowych.

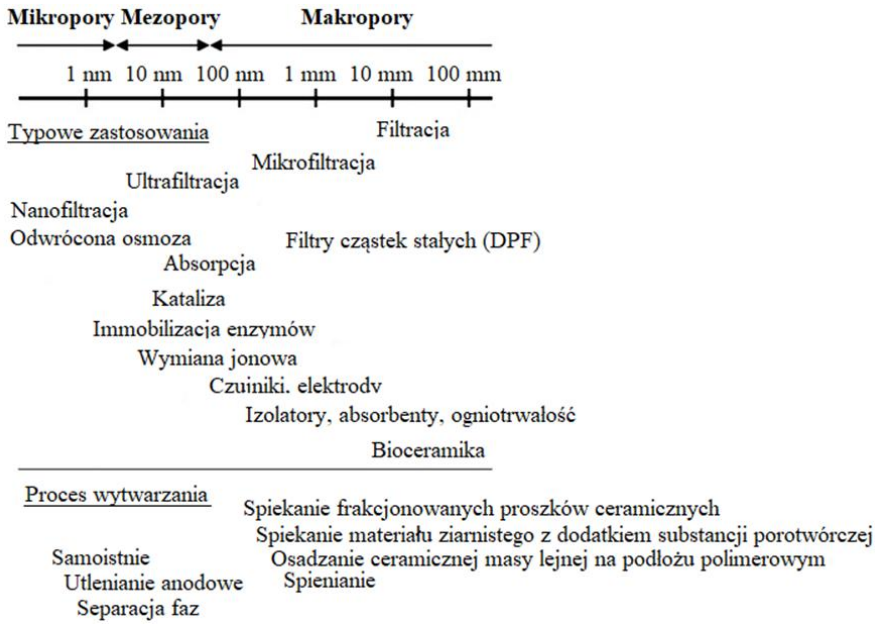
Porowatość pian ceramicznych waha się w zakresie od 20% do ponad 90%, a duży wpływ na wytrzymałość mechaniczną tworzywa ma skład mineralogiczny spieku. Ważny jest również rodzaj i wielkość ziarna, rodzaj i ilość dodanego spoiwa oraz parametry spiekania.

Ceramiczne materiały porowate można podzielić na trzy grupy w zależności od średnicy porów d: makroporowate ($d > 50$ nm), mezoporowate (50 nm $> d > 2$ nm) i mikroporowate ($d < 2$ nm). Na rysunku 13. przedstawiono tę klasyfikację wraz z typowymi zastosowaniami i procesami produkcyjnymi specyficznymi dla danych średnic porów. Jednym z najbardziej reprezentatywnych zastosowań materiałów porowatych jest filtracja [27].

Proces filtracji może być klasyfikowany w zależności od średnicy porów d i odcięcia masy cząsteczkowej (MWCO) na [28]:

- filtrację (zazwyczaj $d > 10$ μm);
- mikrofiltrację (10 μm $> d > 100$nm);
- ultrafiltracjację (100 nm $> d > 1$ nm, MWCO = 103–106);
- nanofiltrację ($d \approx 1$–2 nm, MWCO = 200–103);
- odwróconą osmozę ($d < 1$ nm, MWCO≈100).

Gdy wielkość porów jest duża, jak dla filtracji i mikrofiltracji, separacja odbywa się głównie poprzez efekt przesiewania. W ultrafiltracji, nanofilacji i odwrotnej osmozie, gdzie wielkość porów jest mała, przepuszczalność płynu zależy od powinowactwa rozpuszczalnika i substancji rozpuszczonej.

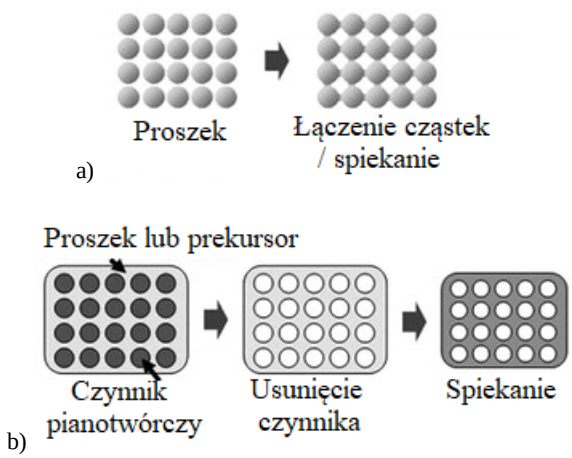


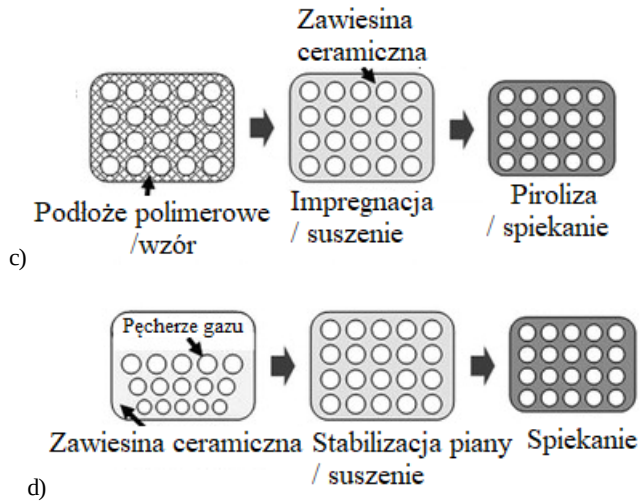
Rysunek 14. Klasyfikacja materiałów porowatych według wielkości porów i odpowiadających im typowych zastosowań i procesów produkcyjnych [27]

Jak widać na rysunku 14. podstawowymi procesami wytwarzania makroporowatej ceramiki są:

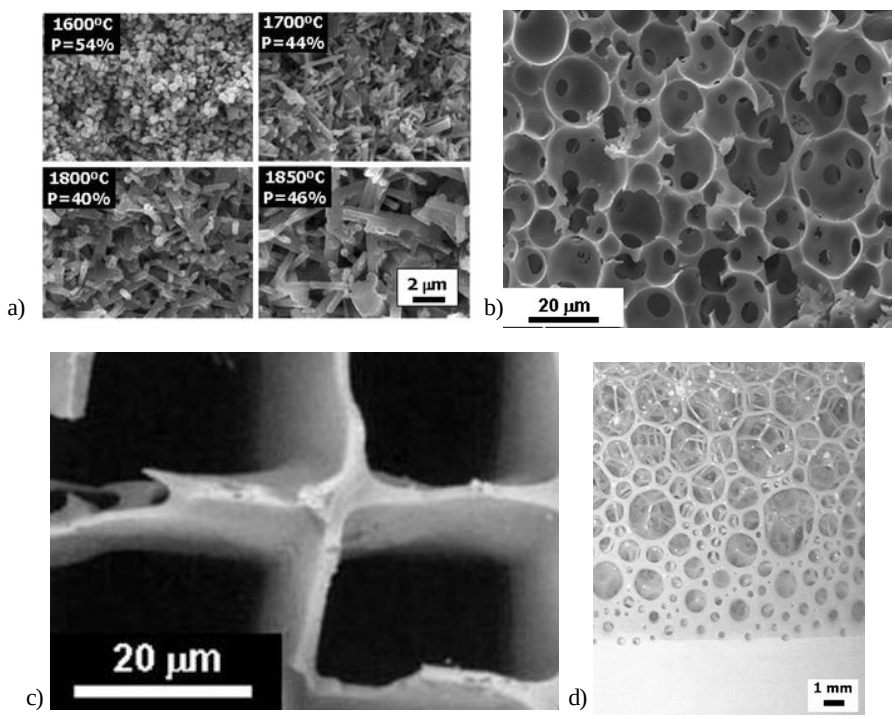
- spiekanie frakcjonowanych proszków ceramicznych z dodatkiem (lub bez) spoiwa;
- spiekanie materiału ziarnistego z dodatkiem substancji porotwórczej;
- osadzanie ceramicznej masy lejnej na podłożu polimerowym;
- spienianie.

Procesy te schematycznie przedstawiono na rysunku 15., natomiast przykładowe struktury otrzymane tymi metodami – na rysunku 16.





Rysunek 15. Schematy metod wytwarzania ceramiki porowatej: a) spiekanie frakcjonowanych proszków ceramicznych z dodatkiem (lub bez) spoiwa, b) spiekanie materiału ziarnistego z dodatkiem substancji porotwórczej, c) osadzanie ceramicznej masy lejnej na podłożu polimerowym, d) spienianie [27]



Rysunek 16. Struktura ceramiki porowatej uzyskana metodą: a) spiekania frakcjonowanych proszków ceramicznych [29], b) spiekania materiału ziarnistego z dodatkiem substancji porotwórczej [30], c) osadzania ceramicznej masy lejnej na podłożu polimerowym [31], d) spieniania [32]

Filtry ceramiczne są obecnie szeroko stosowane w silnikach wysokoprężnych w celu wychwytywania cząstek stałych w strumieniu gazów spalinowych, są to tzw. zwane filtry cząstek stałych (DPF) [33].

Ostatnio duże zainteresowanie wzbudza porowata bioceramika z otwartymi strukturami porów do zastosowań bioimplantacyjnych, w tym regeneracji kości [34]. Komórki kostne są impregnowane na otwarte pory i rosną na ich biokompatybilnych ścianach.

Innym ciekawym i istotnym zastosowaniem ceramiki porowatej są ogniotrwałe cegły pieców wysokotemperaturowych ze względu na ich niską przewodność cieplną i wysoką odporność na wstrząsy termiczne [35]. Z drugiej strony ceramika porowata o właściwościach przewodzących (np. cyrkonu lub węglika krzemu) jest stosowna w wymiennikach ciepła i grzejnikach [36].

2.4. Materiały porowate z druku 3D

Wytwarzanie elementów w technologii druku 3D staje się coraz bardziej popularne z uwagi na coraz większą dostępność tej technologii oraz obniżenie kosztów wytwarzania. Wykorzystuje się ją także jako metodę szybkiego prototypowania – rapid prototyping. Drukarka 3D jest urządzeniem, które potrafi wytworzyć trójwymiarowy rzeczywisty komponent na podstawie danych numerycznych. Model jest tworzony w oprogramowaniu komputerowym CAD (*Computer Aided Design*), a następnie przesyłany do drukarki, która stopniowo, warstwa po warstwie, buduje zadany obiekt.

Rozwój technik druku 3D jest konsekwencją wielu zalet, jakie ona posiada. Są to m.in. [37]:

- możliwość wytworzenia obiektów o niestandardowych kształtach, której nie dałoby się wytworzyć konwencyjnymi sposobami np. kula pusta w środku;
- wytworzenie modelu w jednym kroku. Zastępuje w ten sposób kilka tradycyjnych procesów technologicznych;
- relatywnie niska cena wytworzenia elementów o jednostkowej produkcji;
- wytwarzanie może odbywać się za pomocą różnorodnych materiałów (np. beton, czekolada).

Do najpopularniejszych metod szybkiego prototypowania (ang. *Rapid Prototyping* – RP) można zaliczyć:

- *Fused Deposition Modeling* – FDM – nakładanie stopionego materiału;
- *Selective Laser Sintering* – SLS – wiązka lasera o dużej mocy, spieka małe cząstki z tworzywa sztucznego;
- stereolitografia – SLA – utwardzanie fotopolimeru za pomocą lasera UV.

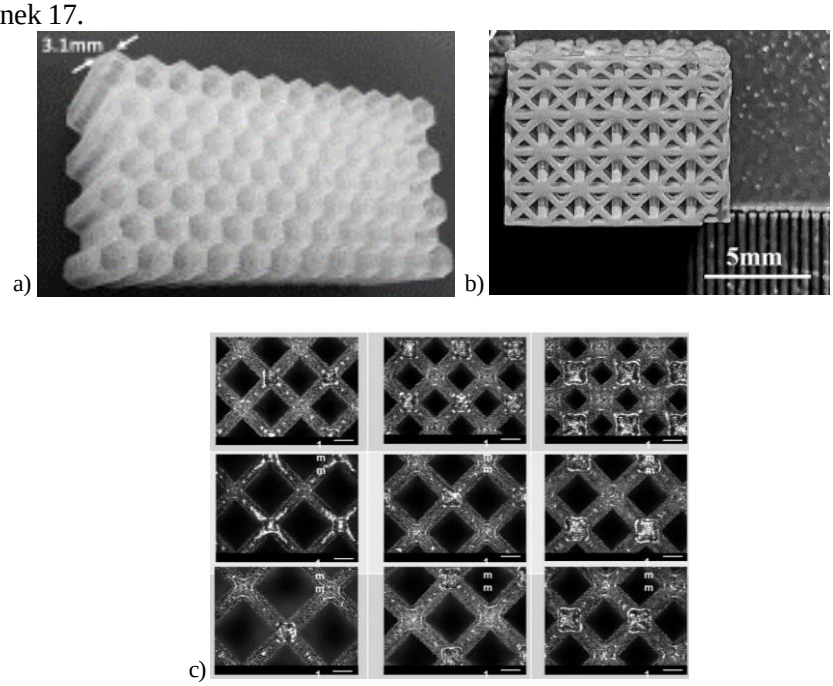
Metoda FDM opiera się o technologię nakładania warstwa po warstwie stopionego tworzywa na podłożu przy pomocy ruchomej głowicy dozującej. Tworzywo sztuczne w postaci filamentu (cienkiego pręcika) dostaje się do głowicy, która go roztapia i poruszając się w płaszczyźnie trójwymiarowej wyciska go na podłoże. Po zetknięciu z chłodniejszym otoczeniem tworzywo schładza się i zestala, tworząc model, w zależności od modelu drukarki może się poruszać również podłoże. Filament może mieć średnicę kilku milimetrów i dla łatwego podawania nawinięty on jest na szpulkę. Elementy wykonane tą technologią nie są najlepszej jakości. Tolerancja wymiarowa i estetyka nie są na najwyższym poziomie. Jednak rekompensują to dobre właściwości mechaniczne, prostota działania i niska cena drukarek [38].

Technologia SLS jest jedną z pierwszych metod szybkiego prototypowania. Obecnie jest najczęściej używana do produkcji krótkoseryjnej. Sposób działania drukarek SLS jest bardzo zbliżony do metody SLA, z tą różnicą, że zamiast płynu wykorzystywany

tu jest proszek. Stosowane mogą być proszki nylonowe, ceramiczne, szklane, a także metalowe. Model buduje się przez spiekanie laserem proszku do powstania kolejnych warstw. Za równomierne rozprowadzenie proszku odpowiada specjalna rolka. Wydrukowane elementy charakteryzują się dobrą dokładnością, jednak mogą być porowate, co powoduje konieczność zastosowania utwardzaczy. Obecnie najpopularniejszymi materiałami stosowanymi w tej technologii są poliamidy (PA), ale także polistyren (PS) oraz mieszaniny poliamidów z wypełniaczami [39].

Drukowanie metodą SLA odbywa się na ruchomej podstawie, która zanurzona jest w ciekłej żywicy. Istnieją drukarki z podstawą zamocowaną nad i pod poziomem żywicy. Naświetlanie laserem UV powoduje utwardzenie warstwy, której grubość zależy od mocy lasera i parametrów zestalenia żywicy. Czynność zamoczenia i naświetlania jest powtarzana, aż do osiągnięcia pożądanego kształtu obiektu. Modele wykonane w tej technologii cechują się wysoką gładkością powierzchni i tolerancją wymiarową. Wykonywany model jest bardzo często wyposażony w podpory z tego samego materiału i tworzone w taki sam sposób jak model docelowy. Służą one do zabezpieczenia krawędzi modelu przed ich zdeformowaniem wynikłym z jego ciężaru. Mają one kształt cienkich pionowych pręcików zwężających się przed powierzchnią modelu. Podpory po procesie druku usuwa się mechanicznie. Po zakończeniu drukowania obiekt zdejmuje się z podstawy, a następnie oczyszcza się w alkoholu izopropylowym. Dla otrzymania najlepszych właściwości wytrzymałościowych wydrukowane elementy wygrzewa się w podwyższonej temperaturze oraz wystawia na działanie światła UV.

Rozwój technologii druku 3D, ich dokładność oraz wskazane powyżej zalety, sprawiają, że jest ona wykorzystywana do wytwarzania struktur porowatych, komórkowych zarówno przez naukowców, jak i komercyjnie. Przykłady takich wydruków pokazano na rysunek 17.



Rysunek 17. Przykłady struktur porowatych wydrukowanych na drukarkach 3D:
a) SLS [40], b) SLA [41], c) SLM [42]

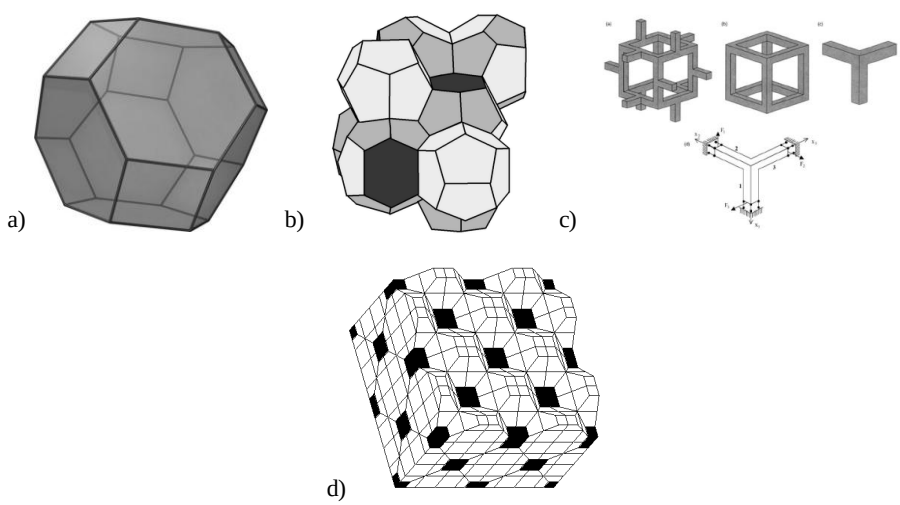
Należy również wspomnieć, że obecnie możliwe jest również drukowanie pian ceramicznych, np. metodą FDM lub SLA [43]. Jednym z przykładów materiałów fotoutwardzalnych do druku 3D metodą SLA jest żywica Ceramic Resin firmy Formlabs. Jest to materiał kompozytowy. Wydrukowana próbka musi zostać wypalona w piecu, gdzie wytapiana jest matryca polimerowa, a w próbce pozostaje ceramika krzemionkowa. Dzięki drukowaniu 3D z żywicy ceramicznej można wykonać złożone obiekty ceramiczne, których nie można byłoby odlewać. Po wypaleniu ceramika drukowana jest odporna na temperaturę powyżej 1000°C [44].

3. Projektowanie materiałów pianowych

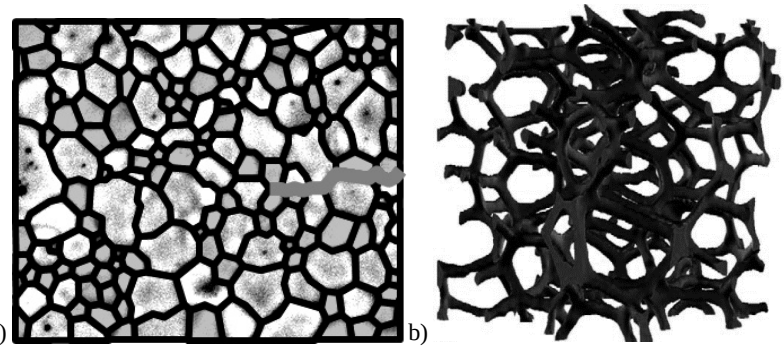
Materiały porowate i piany są obecnie przedmiotem intensywnych badań, zarówno eksperymentalnych, jak i numerycznych. W dziedzinie badań numerycznych spotyka się różnorodne podejścia do zagadnienia modelowania ich struktury. Spotyka się różne modele materiałów/ośrodków ogólnie nazywane „wielokomórkowymi". Są one używane do badań numerycznych ceramik, kompozytów ceramicznych, gruntów, mieszanin, pian, gąbek, itp. Określona geometria może być adaptowana do budowania modeli różnych materiałów [45].

Modelowanie materiałów wiąże się z podejściem wieloskalowym. Aby uzyskać materiał o pożądanych właściwościach jego projektowanie zacząć można od analizy składu chemicznego i budowy cząsteczkowej (poziom nano), następnie rozważać budowę mikrostrukturalną (np. rozkład, wielkość, kształt porów w przypadku materiałów pianowych) oraz właściwości makrostrukturalne. Rozwój metod numerycznych oraz ciągle zwiększająca się moc obliczeniowa komputerów pozwalają na coraz dokładniejszą analizę i projektowanie materiału pod konkretne zastosowania.

Istnieje wiele modeli mikrostrukturalnych materiałów porowatych. Mogą być one oparte na geometrii idealnej (rys. 18) lub rzeczywistej (rys. 19). Te ostatnie są tworzone dzięki wprowadzeniu do badań materiałowych nowoczesnych technik eksperymentalnych, jak tomografia komputerowa czy mikroskopia sił atomowych SEM.



Rysunek 18. Przykłady powtarzalnych struktur komórkowych bazujących na różnych bryłach foremnych [10]: a) czternastościan Kelvina, b) struktura Weaire–Phelana, c) model Gibsona, d) geometria wielokomórkowej wykorzystanej do badania kompozytów ceramicznych [46]



Rysunek 19. Przykłady modeli opartych na analizie struktur rzeczywistych: a) zbudowana na bazie zdjęcia mikroskopowego rzeczywistej struktury [47], b) tomografii komputerowe [48]

Stopień złożoności modeli może być krańcowo różny. Wybór określonego sposobu modelowania zależy w dużym stopniu od posiadanych przez badacza możliwości badań eksperymentalnych, które mogą posłużyć w początkowej fazie prac jako źródło danych do tworzenia modeli numerycznych, a w dalszej – do weryfikacji poprawności zastosowanego modelu.

Nowoczesne narzędzia, takie jak stosowana w przemyśle i nauce metoda elementów skończonych (MES), pozwalają na skuteczną predykcję skutków oddziaływania wymuszeń dynamicznych, w tym szybkozmiennych, jak na przykład fali uderzeniowej pochodzącej od detonacji materiału wybuchowego na struktury ochronne. W celu zapewnienia wysokiej jakości odwzorowania rzeczywistych warunków, czy właściwości stosowanych materiałów, dane do sporządzenia modelu powinny być jak najbardziej wiarygodne. Jednocześnie, prowadzenie czasochłonnych i kosztownych badań eksperymentalnych, które na celu mają wyłonienie niezbędnych parametrów, opóźnia często proces projektowania konstrukcji.

4. Podsumowanie

W rozdziale przedstawiono szeroką gamę nowoczesnych materiałów konstrukcyjnych, jakimi są piany. Rozwój technik wytwarzania pozwolił człowiekowi na wytwarzanie spienionych struktur o kontrolowanych właściwościach z metali, ceramiki, tworzyw sztucznych. Współcześnie można znaleźć też zastosowania tych materiałów w biotechnologii i medycynie.

Największe nadzieje na dalszy rozwój tych materiałów daje z jednej strony rozwój metod obliczeniowych i eksperymentalnych, które pozwalają na wieloskalowe projektowanie pian – od składu chemicznego, przez budowę mikrostrukturalną po właściwości globalne. Z drugiej strony gwałtowny rozwój metod wytwarzania, w tym druku 3D, pozwala na dokładne odwzorowanie zaprojektowanej na drodze obliczeń struktury. Inżynierowie mają możliwość zaprojektowania konkretnego materiału o cechach ściśle dostosowanych do zastosowania. Takie podejście znajduje zastosowanie w przemyśle samochodowym czy chemicznym (np. filtracja), ale także w budownictwie i przemyśle wojskowym (ochrona przed falą wybuchu). Takie podejście ma też aspekt ochrony środowiska – nowe technologie wytwarzana są niskoodpadowe, a materiały mogą być tak zaprojektowane, aby umożliwić ich powtórne wykorzystanie czy recykling.

Literatura

1. www.123rf.com, dostęp 10.03.2021.
2. www.ecospa.pl, dostęp 11.03.2021.
3. www.businessmagnet.co.uk, dostęp 28.02.2021.
4. www.asiapack.pl, dostęp 12.03.2021.
5. Evans A.G. (red.), *Ultralight Metal Structures*. Division of Applied Sciences, Harvard University, USA, Mass, Cambridge 1999.
6. Banhart J., Seeliger H.-W., *Aluminium Foam Sandwich Panels: Manufacture, Metallurgy and Applications*, Advanced Engineering Materials, 10, 9, 2008, s. 793-802.
7. Gale W.F., Totemeier T.C. (red.), *Smithells Metals Reference Book* (8th Edition), Elsevier Ltd., 2004
8. Goodall R., *Porous metals: foams and sponges*, Advances in Powder Metallurgy, 2013, s. 273-307.
9. Neikov O.D., Naboychenko S.S., Murashova I.V., Gopienko V.G, Lotsko D.V. (red.), *Handbook of Non-Ferrous Metal Powders*, Elsevier Ltd., 2009, s. 212-223.
10. Miedzińska D., *Numeryczne badania energochłonności materiałów pianowych z uwzględnieniem ich mikrostruktury*, Wojskowa Akademia Techniczna, 2015.
11. HexWeb® Rigicell™, *Corrosion Resistant Aluminum Corrugated Honeycomb*, Product Data, 2019.
12. McFarland R., *The development of metal honeycomb energy-absorbing elements*, Technical report no. 32-639, California, 1964.
13. Szycher M., Szycher's *Handbook of Polyurethanes*, 2nd ed., CRC Press, USA, New York 2006.
14. Ionescu M. *Chemistry and Technology of Polyols for Polyurethanes*, Rapra Technology Limited, UK, Shawbury 2005.
15. *Polyurethane Global Market Size Forecast 2021*, www.statista.com, dostęp 13.02.2021.
16. Palm E., Svensson Myrin E., *Mapping the Plastics System and Its Sustainability Challenges*, IMES/EESS Report No. 108, Lund University, Lund, Sweden, 2018.
17. *Polyurethane Production, Pricing and Market Demand*, www.plasticsinsight.com, dostęp 13.02.2021.
18. Gama N.V., Ferreira A., Barros-Timmons A., *Polyurethane Foams: Past, Present, and Future*, Materials, 11, 1841, 2018.
19. Okabe A, Boots B., Sugihara K., Chiu S.N., *Spatial Tessellations – Concepts and Applications of Voronoi Diagrams* (2nd edition), John Wiley, USA, 2000.
20. Ashida K., *Polyurethane and Related Foams Chemistry and Technology*, Taylor & Francis Group, Boca Raton, FL, USA, 2007.
21. www.iqsdirectory.com/articles/foam/polyurethane-foam.html, dostęp 22.03.2021.
22. Mills N.J., *The high strain mechanical response of the wet Kelvin model for open-cell foams*, International Journal of Solids and Structures, 44, 1, 2007, s. 51-65.
23. Fridrihsone A., Stirna U., Lazdiņa B., Misāne M., Vilsone D., *Characterization of polyurethane networks structure and properties based on rapeseed oil derived polyol*, European Polymer Journal, 49, 2013, s. 1204-1214.
24. Philipp C., Eschig S., *Waterborne polyurethane wood coatings based on rapeseed fatty acid methyl esters*, Progress in Organic Coatings, 74, 2012, s. 705-711.
25. Kong X., Liu G., Curtis J.M., *Novel polyurethane produced from canola oil based poly(ether ester) polyols: Synthesis, characterization and properties*, European Polymer Journal, 48, 2012, s. 2097-2106.
26. Yang W., Dong Q., Liu S., Xie H., Liu L., Li J., *Recycling and disposal methods for polyurethane foam wastes*, Procedia Environmental Sciences, 16, 2012, s. 167-175.

27. Ohji T., Fukushima M., *Macro-porous ceramics: processing and properties,* International Materials Reviews, 57, 2, 2012, s. 115-131.
28. Wang A., *Fabrication of $La_{0.8}Sr_{0.2}FeO_{3-\delta}$ Porous Ceramics Using a Phase-Leaching Method*, Journal of the American Ceramic Society, 91, 12, s. 4118-4120, 2008.
29. Yang J.F., Deng Z.Y., Ohji T., *Fabrication and characterisation of porous silicon nitride ceramics using Yb_2O_3 as sintering additive*, Journal of the European Ceramic Society, 23, 2, 2003, s. 371-378.
30. Barg S., Soltmann C., Andrade M., Koch D., Grathwohl G., *Cellular Ceramics by Direct Foaming of Emulsified Ceramic Powder Suspensions*, Journal of the American Ceramic Society, 91, 9, 2008, s. 2823-2829.
31. Singh M., Yee B.-M., *Reactive processing of environmentally conscious, biomorphic ceramics from natural wood precursors*, Journal of the European Ceramic Society, 24, 2004, s. 209-217.
32. Pyzik A.J., Li C.G., *New Design of a Ceramic Filter for Diesel Emission Control Applications*, International Journal of Applied Ceramic Technology, 2, 6, 2005, s. 440-451.
33. Le Guehennec L., Layrolle P., Daculsi G., *A review of bioceramics and fibrin sealant,* European Cells & Materials Journal, 8, 2004, s. 1-11.
34. Litovsky E.Y., Shapiro M., *Gas Pressure and Temperature Dependences of Thermal Conductivity of Porous Ceramic Materials: 1. Refractories and Ceramics with Porosity Below 30-Percent*, Journal of the American Ceramic Society, 75, 12, 1992, s. 3425-3439.
35. Scheffler M., Colombo P. (red.), *Cellular ceramics: structure, manufacturing, properties and applications*, Weinheim, Wiley-VCH Verlag GmbH, 2005.
36. Leniowski R., Nitek S., Rzeszutek M. at al., *Modelling of endoprostheses – applications of 3D and 5D printers in medicine,* Acta Bio-Optica et Informatica Medica, Inżynieria Biomedyczna, 22, 4, 2016, s. 226-233.
37. www.swiatdruku3d.pl, dostęp 16.12.2020.
38. www.drukarki3d.pl, dostęp 18.01.2021.
39. Naddeo F., Naddeo A., Cappetti N., *Novel "load adaptive algorithm based" procedure for 3D printing of lattice-based components showing parametric curved micro-beams,* Composites Part B: Engineering, 115, 2017, s. 51-59.
40. Shuai X., Zeng Y., Li P. et al., *Fabrication of fine and complex lattice structure Al2O3 ceramic by digital light processing 3D printing technology*, Journal of Materials Science, 55, 2020, s. 6771-6782.
41. Płatek P., Sienkiewicz J., Janiszewski J., Jiang F., *Investigations on Mechanical Properties of Lattice Structures with Different Values of Relative Density Made from 316L by Selective Laser Melting (SLM)*, Materials, 13, 2204, 2020.
42. Hwa L.C., Rajoo S., Noor A.M., Ahmad N., Uday M.B., *Recent advances in 3D printing of porous ceramics: A review,* Current Opinion in Solid State and Materials Science, 21, 6, 2017, s. 323-347.
43. www.formlabs.com, dostęp 28.03.2021.
44. Ashby M.F., Evans A.G., Fleck N.A. et al., *Metal foams: A Design Guide*, Butterworth-Heinmann, USA, 2000.
45. Szymczyk W., Włodarczyk J., *Influence of Chosen Microstructure Features on Residual Stresses Distribution in FGM Surface Coating System with the Use of FEM Micromechanical Random Models*, Journal of KONES Powertrain and Transport, 14, 2007, s. 123-128.
46. Niezgoda T., Miedzińska D., *Modelling of Random Porosity Foam Microstructures in Aspect of Security and Safety Applications,* The International Symposium on Stochastic Models in Reliability Engineering, Life Science 2010, SMRLO'10, Beer Sheva, Izrael, CD Conference Proceedings, 2010, s. 1-10.
47. Miedzińska D., Niezgoda T., Gieleta R., *Numerical and experimental aluminum foam microstructure testing with the use of computed tomography,* Computational Materials Science, 64, 2012, s. 90-95.

Rozwój nowych technologii materiałowych – materiały porowate i cellulary w przemyśle XXI wieku

Streszczenie

Mimo że materiały porowate, takie jak piany metaliczne, piany polimerowe czy struktury ulowe (cellulary) są znane od kilkudziesięciu lat, rozwój technologii produkcji i metod komputerowych umożliwia postęp technologiczny w ich projektowaniu i wykorzystaniu. Metody obliczeniowe umożliwiają projektowanie materiałów od składu chemicznego, budowy mikrostrukturalnej do makrostrukturalnego obiektu z uwzględnieniem specyficznych wymagań produktu końcowego. Materiały porowate, w tym cellulary, znajdują obecnie coraz szersze zastosowanie – od obiektów ochronnych (np. pancerze pojazdów bojowych) do opakowań. W rozdziale opisane zostaną nowe metody projektowania, produkcji oraz zastosowań tych materiałów, z uwzględnieniem coraz szerszej gamy materiałów wyjściowych, jak ceramika czy materiały wytwarzane na drukarkach 3D.

Słowa kluczowe: materiały porowate, piany metaliczne, ceramika porowata, piana poliuretanowa, druk 3D

Development of New Material Technologies – Porous and Cellular Materials in the Industry of the 21st Century

Abstract

Although porous materials such as metallic foams, polymer foams or cellular structures have been known for several decades, the development of production technology and computer methods enables technological progress in their design and use. Computational methods enable the design of materials from chemical composition, microstructural to macrostructural object, taking into account the specific requirements of the final product. Porous materials, including cellular materials, are now increasingly used – from protective objects (e.g. armor of combat vehicles) to packaging. This chapter will describe new methods of designing, producing and using these materials, taking into account an increasingly wide range of starting materials, such as ceramics or materials produced on 3D printers.

Keywords: porous materials, metallic foams, porous ceramics, polyurethane foam, 3D printing

Barbara Mrzygłód[1], Krzysztof Regulski[2], Andrzej Opaliński[3]

Uczenie maszynowe w materiałoznawstwie

1. Wstęp: 4. paradygmat nauki w 4. rewolucji przemysłowej

Naukę można zdefiniować za pomocą czterech paradygmatów: 1: dowody empiryczne, 2: teoria naukowa, 3: informatyka i 4. paradygmat: nauka o danych [1]. Wkraczając w erę Przemysłu 4.0, wymuszającego automatyczne sterowanie i autonomiczne podejmowanie decyzji w systemach cyberfizycznych, dwa ostatnie paradygmaty nabierają nowego znaczenia. Tak jak nie ma modelowania numerycznego bez obliczeń komputerowych, tak samo nie ma autonomicznego podejmowania decyzji bez uczenia maszynowego. Data Science definiuje sztukę przetwarzania danych i zdobywania wiedzy. Możliwości, jakie daje nam uczenie maszynowe, wydają się dostarczać nowych instrumentów badawczych, które, jak kiedyś mikroskop, mogą dostarczyć całkiem nowych spostrzeżeń wielu naukowców. Nie zawsze wiąże się to (choć jest to najbardziej oczywisty i niezawodny sposób) z pozyskiwaniem nowych danych prosto z procesów przemysłowych. Często dane z już przeprowadzonych badań eksperymentalnych, ale odpowiednio zsyntetyzowane i połączone, mogą dostarczyć nowych wniosków i wyników podczas eksploracji istniejącego archiwum. Nie byłoby to jednak możliwe bez ustrukturyzowanej procedury pozyskiwania tych danych zapewnianej przez reżim data science, a także dalszej interpretacji w pryzmacie wiedzy dziedzinowej.

Wspomaganie decyzji, będące paradygmatem Industry 4.0 dotyczy zarówno procesów wytwarzania, jak i przetwórstwa metali. Rozpatrując koncepcję Przemysłu 4.0 z perspektywy metalurgii, istotna jest charakterystyka tego obszaru. Postrzegając przemysł metalurgiczny przez pryzmat procesów, należy zauważyć, że produkcja jest skomplikowana i złożona z wielu etapów, bardzo duża liczba parametrów procesu determinuje jakość wyrobu, a zjawiska zachodzące w trakcie przetwarzania materiału mają najczęściej charakter nieliniowy i trudny do modelowania z użyciem metod deterministycznych czy numerycznych. Wszędzie tam, gdzie procesy metalurgiczne nie dają się modelować z zastosowaniem narzędzi typu CAx lub jest to zbyt kosztowne obliczeniowo i czasochłonne, można tworzyć modele stochastyczne wspomagające podejmowanie decyzji z użyciem technik eksploracji danych.

Realizowane obecnie projekty z zakresu Industry 4.0 w zakładach związanych z wytwarzaniem produktów metalowych dotyczą systemów wielopoziomowych, gdzie warstwa integracji danych pochodzących z pomiarów i kontroli jakości integrowana jest z warstwą kontroli i monitorowania procesów oraz z warstwą wspomagania decyzji. W takich systemach problemy związane są zarówno z problemem skali, czyli BigData, ale także z integracją tych danych i modeli uczenia maszynowego związanymi z pomiarami jakości. Najczęściej dane procesowe i dane dotyczące wskaźników jakości pochodzą z innych źródeł, o innym próbkowaniu, wskaźnikach referencyjnych i innym sposobie identyfikacji. Wyzwania leżą też w zakresie tworzenia odpowiednich archi-

[1] AGH w Krakowie.
[2] AGH w Krakowie.
[3] AGH w Krakowie.

tektur układów sensorów, zazwyczaj bezprzewodowych, łączących się w sieci IoT, integracji różnych formatów danych, tworzeniu zbiorów uczących, a po zrealizowaniu modeli data-driven, wdrożeniu do procesu i wspomagania decyzji.

Metody uczenia maszynowego stosowane w inżynierii materiałowej nie mogą być realizowane bez odpowiedniej infrastruktury sprzętowo-programowej, zapewniającej wydajność i skuteczność ich działania, oraz bezpieczeństwo danych, w oparciu o które działają. Najczęściej realizowane są one w ramach systemów informatycznych, wdrożonych i zintegrowanych w infrastrukturze środowiska specyficznego dla danej dziedziny przemysłu. W zależności od tego kontekstu, cechy specyficzne takich systemów mogą się różnić, jednak można wyodrębnić jeden uniwersalny model, przedstawiający strukturę tego typu systemów. Składa się on zwykle z trzech głównych warstw. Najwyższa warstwa tego typu systemów to jego interfejsy, pozwalające na wprowadzanie danych do systemu, zarządzanie nim oraz wykorzystywanie rezultatów zwracanych przez metody w nich wykorzystywane. Dane do systemu mogą być wprowadzane w oparciu o graficzny interfejs użytkownika, kiedy to administrator systemu lub ekspert technologiczny wprowadza dane na podstawie przygotowanych wcześniej, dedykowanych zestawach formularzy.

Druga grupa metod nie wymaga bezpośredniego udziału człowieka. Pierwszym sposobem może być pozyskiwanie danych w oparciu o dostarczone arkusze z danymi w formie elektronicznej (np. eksperymentalnymi, wynikami badań). Metody automatyczne wymagają zwykle zastosowania dedykowanych *plugin*-ów (automatycznych bibliotek softwarowych, umożliwiających akwizycję danych z zewnętrznych systemów informatycznych). Trzeci sposób akwizycji danych opiera się na sprzętowych sensorach, zasilających bezpośrednio warstwę bazodanową. Metody automatycznej akwizycji danych korzystają zwykle z funkcjonalności realizowanej już w drugiej warstwie systemu (warstwie logiki biznesowej), jakimi są *parsery* danych surowych, niezbędne do pozyskania oczyszczonych, znormalizowanych i odfiltrowanych danych ze źródeł w oparciu, o które działają. Tego typu mechanizmy zapewniają dodatkowo spójność i kompletność danych oraz monitorują źródła pod kątem ich dostępności. W warstwie logiki biznesowej systemu informatycznego znajdują się także kluczowe komponenty od strony metod uczenia maszynowego. W zależności od zastosowanej w systemie metody mogą znajdować się tam metody trenowania, testowania, walidacji klasyfikatorów, sieci neuronowych czy drzew decyzyjnych, czy innych, takich jak omawiane w tym opracowaniu maszyny wektorów nośnych (SVM – *Support Vector Machine*) czy algorytmy analizy skupień. Komponenty tego poziomu wykorzystują natomiast jako bazę do swojego funkcjonowania struktury i zawartość trzeciej, najniższej warstwy systemów informatycznych tego typu. Jest to warstwa związana z danymi oraz modelami przechowywania informacji. Możemy tam znaleźć zarówno dane surowe pochodzące np. z sensorów czy pomiarów bieżących, dane umieszczone w modelu uniwersalnym, będącym rezultatem działania mechanizmów parsujących czy dane specyficzne dla konkretnych metod uczenia maszynowego, wygenerowane na potrzeby optymalnego i efektywnego procesu przetwarzania tych danych przez te metody. Rezultaty działania tego typu systemów są zwracane dla użytkownika z warstwy logiki biznesowej (jako rezultaty działania metod uczenia maszynowego – *Machine Learning (ML)*) do najwyższej warstwy systemu (warstwy interfejsu), gdzie prezentowane są zwykle w graficznym interfejsie użytkownika (dla użytkowników systemu) lub w postaci

cyfrowej (np. w formie *webserwis*-ów) do dalszego, automatycznego wykorzystania w bardziej rozbudowanych systemach produkcyjnych (działających np. w oparciu o model Industry 4.0). Tego typu ogólny schemat architektury systemu może być rozbudowywany o dodatkowe komponenty, umożliwiający ich integrację z dodatkowymi, specyficznymi źródłami danych, jak również w celu ich dalszego udostępnienia w szerszym procesie przetwarzania informacji.

Jak widać z powyższych rozważań, 4. paradygmat nauki w 4. rewolucji przemysłowej to przede wszystkim dane: pozyskiwanie, przetwarzanie, przechowywanie, *data-mining*, *data-modelling* i w konsekwencji wdrażanie modeli *data-driven*. Niniejsze opracowanie koncentruje się przede wszystkim jednak na tworzeniu modeli i uczeniu maszynowym, a także na późniejszej formalizacji pozyskanej wiedzy w celu późniejszego wykorzystania i re-użytkowania.

2. Metody

Zazwyczaj literatura dotycząca metod uczenia maszynowego wymienia metody w podziale nadzorowanym i nienadzorowanym [2-4]. Uczenie nadzorowane umożliwia przewidywanie odpowiednio zmiennych jakościowych (klasyfikacja) lub ilościowych (regresja). Taka sytuacja ma miejsce, gdy mamy wyniki (odpowiedzi, pomiary) zmiennej zależnej i jesteśmy w stanie sformułować odpowiednie hipotezy. W materiałoznawstwie najczęstszym przykładem nadzorowanego uczenia się jest przewidywanie właściwości materiałów w zależności od ich specyfiki w postaci mikrostruktury lub składu chemicznego. Gdy szukamy podobnych obiektów, grup sygnałów lub obszarów o podobnej gęstości, mówimy o metodach uczenia nienadzorowanego – takie przykłady można znaleźć w analizie morfologicznej, w grupowaniu materiałów o podobnej charakterystyce czy w rozpoznawaniu składu fazowego mikrostruktury. Przykłady takich badań zostaną omówione w kolejnych rozdziałach. W szczególnych przypadkach jednak metody z różnych grup są zwykle stosowane jako rozwiązania kombinowane lub hybrydowe. Ponadto jako odrębną grupę stosowanych metod sztucznej inteligencji opiszemy różne formalizmy reprezentacji wiedzy, które mają różne właściwości w zakresie ich dalszego wykorzystania i ponownego wykorzystania. Modele w postaci regułowej (symbolicznej) zostały wykorzystane jako wynik algorytmów indukcji reguł, a także modele wnioskowania oparte na logice rozmytej lub modele w postaci subsymbolicznej modeli neuronowych lub neuro-rozmytych, na ontologiach kończąc.

3. Materiały – Procesy – Formalizacja wiedzy

W rozdziale przyjęto kolejność sekwencji badawczych od materiałów (laboratorium) poprzez procesy (warunki częściowo kontrolowane) aż do realizacji modeli w postaci modeli wnioskowania. Każdy blok badań ma inne uwarunkowania i często punkt widzenia analityka jest inny, podobnie jak cel modelowania. Niniejszy rozdział stanowi podsumowanie wieloletnich prac autorów nad zastosowaniem metod uczenia maszynowego, eksploracji danych i technik sztucznej inteligencji w pozyskiwaniu i formalizacji wiedzy dla potrzeb problemów inżynierii materiałowej i metalurgii.

Algorytmiczne pozyskiwanie wiedzy polega na tworzeniu modeli na podstawie danych (*Data-Driven Models*). Modele te służą do odzwierciedlenia zależności parametrów, opisu ilościowego ich wzajemnego wpływu. Określona na podstawie wiedzy eksperta zmienna decyzyjna (zależna) opisana zostaje poprzez uchwycenie

reguł zmienności tej cechy w zależności od zmian zachodzących w pozostałych parametrach. Zależności stochastyczne mogą zostać odzwierciedlone w postaci modeli regresji, w postaci reguł 'IF-THEN' lub w postaci sub-symbolicznej jak w przypadku sztucznych sieci neuronowych. W prowadzonych badaniach wykorzystano w szczególności algorytmy:

- indukcji drzew decyzyjnych: CART, CHAID, losowy las (*Random Forest),* drzewa wzmacniane (*Boosted Trees*);
- analizy skupień (*k-means*);
- teorii zbiorów przybliżonych (*Rough Sets*);
- uczenie maszynowe (*Machine Learning*, w tym: metodę wektorów nośnych – *Support Vector Machine* SVM, sztuczne sieci neuronowe – *Artificial Neural Networks* ANN).

Pozyskane modele mogą zostać wykorzystane do wnioskowania w systemach wspomagania decyzji. Stanowią one modele wnioskowania w kategoriach systemów zarządzania wiedzą. Różnią się między sobą strukturą, często też precyzją, jednak wzajemnie się uzupełniają, zapewniając użytkownikowi końcowemu zróżnicowaną użyteczność. Omówienie ich funkcjonalności i korzyści stosowania poszczególnych metod czytelnik znajdzie w wymienionych publikacjach [5-7]. Ramą, która zamyka badania związane z ekstrakcją wiedzy z danych, jest formalizacja tej wiedzy.

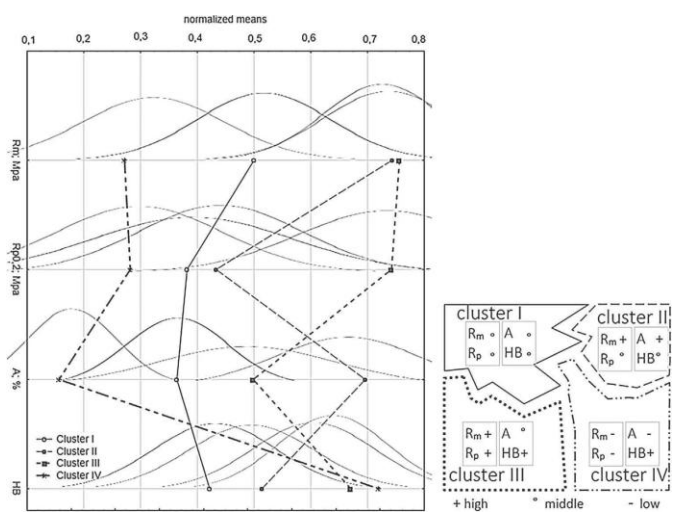
3.1. Materiały

W zakresie badań materiałowych autorzy zrealizowali szereg badań ukazujących przykłady aproksymacji właściwości na podstawie: składu chemicznego, parametrów przetwarzania czy mikrostruktury oraz badania związane z identyfikacją składników stopowych i klasyfikacją materiałów. Zastosowane narzędzia to: (1) narzędzia statystyczne, regresja, (2) odcinkowe modele regresji, (3) MARSplines, (4) drzewa decyzyjne, (5) las losowy, (6) drzewa wzmocnione, (7) sztuczne sieci neuronowe, (8) SVM – maszyna wektorów nośnych, (9) teoria zbiorów przybliżonych, (10) klastrowanie i wreszcie (11) modele neuro-rozmyte [5, 6].

Z perspektywy inżynierii materiałowej najczęściej spotykamy się z problemem przewidywania przyszłych właściwości wyrobu gotowego celem kontroli i optymalizacji procesu. Jakość wyrobu może być określona poprzez właściwości mechaniczne lub pośrednio poprzez szacowanie składu chemicznego i mikrostruktury. Idea Industry 4.0 postuluje monitorowanie procesów poprzez porównywanie bieżących operacji z równolegle obliczanym z modelu wirtualnego oczekiwanym wynikiem w celu utrzymania oczekiwanego poziomu jakości. Innymi słowy, jest to ciągłe przewidywanie rezultatów bieżących operacji. Tworzenie modeli predykcyjnych w tym zakresie opiera się w pierwszym kroku na poszukiwaniu zależności pomiędzy parametrami procesu, a następnie na zastosowaniu algorytmu, który, analizując te zależności, będzie w stanie przewidzieć nieznaną wartość parametrów istotnych dla efektywności procesu (zmiennej zależnej). Badania materiałów niejednokrotnie opierają się na badaniu właściwości mechanicznych i próbie przewidywania tych właściwości w oparciu o skład chemiczny, mikrostrukturę czy parametry obróbki danego materiału.

Przykładem wykorzystania technik sztucznej inteligencji jest pozyskiwanie wiedzy na temat zmian mikrostruktury siluminów po wprowadzeniu pierwiastków stopowych (Cr, Mo, V oraz W) bezpośrednio z wyników badań eksperymentalnych z zastoso-

waniem mieszanych technik *Data Mining*. W ramach badań opracowano procedurę postępowania w analizie małych zbiorów danych eksperymentalnych dla modeli wieloetapowych, wielowartościowych o wielu zmiennych zależnych (*Multi-stage, Multi-veriate, Multi-variable*). Stworzenie procedury głosowania w oparciu o wyniki drzew klasyfikacyjnych i analizę skupień umożliwiło na podstawie badań eksperymentalnych na stworzenie modeli drzew decyzyjnych, których jakość poprawiono z wykorzystaniem dyskretyzacji. Zastosowano analizę skupień, by zniwelować wielowymiarowość wektora wyjść (rozwiązanie zagadnienia wielu zmiennych zależnych – *Multi-variable*). Wykazano skuteczność zastosowania dyskretyzacji zarówno zmiennych zależnych, jak i predyktorów. W ten sposób powstała wieloetapowa procedura klasyfikacji z zastosowaniem dyskretyzacji i analizy skupień. Badania wykazały również, że przeprowadzone postępowanie wymaga na każdym etapie decyzji analityka, ciągłych poszukiwań wśród dostępnych narzędzi, elastycznego dobierania modeli i wnioskowania na bardziej wyrafinowanym poziomie, niż zwykłe techniki *Machine Learning*. Przywoływane przykłady publikacji autorów prezentują, w jaki sposób kształtuje się balans pomiędzy wygodą użytkowania a skutecznością w przypadku pozyskiwania wiedzy z małych zbiorów. Badania wykazały, że drzewa decyzyjne są wygodnym narzędziem odkrywania wiedzy i dobrze radzą sobie z problemami silnie nieliniowymi, a wprowadzenie dyskretyzacji poprawia ich działanie. Zastosowanie metod analizy skupień umożliwiło też wyciąganie bardziej ogólnych wniosków, przez co udowodniono tezę, że właściwa granulacja informacji pozwala znaleźć wzorce nawet w małych zbiorach danych.



Rysunek 1. Diagram rozkładów charakterystyk dla poszczególnych klastrów oraz uproszczony diagram charakterystyk dla wyników analizy skupień [5]

Opracowane reguły dotyczące wpływu poszczególnych dodatków stopowych i mieszanek na właściwości siluminów stanowią bazę wiedzy, która umożliwia wspomaganie decyzji i podniesienie jakości i skrócenie czasu badań materiałowych [5]. Badania wykazały, że przewidywanie właściwości na podstawie składu chemicznego z eksperymentalnej bazy danych (wprost) może zostać istotnie ulepszone dzięki zastosowaniu analizy skupień do segmentacji przestrzeni wyników ($R^2 = 0{,}986$).

Zastosowanie teorii zbiorów przybliżonych w analizie danych z badań wpływu obróbki cieplnej na właściwości eksperymentalnego stopu brązu jest kolejnym przykładem wykorzystania technik uczenia maszynowego [6]. Mechanizm wnioskowania wspomagający dobór odpowiednich parametrów obróbki cieplnej i modyfikatorów w celu uzyskania pożądanych właściwości brązu oparty został na indukcji reguł algorytmami wykorzystującymi teorię zbiorów przybliżonych, jak również drzew decyzyjnych oraz analizie eksploracyjnej wyników badań, w tym analizie statystycznej oraz analizie wariancji (ANOVA). Zastosowanie narzędzi, jakimi są zbiory przybliżone, pozwoliło rozwiązać problem granularności informacji, niepewności i niekompletności danych w warunkach sygnałów dyskretnych i punktowych.

Zastosowanie narzędzi uczenia maszynowego znalazło też miejsce w analizie porównawczej właściwości materiałów w szczególności żeliwa ADI oraz żeliwa sferoidalnego z węglikami (*nodular cast iron with carbides*). Zastosowano maszynę wektorów nośnych (*Support Vector Machine – SVM*), sztuczne sieci neuronowe (*Artificial Neural Networks – ANN*), klasyfikator minimalno-odległościowy (*kNN*) i drzewa klasyfikacyjne (CART oraz CHAID) do rozwiązania problemu doboru materiału do produkcji wyrobu o zadanych właściwości mechanicznych. Modele klasyfikatorów pozwoliły na wspomaganie decyzji technologicznych na etapie projektowania wyrobów odlewniczych. Przeprowadzono dwutorowy przebieg analizy. Z jednej strony badany był problem klasyfikacyjny na podstawie surowych danych na temat właściwości, a z drugiej strony usprawnienie tej metody poprzez wykorzystanie wstępnej analizy skupień na bazie właściwości, czyniąc klasyfikację bardziej zgrubną, ale jednocześnie bezbłędną. Uzyskane w ten sposób rozwiązanie łączy zalety wymienionych wcześniej algorytmów. Zaproponowana metodyka analizy porównawczej zapewnia wygodną ścieżkę eksploracyjną dla analiz rozszerzonych o nowe materiały, których właściwości konstrukcyjne można zestawiać, tworząc w ten sposób wachlarz możliwości dla technologa czy konstruktora.

Rozwinięciem tych prac były badania nad opracowaniem metodyki eksploracyjnej analizy danych pochodzących z badań mikrostrukturalnych żeliwa wermikularnego w celu pozyskania wiedzy na temat czynników sprzyjających powstawaniu ausferrytu [6]. Opracowano reguły umożliwiające szacowanie zawartości ausferrytu na podstawie składu chemicznego. Opracowano modele z wykorzystaniem: (1) narzędzi statystycznych, regresji, (2) regresyjnych drzew wzmacnianych (*boosted trees*), (3) losowego lasu (*random forest*), (4) sztucznych sieci neuronowych, (5) modeli regresji składanej (*piecewise regression models*), (6) algorytmu CART, (7) MARSplines. Procedura etapowego budowania modeli regresyjnych na ograniczanych iteracyjnie zbiorach umożliwiła z jednej strony poprawę precyzji prognozowania, a z drugiej pozyskanie większej wiedzy na temat powstawania ausferrytu.

Wraz z procesem filtrowania danych zmieniła się precyzja modeli. Po usunięciu przypadków stosunkowo prostych do nauczenia (bez ausferrytu) modele pogorszyły współczynniki dopasowania, ale jednocześnie precyzja w zakresie przewidywania faktycznych zawartości się poprawiła. W przypadku drzew wzmacnianych można wyraźnie zauważyć zmianę na skutek redukcji zbioru danych. Na pełnym zbiorze algorytm wyznaczył ponad 800 drzew, uzyskując błąd MSE na poziomie 30 jednostek, natomiast na zbiorze zredukowanym wystarczyło ok. 320 drzew, jednak błąd MSE zwiększył się prawie dwukrotnie. Wielokrotne badanie istotności wpływu poszczegól-

nych czynników w różnych modelach regresyjnych pozwoliło wytypować najważniejsze zmienne mające wpływ na zawartość ausferrytu w różnych obszarach zmienności parametrów. Domknięciem tych badań było zbudowanie modelu aproksymacji mikrostruktury żeliwa wermikularnego (*CGI – Compacted Graphite Iron*) na podstawie składu chemicznego i grubości ścianki odlewów, a tym samym predykcja udziału poszczególnych składników osnowy (austenitu, ausferrytu, martenzytu, węglików, perlitu i ferrytu) [7]. Badania są przykładem zastosowania kilku algorytmów uczenia maszynowego w akwizycji wiedzy o mikrostrukturze CGI. Zastosowanie wielu algorytmów, jak również przeprowadzenie analizy dwoma scenariuszami – zarówno klasyfikacji w oparciu o surowe dane, jak i z zastosowaniem wcześniejszej analizy skupień, pozwoliło na porównanie wyników tych procedur – wcześniejsza analiza skupień poprawia jakość klasyfikacji. Ponadto badanie to wykazało, że w sytuacjach, gdy ważniejsza jest precyzja niż możliwość interpretacji, sztuczne sieci neuronowe oraz metoda wektorów nośnych daje znacznie lepsze wyniki niż drzewa decyzyjne.

W pracy [7] przedstawiono również przykłady wykorzystania doświadczalnych danych źródłowych do budowy modeli prognozujących skład objętościowy poszczególnych składników mikrostruktury odlewów na podstawie składu dodatków stopowych – badania te opracowane z wykorzystaniem *Adaptive Neuro Fuzzy Inference System* (ANFIS) [7]. W ramach modelu neuro-fuzzy systemy wnioskowania rozmytego FIS dostarczają schematu wnioskowania, a więc sposobu konstrukcji reguł logicznych, natomiast wyznaczenie parametrów tego modelu odbywa się według algorytmu zaczerpniętego z teorii sieci neuronowych (ANN), w oparciu o analizowane dane. Wzorcowe przypadki uczące stanowią zebrane w jednej bazie danych wyniki obserwacji i pomiarów zawartości poszczególnych składników mikrostruktury żeliwa wermikularnego, w zależności od zawartości poszczególnych dodatków stopowych. Opracowany z wykorzystaniem algorytmu ANFIS system, jako zmienną objaśnianą (zależną) przyjmuje procentową zawartość ausferrytu w mikrostrukturze, w zależności od takich czynników, jak: zawartość Cu [% wag], zawartość Ni [% wag], zawartość Mo [% wag], dla różnej grubości ścianki odlewanego materiału [mm]. W systemie działa 108 reguł rozmytych typu SUGENO, w których przesłanki odwołują się do opracowanych zbiorów rozmytych, a konkluzje stanowią funkcyjne zależności pomiędzy przyjętymi zmiennymi. Opracowany system FIS działa z błędem uczenia na poziomie 5% i błędem testowania na poziomie ok. 9%. Wyniki potwierdziły, że opracowany model może zostać wykorzystany przy doborze składu chemicznego odlewów z żeliwa wermikularnego, charakteryzującego się wysokimi właściwościami wytrzymałościowymi, a co za tym idzie, zmniejszenie nakładów na badania eksperymentalne, które mają na celu wykazanie możliwości wyprodukowania żeliwa o ausferrytycznej mikrostrukturze osnowy bez stosowania obróbki cieplnej, a jedynie za pomocą modyfikacji jego składu chemicznego przy uwzględnieniu grubości ściany odlewu.

Przeprowadzone badania pozwoliły również dostrzec bardzo duży potencjał adaptacyjnego układu wnioskowania neuronowo-rozmytego (ANFIS). Umożliwia on optymalizację (strojenie) parametrów modelu rozmytego na bazie danych pomiarowych. Umożliwia korektę niedokładnych modeli rozmytych sformułowanych przez ekspertów, posiada strukturę i parametry zrozumiałe dla człowieka (zapis wiedzy w postaci lingwistycznych reguł wnioskowania). Umożliwia uogólnienie wiedzy zawartej w zaszumionych danych pomiarowych i przedstawienia jej w zrozumiałej dla człowieka formie.

3.2. Procesy

Zagadnienia związane z tworzeniem modeli procesów w wyniku uczenia maszynowego posiadają schemat działania podobny do podejścia związanego z badaniem właściwości materiałów. W każdym przypadku modele mają na celu z jednej strony umożliwienie prognozowania (dla określonego parametru decyzyjnego), a z drugiej strony ocenę, w jaki sposób poszczególne parametry wejściowe procesu zmieniają wartość na wyjściu (analiza wrażliwości).

Zastosowanie drzew decyzyjnych wspierają np. badania nad procesem produkcji śrub w ramach łańcucha produkcyjnego. Na podstawie danych dotyczących procesu można opracować modele drzew decyzyjnych CART i CHAID [6]. Analiza danych procesowych pozwoliła na ocenę ważności (siły wpływu) zmiennych na klasę jakości śruby. Osiągnięciem było wygenerowanie reguł pozwalających na detekcję możliwości występowania wad produkcyjnych. Wykorzystanie drzew decyzyjnych dało możliwość interpretacji modelu (w całości, jak również we fragmentach – poszczególnych reguł) przez człowieka. To bardzo ważny aspekt eksploracji danych, a niestety niedający się zrealizować z użyciem innych technik. Przejrzystość modeli uzyskanych z pomocą indukcji drzew decyzyjnych, okupiona jest jednak często zmniejszoną precyzją albo przeuczeniem modelu (uwzględnieniem nieistotnych odchyleń). W prowadzonych badaniach wykazano jednak, że da się wykorzystywać to narzędzie w sterowaniu procesami jako wspomaganie decyzji z dokładnością nawet do 94%.

W badaniach dotyczących procesu walcowania pasm ze stali dwufazowych referowanych również zastosowano metodę indukcji drzew decyzyjnych w celu optymalizacji procesu [8]. Zaprezentowano rozwiązanie umożliwiające badanie wpływu poszczególnych parametrów procesu (22 zmienne niezależne w modelu, w tym temperatura wsadu, czasy międzyoperacyjne dla sześciu przepustów, prędkość obrotowa walców, współczynnik wymiany ciepła, etc.) na parametry wyjściowe (temperatura walcowania w poszczególnych przepustach, wielkość ziarna). Wyniki analizy statystycznej, jak również eksploracyjnej porównano z wynikami analizy wrażliwości [9]. Wykazano, że zastosowanie zaproponowanych technik pozwala na osiągnięcie tych samych wyników przy wykorzystaniu zaledwie 1% danych potrzebnych do przeprowadzenia analizy wrażliwości konwencjonalnymi technikami obliczeniowymi. Takie rozwiązanie zapewnia możliwość szybkiej analizy, realnej nawet w czasie rzeczywistym.

Innym przykładem modelowania opartego na danych są badania wilgotności mas formierskich [10]. Opracowano w ramach tych badań sposób poprawy jakości procesu przygotowania mas formierskich w oparciu o predykcję wilgotności. Badania eksploracyjne dały możliwość określenia zależności pomiędzy czynnikami określającymi właściwości mas formierskich. Opracowano dwustopniowe podejście do dopasowania predykcji (zarówno z zastosowaniem indukcji drzew, jak i analizy skupień) i wykazano jego skuteczność w ocenie wilgotności masy formierskiej. Pozyskano w ten sposób wiedzę w postaci reguł przydatną przy tworzeniu baz wiedzy dla systemów wspomagających decyzje w diagnostyce odświeżania mas formierskich.

W badaniach dotyczących analizy trwałości narzędzi kuźniczych wykorzystano logikę rozmytą oraz sztuczne sieci neuronowe do opracowania systemów przewidujących zużycie narzędzi kuźniczych dla wybranego procesu kucia matrycowego [11-15]. Trwałość narzędzi kuźniczych jest istotnym i zarazem trudnym zagadnieniem w aspekcie naukowym oraz ekonomicznym w przypadku zakładów zajmujących się produkcją

odkuwek matrycowych na gorąco. Wynika to ze znacznego udziału kosztów narzędzi wśród wszystkich kosztów liczonych w oparciu o wytworzenie pojedynczej sztuki odkuwki. Opracowane systemy umożliwiają obliczenie ubytku badanego narzędzia dla zadanych wartości jego parametrów eksploatacyjnych (liczba odkuwek, temperatura narzędzia w wybranych punktach, rodzaj nałożonej warstwy ochronnej, ciśnienie i droga tarcia) oraz oszacowanie intensywności występowania typowych mechanizmów niszczenia narzędzi, tj. zmęczenia cieplno-mechanicznego, zużycia mechanicznego, zużycia ściernego i odkształcenia plastycznego. Jako dane źródłowe wykorzystano wyniki wieloletnich badań eksperymentalnych i modelowania numerycznego, które dotyczyły wpływu określonych parametrów procesu kucia na eksploatację i zużycie narzędzia. Badania te obejmowały m.in.: obserwację eksploatacyjną procesu kucia, analizę makroskopową, połączoną ze skanowaniem powierzchni roboczych narzędzi, pomiarami mikrotwardości, analizę mikrostrukturalną oraz modelowanie numeryczne procesu kucia. Zmiennymi wejściowymi (objaśniającymi) były: liczba wykonanych odkuwek, naciski, temperatura na wybranych powierzchniach narzędzia, droga tarcia oraz rodzaj zastosowanej warstwy ochronnej narzędzia. Natomiast zmiennymi wyjściowymi (objaśnianymi) były: ubytek geometryczny materiału narzędzia oraz procentowy udział czterech głównych mechanizmów niszczących. Podczas tych badań rozważano wiele formalnych metod komputerowej reprezentacji wiedzy. Podstawą były wyniki wieloletnich badań przemysłowych zgromadzone w opracowanej bazie danych. Ponieważ wyniki eksperymentów materiałowych są najczęściej niekompletne i mogą być obarczone błędem dlatego analiza została przeprowadzona głównie z wykorzystaniem metod formalnych, które w swojej naturze pozwalają na modelowanie w oparciu o dane niepewne i niepełne. Przeprowadzono badania z wykorzystaniem logiki rozmytej [11] oraz neuro-rozmytego algorytmu ANFIS [12]. Najlepsze dopasowanie i najmniejszy błąd uzyskano jednak przy użyciu sztucznych sieci neuronowych do opracowania systemowej reprezentacji wiedzy [13-15]. W systemie działa dziewięć sieci neuronowych: pięć z nich określa wartość ubytku geometrycznego materiału, dla narzędzi pracujących z warstwami ochronnymi (azotowana, napawana oraz warstwy hybrydowe: Al-CrTiSiN, Cr/CrN, Cr/AlCrTiN), cztery sieci określają intensywność występowania mechanizmów niszczących: zmęczenie cieplno-mechaniczne, zużycie mechaniczne, zużycie ścierne, odkształcenie plastyczne).

3.3. Formalizacja wiedzy

W modelowaniu procesów nierzadko wykorzystuje się bazy danych materiałowych. Do podejmowania decyzji, a także przewidywania przyszłości w procesach produkcyjnych, wykorzystuje się wiedzę składającą się nie tylko z danych, ale przede wszystkim umiejętności technologicznych. W opisanych wcześniej pracach wykazano, że wiedzę o procesach można pozyskiwać również w sposób automatyczny z wykorzystaniem uczenia maszynowego. Aby w pełni wykorzystać potencjał, jaki dają stosowane metody sztucznej inteligencji, należy jednak pozyskaną wiedzę w postaci reguł, modeli czy wzorców zapisać, zachować, czyli skodyfikować. By zachowana wiedza była dodatkowo zrozumiała dla ludzi i przetwarzalna dla maszyn, konieczne jest odpowiednie opisanie tej wiedzy.

Część badań autorów traktuje właśnie o metodach formalizacji wiedzy – zaprezentowano proces odkrywania wiedzy i wnioskowania z użyciem modeli uczenia

maszynowego z zastosowaniem technik semantycznych (ontologii) celem formalizacji wiedzy i kodyfikacji w semantycznych bazach wiedzy. Wykorzystanie technik tworzenia ontologii, czyli modelowania semantycznego, można zaprezentować na przykładzie ontologii do opisu modeli wieloskalowych termomechanicznej obróbki stopów [16]. Model semantyczny, czyli ontologia, umożliwiła walidację modeli wieloskalowych poprzez badanie kompletności zmiennych wejściowych. Opis taki stanowi również narzędzie jednoznacznej komunikacji pomiędzy inżynierami i programistami.

Stworzono system umożliwiający automatyczne katalogowanie (w oparciu o ontologię) i przeszukiwanie merytoryczne dokumentów tekstowych z repozytorium dokumentów dotyczących technologii przetwórstwa metali. Rozwiązanie to umożliwia stworzenie semantycznej bazy wiedzy w zakresie metalurgii i przetwarzania metali z użyciem technik akwizycji i kodyfikacji dających się ponownie wykorzystać artefaktów komputerowych generowanych na podstawie dokumentów. Metodyka obejmuje: algorytm TF-iDF dla określenia siły dyskryminacyjnej słów, ukryte indeksowanie semantyczne LSI celem grupowania dokumentów, tworzenie grup wyrażeń bliskoznacznych oraz integrację z ontologią poprzez przypisanie klas do grup tematycznych. Opracowanie systemu wymusiło proces formalizacji wiedzy, który jest nieodzownym elementem tworzenia systemów wspomagających podejmowanie decyzji. Rozwiązanie polegało na adaptacji klasyfikatorów opartych na teorii zbiorów przybliżonych i analizy asocjacji w eksploracji tekstu dokumentów dotyczących technologii produkcji i przetwarzania metali [17].

4. Podsumowanie

Opracowana metodyka dostarcza uniwersalnego narzędzia pozyskiwania wiedzy – z jednej strony precyzja sieci neuronowych czy maszyny wektorów nośnych, a z drugiej – intuicyjność i przejrzystość drzew decyzyjnych i możliwość wnikliwej analizy zależności. Stworzone w ten sposób bazy wiedzy, formalizowane z wykorzystaniem najnowszych sposobów reprezentacji wiedzy, odpowiadają potrzebom użytkowników i uczestników Industry 4.0 w możliwie szerokim zakresie. Wykazano również, że można je wykorzystać w wytwarzaniu różnych rodzajów materiałów, o bardzo różnej specyfice. Prowadzone badania wykazały, że jest możliwe wykorzystanie tych samych metod sztucznej inteligencji (uczenia maszynowego) do modelowania różnych procesów w obszarze przetwórstwa metali. Dzięki proponowanym badaniom udało się zaproponować zestaw narzędzi analitycznych pozwalających na rozwiązywanie i modelowanie różnych, często skomplikowanych problemów przetwórstwa metali jednocześnie pozyskując wiedzę na ich temat w formie zrozumiałej dla człowieka i przetwarzalnej dla maszyny, co jest warunkiem koniecznym wspierania idei Industry 4.0.

Podziękowania

Praca została zrealizowana jako część badań podstawowych finansowanych z subwencji Ministerstwa Nauki i Szkolnictwa Wyższego nr 16.16.110.663 – B. Mrzygłód. Dziękujemy również za wsparcie finansowe z Programu Operacyjnego Inteligentny Rozwój: POIR.01.01.01-00-0996/19 – K. Regulski oraz POIR.01.02.00-00-0091/19 – A. Opaliński.

Literatura

1. Hey T., Tansley S., Tolle K., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009, ISBN: 978-0-9825442-0-4.
2. Mueller T., Kusne A.G., Ramprasad R., *Machine learning in materials science: Recent progress and emerging applications*, Reviews in Computational Chemistry 29, 2016, s. 186-273.
3. Bartók A.P., Poelking C., Bernstein N., Kermode J.R., Csányi G., Ceriotti M., *Machine learning unifies the modeling of materials and molecules. Science advances*, 3(12), 2017, s. e1701816.
4. Butler K.T., Davies D.W., Cartwright H., Isayev O., Walsh A., *Machine learning for molec-ular and materials science*, Nature 559(7715), 2018, s. 547-555.
5. Regulski K., Wilk-Kołodziejczyk D., Kluska-Nawarecka S., Szymczak T., Gumienny G., Jaskowiec K., *Multistage discretization and clustering in multivariable classification of the impact of alloying elements on properties of hypoeutectic silumin*, Archives of Civil and Mechanical Engineering, 19(1), 2019, s. 114-126.
6. Regulski K., *Data mining and machine learning in aspects of acquiring knowledge about the production and processing of metals for the needs of Industry 4.0*, Hutnik 2020(4), 10.15199/24.2020.4.3.
7. Mrzygłód B., Gumienny G., Wilk-Kołodziejczyk D., *Application of Selected Artificial Intelligence Methods in a System Predicting the Microstructure of Compacted Graphite Iron*, J. of Materi Eng and Perform 28, 2019, s. 3894-3904, https://doi.org/10.1007/s11665-019-03932-4.
8. Regulski K., Wilk-Kołodziejczyk D., Szymczak T., Gumienny G., Gietka T., Pirowski Z., Jaskowiec K., Kluska-Nawarecka S., *Data Mining Methods for Prediction of Multi-Component Al-Si Alloy Properties Based on Cooling Curves. Journal of Materials Engineer-ing and Performance (JMEP*), 28, 2019, s. 7431-7444, doi:10.1007/s11665-019-04442-z.
9. Szeliga D., Kusiak J., Rauch Ł., *Sensitivity analysis as Support for Design of Hot Rolling Technology of Dual Phase Steel Strips*, 2012, [w:] Kusiak J., Majta J., Szeliga D. (red.), *Metal Forming 2012: proceedings of the 14th international conference on Metal Forming*, Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA (Steel Research International) 2021, s. 1275-1278.
10. Rauch Ł., Kusiak J., Regulski K., *Artificial Intelligence in steel industry – from casting to final product*, [w:] *The Metal Forming Conference MEFORM*, 2020, s. 11-14, ISBN 978-3-86012-632-5.
11. Gronostajski Z., Hawryluk M., Kaszuba M., *The expert system supporting the assessment of the durability of forging tools*, Int J Adv Manuf Technol 82, 2016, s. 1973-1991, https://doi.org/10.1007/s00170-015-7522-3.
12. Hawryluk M., Mrzygłód B., *Application of Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict the wear of forging tools*, Metal, 2016, 25, s. 90, [w:] International Conference on Metallurgy and Materials: May 2016, Brno, Czech Republic, Eu: list of abstracts, Ostrava: TANGER Ltd., cop. 2016, ISBN: 978-80-87294-66-6.
13. Mrzygłód B., Hawryluk M., Gronostajski Z., Opaliński A., Kaszuba M., Polak S., Widomski P., Ziemba J., Zwierzchowski M., *Durability Analysis Of Forging Tools Af-ter Different Variants Of Surface Treatment Using A Decision-Support System Based On Artificial Neural Networks*, Archives Of Civil And Mechanical Engineering, 18 (4), 2018, s. 1079-1091, https://doi.org/10.1016/j.acme.2018.02.010
14. Hawryluk M., Mrzyglod B., *A system of analysis and prediction of the loss of forging tool material applying artificial neural networks*, Journal of Mining and Metallurgy, Section B: Metallurgy, 54 (3), 2018, s. 323-337, https://doi.org/10.2298/JMMB180417023H.

15. Mrzygłód B., Hawryluk, M., Janik, M., *Sensitivity analysis of the artificial neu-ral networks in a system for durability prediction of forging tools to forgings made of C45 steel*, Int J Adv Manuf Technol 109, 2020, s. 1385-1395, https://doi.org/10.1007/s00170-020-05641-y.
16. Macioł P., Regulski K., *Development of semantic description for multiscale models of thermo-mechanical treatment of metal alloys*, The Journal of The Minerals JOM 68, 2016, s. 2082-2088.
17. Regulski K., *Formalization of technological knowledge in the field of metallurgy using document classification tools supported with semantic techniques*, Archives of Metallurgy and Materials 62(2), 2017, s. 715-720.

Uczenie maszynowe w materiałoznawstwie

Streszczenie
Badania materiałoznawcze rozpoczynają się w laboratoriach od badań właściwości metali i ich stopów, właściwości materiału w zależności od dodatków i mikrostruktury oraz zmian tych właściwości pod wpływem obróbki. Kolejnym krokiem jest modelowanie procesów, symulacje tych procesów w celu wpłynięcia na możliwość sterowania i monitorowania procesów w warunkach produkcyjnych. Część badań dotyczy trwającego procesu, a następnie badania skupiają się na kontroli jakości procesu, optymalizacji i wykrywaniu nieprawidłowości oraz wad produktu. Na wszystkich etapach badań możliwe jest zastosowanie metod uczenia maszynowego – w zakresie wybranym przez analityka lub eksperta. Mogą to być metody służące do pozyskiwania wiedzy o występujących zjawiskach, wykorzystania tej wiedzy w planowaniu badań i projektowaniu procesów produkcyjnych (zgodnie z 4. paradygmatem nauki), ale mogą to być również modele oparte na danych, dająca możliwość autonomii kontrola wybranego aspektu produkcji (zgodnie z ideą 4. rewolucji przemysłowej). W rozdziale przedstawiono przegląd metod ML na przykładach przypadków materiałoznawstwa w aspekcie MATERIAŁY – PROCESY – FORMALIZACJA WIEDZY.
Słowa kluczowe: uczenie maszynowe, materiałoznawstwo, formalizacja wiedzy, zastosowanie sztucznej inteligencji, Industry 4.0

Machine learning in material science

Abstract
Materials science research dissolve in laboratories from research on the properties of metals and their alloys, the properties of the material depending on the additives and microstructure and changes in these properties under the influence of processing. The next step is processes modeling, simulations of these processes to affect the possibility of control and monitoring of process-es in production conditions. Some studies relate to an ongoing process, and then the research focuses on quality control of the process, optimization and detection of irregularities and product defects. At all stages of research it is possible apply the methods of machine learning – to the extent chosen by the analyst or expert. These can be methods used to obtain knowledge about occurring phenomena, to use this knowledge in research planning and designing production processes (in accordance with the 4^{th} paradigm of science), but they can also be data-driven models giving the possibility of autonomous control of a selected aspect of production (in accordance with the idea of the 4^{th} industrial revolution). The paper presents an overview of ML methods on examples of material science cases in terms of MATERIALS – PROCESSES – KNOWLEDGE FORMALIZATION.
Keywords: machine learning, materials science, knowledge formalization, application of artificial intelligence, Industry 4.0

Grzegorz Słowiński[1]

Uczenie maszynowe z wykorzystaniem języka Python i jego bibliotek

1. Wstęp

Od roku 2010 następuje intensywny rozwój sztucznej inteligencji, w tym w sposób szczególny technologii uczenia maszynowego i uczenia głębokiego (sztuczne sieci neuronowe). W całej szeroko pojętej informatyce obserwuje się deficyt specjalistów. Podobny deficyt specjalistów występuje w obszarze sztucznej inteligencji. Przewaga popytu nad podażą powoduje, że zarobki tzw. *data scientists* i inżynierów uczenia głębokiego są zdecydowanie wyższe, niż średnie w gospodarce [1].

Dwa języki programowania uzyskały szczególną popularność w tym obszarze: język R i język python [2]. W niniejszej pracy postarano się przedstawić (z konieczności ograniczone) wprowadzenie w zagadnienie sztucznej inteligencji, uczenia maszynowego i możliwości użycia języka Python w programowaniu rozwiązań z tego obszaru.

2. Sztuczna inteligencja, uczenie maszynowe i uczenie głębokie

Termin sztuczna inteligencja (z ang. *Artificial Intelligence*) prawdopodobnie utworzył i użył po raz pierwszy w roku 1955 John McCarthy [3] w zgłoszeniu letniego projektu badawczego. Celem projektu było sprawdzenie hipotezy, czy ludzkie uczenie się lub inny przejaw inteligencji można co do zasady opisać tak precyzyjnie, że będzie możliwe stworzenie maszyny symulującej ten proces.

Sztuczna inteligencja (SI) jest pojęciem szerokim i nieostrym. Nie ma ścisłej definicji tego terminu. Z jednej strony terminem sztucznej inteligencji określa się niekiedy stosunkowo proste algorytmy, które w założeniu naśladują sposób myślenia i działania człowieka, z drugiej, terminem tym określa się nieistniejące jeszcze rozwiązanie informatyczne dysponujące samoświadomością [4].

Słownik języka polskiego definiuje sztuczną inteligencję następująco: *dział informatyki badający reguły rządzące zachowaniami umysłowymi człowieka i tworzący programy lub systemy komputerowe symulujące ludzkie myślenie* [5].

Wydaje się, że w praktyce termin SI używany jest również na określenie takiego informatycznego rozwiązania, które symuluje (naśladuje) ludzkie myślenie.

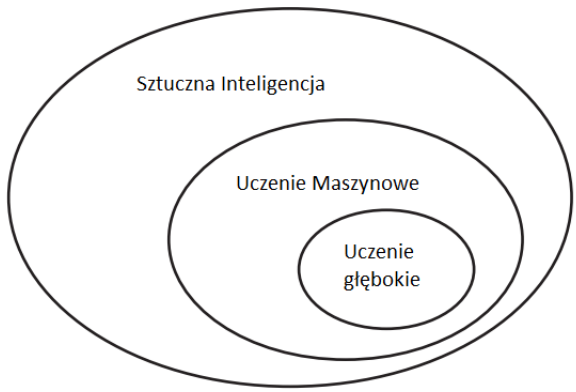
W literaturze wyróżnia się pewne stopnie rozwoju SI i mówi się o wąskiej SI, ogólnej SI i sztucznej nadinteligencji (ang. *Narrow, General i Super Intelligence*) [6].

Nie można zapominać, że czasami termin SI lub sam przymiotnik inteligentny bywa używany na wyrost, w celu marketingowym, aby wykorzystując pozytywną konotację słowa „inteligentny" podnieść znaczenie oferowanego produktu lub usługi w oczach potencjalnego klienta. Stąd spotykamy inteligentne z nazwy: domy, samochody, sprzęty AGD i inne rozwiązania, w których element „inteligencji" może być dyskusyjny.

[1] grzegorz.slowinski@uth.edu.pl, Katedra Informatyki, Wydział Inżynieryjny, Uczelnia Techniczno-Handlowa im. Heleny Chodkowskiej w Warszawie, www.uth.edu.pl.

2.1. Uczenie maszynowe

Uczenie maszynowe (z ang. *Machine Learning*) jest częścią SI. Zależność pomiędzy SI, uczeniem maszynowym (UM) i uczeniem głębokim (UG, z ang. *Deep Learning*) prezentuje rysunek 1. Uczenie maszynowe jest pewnym podzbiorem technik należących do SI, ale jej nie wyczerpuje. Podobnie uczenie głębokie jest pewnym szczególnym podzbiorem technik z zakresu uczenia maszynowego.



Rysunek 1. Związek między sztuczną inteligencją, uczeniem maszynowym i uczeniem głębokim [7]

Istnieje również termin „uczenie płytkie", który określa te wszystkie techniki uczenia maszynowego niezaliczane do uczenia głębokiego.

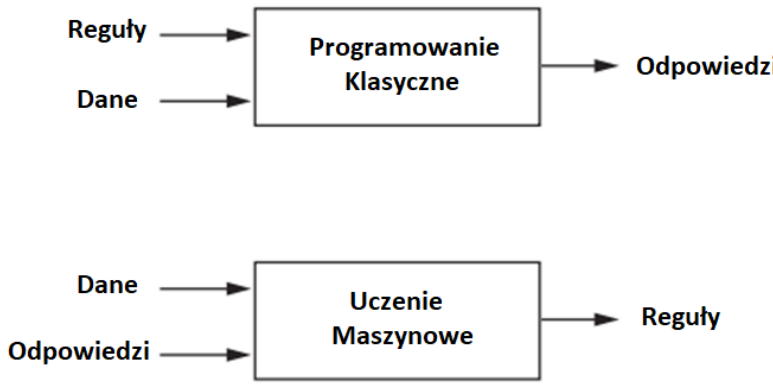
Do uczenia maszynowego zalicza się pewną ilość technik, które potrafią się dość znacznie różnić między sobą. Pojawia się pytanie, w jaki sposób można ocenić, czy daną technikę można zaliczyć do kategorii uczenia maszynowego. Zasadniczą różnicę pomiędzy uczeniem maszynowym a klasycznym sposobem rozwiązywania problemu (klasycznym programem) przedstawia rysunek 2.

W klasycznym podejściu do rozwiązywania problemów decyzyjnych staramy się poznać i opisać reguły podejmowania decyzji, otrzymywania odpowiedzi. Na podstawie reguł powstaje algorytm, który zwykle zostaje wdrożony w postaci programu komputerowego.

Wiele problemów można z powodzeniem rozwiązać, stosując opisane wyżej podejście. Istnieje jednak wiele zagadnień, gdzie nie jest jasne, jakie reguły należy zastosować, aby otrzymać właściwą odpowiedź. Nierzadko, choć nie zawsze są to zagadnienia trywialne dla człowieka. Przykładowo klasyfikacja, czy dana fotografia zawiera podobiznę psa, czy kota jest trywialne człowieka (Zagadnienie to jest popularnym ćwiczeniem podczas nauki budowy modeli UM, zbiór zdjęć kotów i psów można znaleźć w [8]). Jednak ustalenie reguł, jakimi należy się kierować w rozróżnieniu psa od kota i wyrażenie ich przy pomocy ścisłego algorytmu jest ekstremalnie trudne, jeśli w ogóle możliwe.

Uczenie maszynowe polega na ustaleniu reguł, według których należy rozwiązywać dane zagadnienie, poprzez porównanie wielu zestawów danych z właściwymi odpowiedziami. Nawiązując do wspomnianego przykładu klasyfikacji zdjęć kotów i psów – maszyna obserwuje duże ilości zdjęć i porównuje je z przypisanymi do zdjęć właściwymi odpowiedziami (etykietami).

U ujęciu matematycznym model uczący się maszynowo jest swego rodzaju funkcją matematyczną o pewnej ilości współczynników, których wartości ulegają ustaleniu (poprawieniu) w trakcie procesu uczenia, czyli przeglądania kolejnych próbek danych i porównywania ich z odpowiedziami.



Rysunek 2. Klasyczne rozwiązywanie problemu, a uczenie maszynowe [7]

2.2. Uczenie płytkie, a uczenie głębokie

Termin uczenie głębokie (*Deep Learning*) jest często używany, ale rzadziej tłumaczony. Zwykle wiąże się go ze sztucznymi sieciami neuronowymi (SSN), które posiadają co najmniej jedną warstwę ukrytą. Według F. Choletta (twórcy frameworku Keras) przymiotnik „głęboki" w tym kontekście oznacza wielowarstwową reprezentację danych. Każda warstwa modelu SSN tworzy nową reprezentację (przekształcenie) danych wejściowych. „Głębia" w tym kontekście oznacza dużą ilość następujących przekształceń. Nie należy się według Chollet'a doszukiwać w tym terminie sugestii, że technika SSN osiąga jakieś głębsze rozumienie analizowanego zagadnienia [7]. Podstawowe różnice pomiędzy uczeniem płytkim i uczeniem głębokim zgromadzono w tabeli 1.

Tabela 1. Porównanie charakterystyki uczenia płytkiego i uczenia głębokiego

	Uczenie płytkie	Uczenie głębokie
Dane	Nie wymaga dużych zbiorów danych	Duże zbiory danych
Czas uczenia	Krótki	Długi
Wymagany sprzęt	CPU	GPU, proces uczenia z użyciem wyłącznie CPU będzie znacznie dłuższy

Źródło: Opracowanie własne

Zasadniczo uczenie płytkie należy stosować, gdy mamy mniej danych, dane są prostsze (mniej liczb opisuje daną próbkę). Wymaga ono mniejszej mocy obliczeniowej. Modele uczenia płytkiego mają mniej stopni swobody, mniej współczynników, które są dostrajane w trakcie procesu uczenia.

Modele uczenia głębokiego – sieci neuronowe mogą mieć zdecydowanie więcej współczynników, które wymagają dostrojenia w trakcie procesu nauki. Ze względu na naturę procesu, który można w pewnej części zrównoleglić, warto korzystać w obli-

czeniach z kart graficznych (GPU), które przyśpieszają obliczenia. Ze względu na dużą ilość stopni swobody SSN dość łatwo jest wywołać zjawisko przeuczenia, które polega na tym, że modele świetnie operują na danych treningowych, ale słabo na nowych, niewidzianych uprzednio danych testowych.

3. Podstawowe zagadnienia rozwiązywane przy pomocy uczenia maszynowego

Podstawowe zagadnienia rozwiązywane przy pomocy uczenia maszynowego sprowadzają się do zagadnienia klasyfikacji i regresji.

Zagadnienie klasyfikacji polega na przypisaniu danego obiektu do jednej z kilku kategorii, czy klas. Przykładowe zagadnienia tego typu to: rozpoznawanie odmiany irysów po budowie geometrycznej kwiatu, przewidywanie, czy dany pasażer przeżył katastrofę Tytanica, rozpoznawanie odręcznie pisanych cyfr.

Drugi typ zagadnienia to regresja. To zadanie polega na dobraniu właściwej dla danej próbki odpowiedzi z ciągłego przedziału wartości. Przykładowe zagadnienia to: określenie wytrzymałości twardości betonu na podstawie składu i czasu sezonowania, wycena domów, ocena jakości wina.

Niektóre zagadnienia, w zależności od sposobu sformułowania, można przedstawić zarówno jako regresję i klasyfikację. Przykładowo w przypadku przewidywania wytrzymałości betonu możemy zagadnienie sformułować na sposób regresji i opracować model, który będzie przewidywał wytrzymałość określoną liczbą np. w MPa lub zadowolić się określeniem, czy dany beton będzie: mocny, przeciętny, słaby i sformułować zagadnienie w postaci klasyfikacji.

4. Przegląd wybranych metod uczenia płytkiego

4.1. Naiwny Bayes

Naiwne modele Bayesa oparte są na twierdzeniu Bayesa. Są niezwykle szybkie i proste, ale z drugiej strony ich wydajność jest zwykle ograniczona. Można je wykorzystać jako punkt odniesienia dla problemów klasyfikacyjnych [9, s. 382].

4.2. Drzewo decyzyjne

Drzewo decyzyjne (DT) należy do klasy tzw. algorytmów nieparametrycznych. Termin nieparametryczny może wprowadzać w błąd. W rzeczywistości drzewo decyzyjne ma parametry, ale ich liczba nie jest stała.

Podczas fazy uczenia drzewo decyzyjne próbuje znaleźć najlepsze pytania dzielące zbiór danych w celu zmniejszenia nieczystości informacji (miarą jest indeks Giniego lub entropia informacji). Wielką zaletą drzew decyzyjnych jest to, że są niezwykle intuicyjne. Z drugiej strony drzewo decyzyjne nie ma ograniczonych stopni swobody, więc łatwo jest doprowadzić do przetrenowania (jeśli użytkownik nie jest świadomy tego niebezpieczeństwa). Podziały dokonywane przez drzewo decyzyjne są zawsze ortogonalne (dokonywane na jednym elemencie na raz), więc drzewo decyzyjne jest bardzo wrażliwe na rotację danych [10, s. 188].

4.3. Las losowy

Idea losowego lasu jest następująca: Należy wziąć wiele drzew decyzyjnych (wykorzystując pewną losowość, więc drzewa się różnią) i pozwolić im głosować. Zatem decyzja klasyfikacyjna podjęta przez losowy las jest decyzją podjętą przez najliczniejszą grupę drzew decyzyjnych w losowym zbiorze drzew.

Zwykle losowy las działa lepiej niż pojedyncze drzewo decyzyjne. Jednak losowy las jest uważany za model „czarnej skrzynki", który jest bardzo trudny do interpretacji [9, 10].

4.4. Maszyna wektorów nośnych

Maszyny wektorów nośnych (SVM), które mogą być używane jako regresory lub klasyfikatory, są uważane za bardzo wydajne i elastyczne algorytmy. Z drugiej strony. mogą potrzebować dużej mocy obliczeniowej (patrz [9, s. 405]).

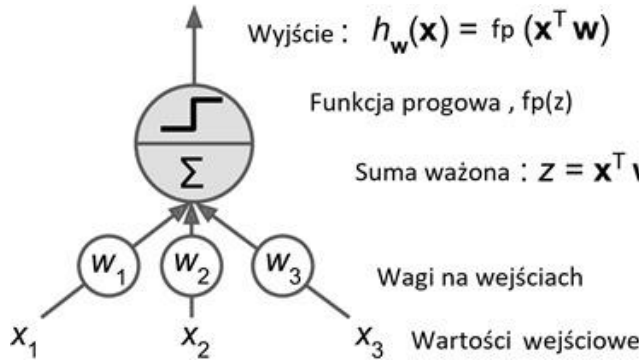
Zasada SVM polega na podzieleniu klas przez „rysowanie linii" (lub płaszczyzny) w sposób, który maksymalizuje margines między klasami. Ponieważ linie proste (lub płaszczyzny) zwykle nie dają najlepszego rozwiązania, SVM może stosować różne jądra (wielomianowe, promieniste i inne). SVM jest szerzej wyjaśnione w [9, 10].

4.5. Regresory

Regresory oparte są o metodę najmniejszych kwadratów (NK) lub jej modyfikacje. Metoda najmniejszych kwadratów znajduje współczynniki równania linii prostej poprowadzonej w taki sposób, aby suma kwadratów różnic pomiędzy wartościami w faktycznych punktach pomiarowych a wartościami obliczonymi z równania prostej była minimalna. Użyta bezpośrednio metoda NK pozwala na analizę zagadnień liniowych. Aby móc zastosować tę metodę do zależności nielinowych dokonuje się transformacji danych przy użyciu dodatkowej funkcji (najczęściej stosuje się wielomian n-tego stopnia, ale można i inne), aby zagadnienie nieliniowe sprowadzić do zagadnienia linowego. Regresory stosuje się w zagadnieniach, gdzie oczekujemy konkretnej wartości liczbowej np. wyznaczenie wytrzymałości betonu, ceny domu, ilości wypożyczonych rowerów itp.

5. Sztuczne sieci neuronowe – uczenie głębokie

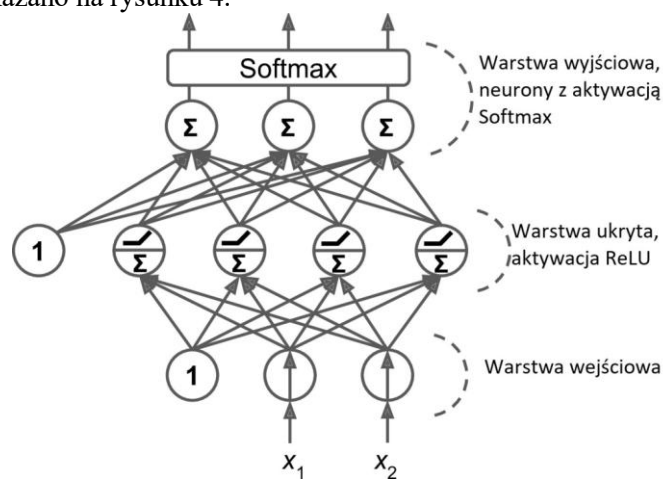
Uczenie głębokie wiąże się nierozerwalnie ze sztucznymi sieciami neuronowymi. Sztuczne sieci neuronowe zostały zainspirowane naturalnymi sieciami neuronowymi. Składają się ze sztucznych neuronów (nazywanych też perceptronami). Neuron wykonuje operację sumowania ważonego sygnałów wejściowych (każdy sygnał jest mnożony przez wagę, czyli współczynnik na danym wejściu). Następnie tak otrzymana suma jest podstawiana jako parametr do funkcji aktywacji. Tak otrzymana wartość jest przesyłana na wejście neuronów w kolejnych warstwach (lub na wyjście). Najczęściej stosowane funkcje aktywacji to ReLU (jednostronnie obcięta funkcja liniowa), sigmoidalna, softmax (znormalizowana funkcja wykładnicza). Model sztucznego neuronu wyposażonego w progową funkcję aktywacji ukazano na rysunku 3.



Rysunek 3. Neuron z progową funkcją aktywacji, źródło [10]

5.1. Sieci gęste

W sieci gęstej neuron z danej warstwy przyjmuje na wejście sygnały z wyjść wszystkich neuronów z warstwy wcześniejszej. Podobnie wartość obliczona w neuronie przekazywana jest na wejście do wszystkich neuronów kolejnej warstwy. Schemat sieci gęstej ukazano na rysunku 4.

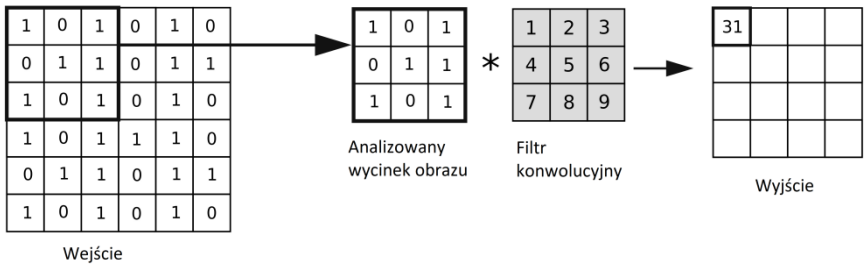


Rysunek 4. Schemat współczesnej sieci neuronowej stosowanej w zagadnieniach klasyfikacji. Neurony w warstwie ukrytej są aktywowane ReLU, a w warstwie wyjściowej Softmax, źródło [10]

5.2. Sieci konwolucyjne

Sieci konwolucyjne zyskały popularność w ostatnich latach. Okazują się niezwykle skuteczne w analizie obrazów, ale również sprawdzają się w analizie tekstów i szeregów czasowych (np. zmiana cen w czasie). Neuron w sieci konwolucyjnej jest co do zasady zbudowany w taki sam sposób jak neuron w sieci gęstej. Ma jednak mniej połączeń na wejściu i wyjściu. Różnicę najlepiej chyba wyjaśnić na przykładzie warstwy wejściowej. W przypadku analizy obrazu siecią gęstą neuron w warstwie wejściowej otrzyma informacje ze wszystkich pikseli obrazu (traci jednak informacje o sąsiedztwie tych pikseli między sobą). Neuron konwolucyjny analizuje naraz kilka pikseli położonych obok siebie. W ten sposób lepiej „zauważa" szczegóły obrazu, abstrahując od ich położenia na ekranie lub lepiej zauważa szczególne sąsiedztwo słów w tekście lub gwałtowną

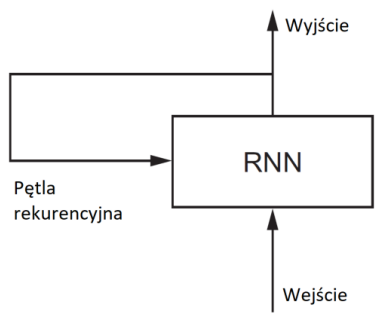
zmianę ceny na wykresie giełdowym. Przypomina to trochę sytuację, gdyby człowiek oglądał obraz, przykładając do niego kartkę z wyciętą w niej mała dziurką i przesuwając tę dziurkę po całym obrazie.



Rysunek 5. Schemat działania neuronu/filtra konwolucyjnego, źródło [11]

5.3. Sieci rekurencyjne

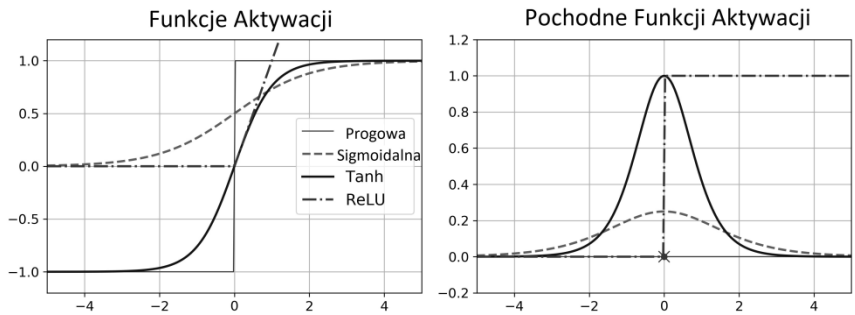
Sieci rekurencyjne nadają się do przetwarzania tekstów lub innych szeregów czasowych (czy podobnych), gdzie ważna jest kolejność przetwarzanych elementów. Podobnie jak w zdaniu kolejne słowa zmieniają znaczenie w zależności od poprzedzającego je kontekstu (wcześniejszych słów), tak w sieci rekurencyjnej aktualny stan wyjścia zależy nie tylko od bieżącego stanu wejść, ale od stanu wyjściowego osiągniętego w poprzednim kroku, który zostaje podany na wejście. W ten sposób sieć niejako zapamiętuje, sens tego, co było analizowane uprzednio. Schemat działania sieci rekurencyjnej ukazuje rysunek 6.



Rysunek 6. Schemat działania sieci rekurencyjnej. W kolejnym kroku analizy częścią danych wejściowych staje się wynik otrzymany w poprzednim kroku, źródło [11]

5.4. Znaczenie funkcji aktywacji ReLU

Gwałtowny rozwój SSN, który nastąpił po roku 2010 ma kilka przyczyn. Na pewno przyczynił się do tego rozwój sprzętu komputerowego (większe, tańsze moce obliczeniowe), w szczególności nowe lepsze karty graficzne, szczególnie od firmy NVIDIA, rozwój Internetu i powstanie większej ilości dużych zbiorów danych (skomplikowane modele potrzebują wielkich zbiorów treningowych, które w praktyce nie są łatwe do pozyskania). Warto zwrócić jednak szczególną uwagę na wprowadzenie nowych funkcji aktywacji, mianowicie Rectified Linear Unit (ReLU) w miejsce wcześniej używanej funkcji sigmoidalnej. Wykresy funkcji aktywacji i ich pochodne zostały przedstawione na rysunku 7.



Rysunek 7. Funkcje aktywacji i ich pochodne [10]

Pełną analogię do działania neuronu naturalnego uzyskano by używając funkcji progowej. Ta ma jednak tę wadę, że jej pochodna wynosi zero, co zasadniczo uniemożliwiłoby realizację procesu uczenia sieci z użyciem algorytmu propagacji wstecznej. W miejsce funkcji progowej zaczęto używać przypominającej ją kształtem funkcji sigmoidalnej, która ma pochodną wprawdzie niezerową, ale zbliżoną do zera dla wszystkich x odległych od zera. Podczas propagacji wstecznej, proces uczenia głębszych (bliższych wejścia) warstw wymaga wymnożenia pochodnych z warstw płytszych (bliższych wyjścia). W efekcie gradient dla głębszych warstw zbiega bardzo szybko do zera, co czyni proces uczenia głębszych warstw praktycznie niemożliwym. Natomiast, gdy używamy ReLU, pochodna dla $x > 0$ wynosi 1. Jeden podniesione do dowolnej potęgi nadal daje jeden. To powoduje, że możliwe jest uczenie zdecydowanie głębszych sieci, nie kilka, ale kilkadziesiąt lub kilkaset warstw nie jest już niczym nadzwyczajnym. Dodatkowym atutem ReLU jest fakt, że pochodną tej funkcji liczy się szybciej (0 lub 1), niż funkcji sigmoidalnej. Z drugiej strony – ReLU przynosi problem tzw. umierających/martwych neuronów [10 s. 329]. W praktyce jednak okazuje się on zdecydowanie mniej poważny. Można też go w razie potrzeby ominąć, stosując zbliżone do ReLU funkcje Leaky ReLU i ELU.

6. Charakterystyka języka Python

Język python powstał w latach 1989-91 [12]. Obecnie dominuje wersja trzecia. Python nie zachowuje wstecznej zgodności. Kod powstały we wcześniejszych wersjach może wymagać zmian, w celu uzgodnienia z wersją 3. Python stał się w ostatnich latach jednym z najpopularniejszych języków. Jest to język hybrydowy, umożliwia pisanie z użyciem kilku paradygmatów: strukturalno-proceduralnego, obiektowego i funkcyjnego. W programowaniu dla potrzeb uczenia maszynowego w zasadzie wystarczy dość podstawowa znajomość Pythona. Programista raczej ogranicza się do korzystania z funkcji i klas dostarczonych w dedykowanych pakietach, z rzadka pisząc własne. Index tiobe [13] w sierpniu 2021 roku pokazuje, że pierwsza piątka najpopularniejszych języków programowania to: C, Java, Python, C++ i C#. Jak wiadomo C++, C# i Java wyewoluowały w pewnym stopniu z języka C i dziedziczą po nim składnię i pewne inne elementy języka, np. statyczne typy. Prawie każdy programista zetknął się z którymś z tych „C-kształtnych" języków. Python różni się dość wyraźnie od C i jego „dzieci", stąd pierwszy kontakt może być nieco utrudniony. Poniżej postarano się omówić w skrócie najważniejsze różnice między Pythonem a językami „C-kształtnymi".

6.1. Sztywna składnia

Języki z rodziny C poza nielicznymi wyjątkami[2] ignorują znaki białe. Strukturę programu zaznacza się, używając przede wszystkim średników i nawiasów klamrowych. W przypadku Pythona każda instrukcja musi znaleźć się w odrębnym wierszu. Kapitalną funkcję pełnią też wcięcia. Przy ich pomocy zaznacza się zagnieżdżone bloki kodu.

6.2. Język interpretowany

Języki „C-kształtne" są kompilowane, powstają odrębne pliki z kodem źródłowy i z kodem wykonywalnym. Proces kompilacji pozwala na zlokalizowanie błędów. Python jest językiem interpretowanym (językiem skryptowym). Jest tłumaczony i wykonywany na bieżąco. To powoduje, że jest wolniejszy i bardziej podatny na błędy w czasie wykonania. Brak procesu kompilacji pozwala na dokonywanie szybkich modyfikacji w kodzie. W Pytonie się wygodnie eksperymentuje i prototypuje.

6.3. Dynamiczne typowanie

Python jest językiem z dynamicznym typowaniem. Definiując zmienną nie podaje się jej typu. Typ powiązany jest z wartością, a nie zmienną. Dana zmienna może w trakcie programu przechowywać najróżniejsze wartości: liczby całkowite i zmiennoprzecinkowe, teksty, jak również kontenery: listy, słowniki itp., a nawet funkcje. W istocie zmienna jest referencją do obiektu dowolnego typu. Zaś informacja o rodzaju obiektu jest przechowywana w samym obiekcie. Podobny efekt można by uzyskać w języku Java, gdyby ktoś konsekwentnie wszystkim zmiennym nadawał typ Object (oczywiście w praktyce nikt tak w języku Java nie robi). Brak kontroli typów podczas tworzenia programu skutkuje częstszymi błędami podczas wykonywania programów, obniża też w pewnym stopniu jego czytelność.

6.4. Indeksowanie list

Python indeksuje listy od 0. Pozwala na indeksowanie ujemne, od końca listy. Indeks -1 oznacza ostatni element na liście, -2 – przedostatni itd. Wygodnie można pozyskiwać wycinki list. Służy do tego znak dwukropka. Indeks [5:8] oznacza elementy od 5 włącznie do 8 wyłącznie.

6.5. Funkcje

W języku Python nie ma przeciążenia funkcji. Jeżeli funkcja zostanie zdefiniowana kilkukrotnie – obowiązująca będzie jej ostatnia definicja. Funkcjonalność zbliżoną do przeciążenia można uzyskiwać dzięki ciekawej budowie nagłówka funkcji. Otóż w nagłówku pythonowej funkcji są 4 kategorie argumentów: 1) zwykłe zmienne, 2) lista, 3) argumenty kluczowe (klucz-wartość) i 4) słownik.

Dodatkowo Python pozwala podczas definiowania funkcji nadawać zmiennym wartości domyślne. Obie te cechy pozwalają na elastyczne definiowanie funkcji. Zaskakująco mogą wyglądać wywołania funkcji. W przypadku argumentów kluczowych, ich kolejność w wywołaniu nie ma znaczenia – rozpoznawane są po nadanych kluczach.

[2] Porównaj przykładowo: a++ +b; a+ ++b; i niepoprawne a+++b.

7. Biblioteki Python dedykowane uczeniu maszynowemu

„Czysty" język Python nie nadaje się do analizy danych i implementowania algorytmów uczenia maszynowego. Dynamiczne typowanie i skryptowość obniżają szybkość wykonania. Opracowano szereg bibliotek (frameworków), które wspierają pracę z danymi, ich wizualizację i tworzenie modeli

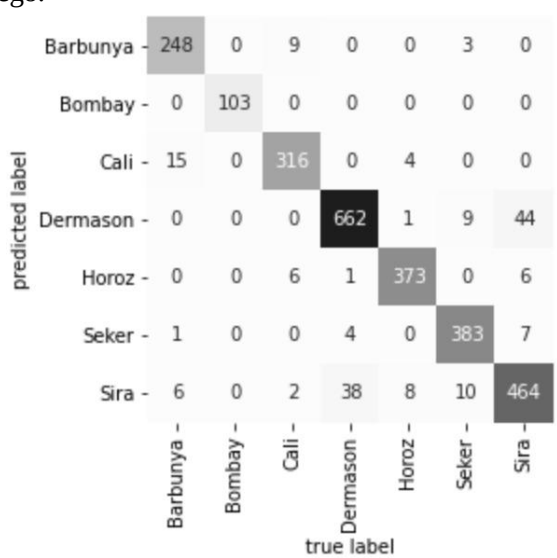
7.1. Numpy i Pandas

Numpy, skrót od Numerical Python [14] jest biblioteką, która dostarcza struktury danych (tablice) i funkcje do operowania na wartościach liczbowych. Tablice dostarczane przez Numpy charakteryzują się tym, że mogą przechowywać wyłącznie numeryczne wartości homogeniczne (czyli w danej tablicy mogą być liczby dokładnie tego samego typu). W gruncie rzeczy Numpy dostarcza tablic, jakie znamy z języka C, czy C++. Heterogeniczne listy z „czystego" Pythona (które w istocie są listami referencji do obiektów) są powolne w użyciu. Wprawdzie od wersji 3.3 Python pozwala na tworzenie homogonicznych list, ale to udogodnienie przynajmniej na razie nie zagroziło popularności frameworku Numpy.

Pakiet Pandas (skrót od Python Data Analysis Library) [15] pozwala na manipulację danymi nienumerycznymi. Oprócz liczb struktury dostarczane przez Pandas (Serie – odpowiadające w przybliżeniu kolumnie danych i DataFrame – odpowiadające tabeli, czy arkuszowi kalkulacyjnemu) mogą przechowywać dane tekstowe.

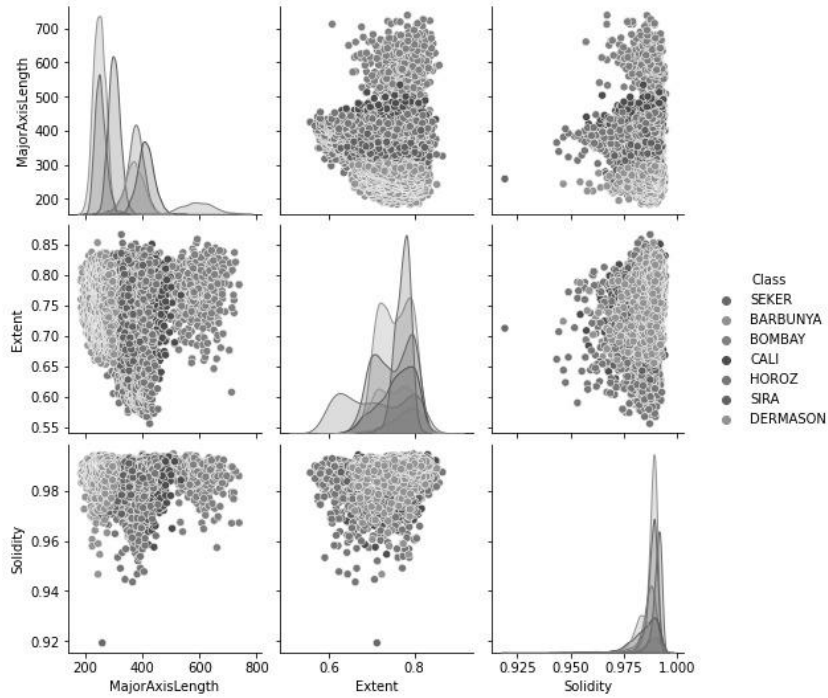
7.2. Matplotlib i Seaborn

Biblioteka Matplotlib [16] pomaga w wizualizacji danych. Nowsza biblioteka Seaborn [17] jest swego rodzaju nakładką na Matplotlib, która umożliwia wygodniejszą pracę z Pandas (Matplotlib powstał wcześniej niż Pandas) oraz pozwala na uzyskiwanie w prosty sposób atrakcyjnych wizualizacji. Na rysunku 8. jako przykład wizualizacji zaprezentowana macierz pomyłki często spotykaną w projektach z zakresu uczenia maszynowego.



Rysunek 8. Macierz pomyłki (confussion matrix) dla procesu klasyfikacji ziaren różnych odmian fasoli, opracowanie własne

Bardzo efektowne (i efektywne) potrafią też być wykresy macierzowe (Pairplot), które niestety tracą dużą część swego uroku, gdy są reprodukowane w odcieniach szarości, rysunek 9.



Rysunek 9. Wykres macierzowy (pairplot) wybranych cech ziaren fasoli z różnych odmian, opracowanie własne

Rysunki 8 i 9 w kolorach wraz z generującym je kodem można obejrzeć w [18, 19].

7.3. Scikitlearn

Scikit-learn (sklearn) [20] jest biblioteką, która udostępnia modele uczenia maszynowego oraz funkcje przydatne w przygotowaniu danych (np. podział na zbiór treningowy i testowy, przeskalowanie/normalizację danych) i analizie wyników (np. obliczenie precyzji modelu). Poniżej ukazano użycie funkcji train_test_split, bardzo użytecznej i często spotykanej.

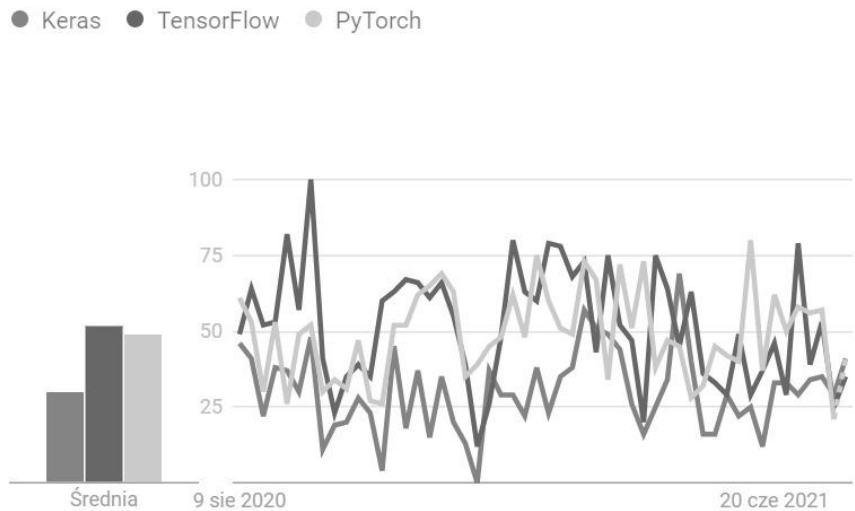
```
Xtrain,Xtest,Ytrain,Ytest=train_test_split(X_beans,
Y_beans,test_size=0.2, random_state=1)
```
[3]

Dostępny jest szeroki wybór modeli uczenia płytkiego jak również istnieje możliwość budowy dość prostych sztucznych sieci neuronowych. Ta część frameworku jest zdaje się rzadziej używana, gdyż ma godnych konkurentów w frameworkach Keras, czy Fastai dedykowanych wyłącznie budowie modeli uczenia głębokiego.

[3] Warto zwrócić uwagę na pythonową specyfikę ukazanego wywołania. Parametry kluczowe test_size i random_state w nawiasach obejmujących listę parametrów i fakt, że po lewej stronie operatora przypisania (znak =) stoją aż 4 zmienne. Możliwość zwrócenia wyników z funkcji do całego szeregu zmiennych może zaskakiwać nowych programistów.

7.4. Frameworki dedykowane uczeniu głębokiemu

Popularne obecnie frameworki do uczenia głębokiego to: TensorFlow, PyTorch i Keras [21-23]. TensorFlow i Keras są związane z Google, PyTorch z Facebookiem. Przybliżoną informację o popularności tych bibliotek daje rysunek 10, na którym ukazano porównanie dokonane z użyciem serwisu Google Trends.



Rysunek 10. Porównanie popularności frameworków uczenia głębokiego: Keras, TensorFlow, PyTorch. TensorFlow i PyTorch cieszą się podobnym zainteresowaniem internautów. Keras pozostaje nieco w tyle

Keras jest wysokopoziomową biblioteką, która podczas swojego działania korzysta z zaplecza obliczeniowego TensorFlow[4]. Poniżej ukazano fragment kodu napisanego w Keras, który definiuje kilkuwarstową siec do rozwiązywania zagadnienia klasyfikacji.

```
#Typ modelu
annRFNS=Sequential()
#warstwa wejściowa,
#określenia wymaga wymiar danych wejściowych
annRFNS.add(Dense(17,input_dim=8,activation = 'relu'))
#warstwy ukryte
annRFNS.add(Dense(12,activation='relu'))
annRFNS.add(Dense(3,activation='relu'))
# warstwa wyjściowa
annRFNS.add(Dense(7,activation='softmax'))
# ustalenie optymalizatora, funkcji straty
annRFNS.compile(optimizer='Adam',
loss='categorical_crossentropy')
```

Uderza zwięzłość kodu, w którym jedną linijką definiuje się od razu całą warstwę sieci. Przekłada się to na łatwość modyfikacji architektury i wynikającą stąd łatwość budowania prototypów.

[4] Od wersji 2.4 Keras korzysta wyłącznie z TensorFlow. Wcześniejsze wersje mogły integrować się z większą ilością bibliotek:TensorFlow, Theano, MXNet i inne.

Pythorch i TensorFlow są w porównaniu do Keras niskopoziomowe (oferują mniejszy poziom abstrakcji). PyTorch bywa oceniany jako mniej zawiły, bardziej zgodny z założeniami języka python i posiadający więcej możliwości niż TensorFlow [24].

Warto wspomnieć o powstałym w środowisku akademickim frameworku fastai [25], którego popularność jest wprawdzie jeszcze umiarkowana, ale który pod pewnymi względami przewyższa dokonania Googla [26].

8. Środowisko pracy. IPython, Jupyter Notebook i chmurowe odpowiedniki

8.1. IPython i Jupyter

Programować w pythonie można w licznych środowiskach programistycznych. Społeczność używająca Pythona do analizy danych i budowania modeli uczenia maszynowego upodobała sobie środowisko IPython (*Interactive Python*), które zwykle używa się z nakładką o nazwie Jupyter uruchamianą w przeglądarce internetowej [27]. IPython pozwala na wygodne przeplatanie fragmentów kodu tekstem. Można również zapisać efekty wykonanych skryptów (wykresy, tabele itp.). W ten sposób można tworzyć bogate w treść notatniki, gdzie tekst objaśnia prowadzone obliczenia. Niektóre podręczniki dotyczące uczenia maszynowego posiadają wręcz swoje odpowiedniki w formie katalogów notatników IPythona dostępnych w serwisie github lub innych lokalizacjach internetowych.

8.2. Notatniki chmurowe

Istnieją usługi, które umożliwiają korzystanie z IPythona w chmurze obliczeniowej (on-line). Tego typu usługi oferują Google Colab, Amazon SageMaker, PaperSpace Gradient. Dostępne są zarówno warianty płatne, jak i bezpłatne. Rozwiązanie chmurowe ma szereg zalet: Uwalnia od konieczności dość żmudnej instalacji, pozwala na skorzystanie z komputera ze wsparciem GPU lub TPU, ułatwia współpracę grupową nad opracowywanymi rozwiązaniami i prezentowanie ich światu. Jak wynika z doświadczenia autora, darmowo dostępne maszyny są na tyle mocne, że z powodzeniem radzą sobie w rozsądnym czasie (sekundy – pojedyncze minuty) z zagadnieniami uczenia płytkiego i prostego uczenia głębokiego (sieci zbudowane z kilku warstw).

8.3. Serwer w chmurze

W przypadku bardziej złożonych obliczeniowo procesów uczenia maszynowego darmowe maszyny mogą okazać się niewystarczające. W takiej sytuacji warto rozważyć skorzystanie z usług chmurowych oferowanych przez wszystkich głównych dostawców (Google Cloud, AWS, Azure) i innych. Ceny za zdalną maszynę kształtują się na poziomie około 0,5-1,5 dolara za godzinę pracy maszyny. Trzeba pamiętać, aby maszynę wyłączyć, a nie po prostu zamknąć okno przeglądarki. Niektórzy dostawcy udostępniają dodatkowo pewną pulę „kredytu” na zachętę dla nowych użytkowników.

9. Podsumowanie

Uczenie maszynowe pozwala na podejmowanie zagadnień, których rozwiązanie w klasyczny sposób byłoby bardzo trudne lub zgoła niemożliwe. Rozwój techniki komputerowej, większy dostęp do mocy obliczeniowej, a także łatwiejszy dostęp do dużych zbiorów danych obniżył barierę wejścia i ułatwił prowadzenie prac w tym obszarze. Wielką pomocą jest język Python i dedykowane uczeniu maszynowemu

frameworki, które pozwalają na szybkie prototypowanie rozwiązań z zakresu UM. Biblioteki obfite w gotowe, dobrze udokumentowane funkcje pozwalają tworzyć skomplikowane modele uczenia maszynowego w kilkunastu, kilkudziesięciu, może kilkuset linijkach kodu. Dostępność podręczników i internetowych instruktaży jest wystarczająca. Programista operuje na wysokim poziomie abstrakcji (ogólności), nie musi dbać o szczegóły prowadzonych obliczeń, czy tworzonych wizualizacji. Łatwość modyfikacji tak tworzonego kodu jest wyśmienita. Zapewnione jest również łatwe wsparcie dla zrównoleglenia obliczeń z użyciem GPU. Fakt, że za głównymi frameworkami do uczenia głębokiego: Keras i PyTorch stoją potężne konkurujące korporacje: Google i Facebook gwarantuje, że w nadchodzących latach możemy się spodziewać dalszego rozwoju tych narzędzi.

Literatura

1. Schroeder B., The *Data Analytics Profession And Employment Is Exploding – Three Trends That Matter*, Forbes, 11.06.2021, https://www.forbes.com/sites/bernhardschroeder/2021/06/11/the-data-analytics-profession-and-employment-is-exploding-three-trends-that-matter/?sh=540b47133f81
2. edX team, 9 Top Programming Languages for Data Science, 24.05.2021, https://blog.edx.org/9-top-programming-languages-for-data-science.
3. McCarthy J., *A proposal for the dartmouth summer research project on artificial intelligence*, August 31, 1955, http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, dostęp 26.07.2021.
4. Zalewski T., *Definicja sztucznej inteligencji,* [w:] *Prawo sztucznej inteligencji*, pod red. Lai L., Świerszczyński M., Beck C.H., 2020.
5. Słownik języka polskiego PWN, https://sjp.pwn.pl/szukaj/sztuczna%20inteligencja.html, dostęp 26.07.2021.
6. Kaplan A., Haenlein M., *Business Horizons*, Volume 62, Issue 1, January-February 2019, s. 15-25.
7. Chollet F., *Deep Learning, Praca z językiem Python i biblioteką Keras*, Helion, 2019.
8. https://www.kaggle.com/c/dogs-vs-cats.
9. VanderPlas J., *Python Data Science Handbook*, O'Reilly, 2017.
10. GéronA., *Handson Machine Learning with Scikit-Learn*, Keras & TensorFlow, O'Reilly, 2019.
11. Reynolds Anh H., *Convolutional Neural Networks* (CNNs), https://anhreynolds.com/blogs/cnn.html, dostęp 07.08.2021.
12. *History and License of Python*, https://docs.python.org/3.9/license.html, dostęp 07.08.2021.
13. *Indeks tiobe*, https://www.tiobe.com/tiobe-index/.
14. Dokumentacja biblioteki numpy, https://numpy.org/doc/, dostęp 07.08.2021.
15. Dokumentacja biblioteki pandas, https://pandas.pydata.org/docs/, dostęp 07.08.2021.
16. Oficjalna strona biblioteki matplotlib, https://matplotlib.org/stable/index.html, dostęp 07.08.2021.
17. Dokumentacja biblioteki seaborn, https://seaborn.pydata.org/.
18. Słowiński G., *Dry Beans Species Classification with Machine Learning Techniques*, notatnik colab, https://colab.research.google.com/drive/11X6VevSMybenGkRqK1Xj_1EJmU3vomFB?usp=sharing.
19. Słowiński G., *Influence of Dimension Reduction and Regularization on Machine Learning Performance*, notatnik colab, https://colab.research.google.com/drive/1l5lH1QgesDX8CbbkqcnmlbqwcXfksGQB?usp=sharing.
20. Dokumentacja biblioteki scikit-learn, https://scikit-learn.org/stable/.
21. Oficjalna strona TensorFlow, https://www.tensorflow.org/.

22. Oficjalna strona PyTorch, https://pytorch.org/.
23. Oficjalna strona Keras, https://keras.io/.
24. Sopyła K., *Dlaczego porzuciłem Tensorflow na rzecz Pytorch*, wpis z 27.01.2019, https://ksopyla.com/machine-learning/pytorch-vs-tensorflow/.
25. Oficjalna strona biblioteki fastai, https://www.fast.ai/.
26. Knight W., *A small team of student AI coders beats Google's machine-learning code*, MIT Technology Review, 10.08.2018, https://www.technologyreview.com/2018/08/10/141098/small-team-of-ai-coders-beats-googles-code/.
27. Oficjalna strona projektu IPython, https://ipython.org/.

Uczenie maszynowe z wykorzystaniem języka python i jego bibliotek

Streszczenie

W pracy wyjaśniono pojęcia sztucznej inteligencji, uczenia maszynowego (UM), uczenia płytkiego (UP) i uczenia głębokiego (UG). Wyjaśniono różnicę między uczeniem nadzorowanym i nienadzorowanym. Zwrócono uwagę na „pułapki" uczenia maszynowego, przede wszystkim na ryzyko przetrenowania.

Przedstawiono podstawowe problemy UM: klasyfikację i regresję. Przedstawiono wybrane metody UP: regresję, drzewa decyzyjne, lasy losowe, maszyny wektorów nośnych.

Przedstawiono sieci neuronowe gęste oraz stosunkowo nowe sieci konwolucyjne, które okazały się szczególnie skuteczne w analizie obrazów. Zwrócono uwagę na funkcję aktywacji rectified linear unit (ReLU), której zastosowanie umożliwiło radykalne zwiększenie ilości warstw ukrytych sieci.

Scharakteryzowano w skrócie język python i jego biblioteki wykorzystywane w UM: numpy, pandas, matplotlib, scikit-learn i keras.

Wspomniano najpopularniejsze narzędzie do budowy skryptów UM: IPython i jupyter notebook oraz jego wersje w chmurze obliczeniowej.

Słowa kluczowe: uczenie maszynowe, uczenie głębokie, python, scikitlearn, keras

Machine Learning with python language and its frameworks

Abstract

The paper explains the concepts of artificial intelligence, machine learning (ML), shallow learning (SL) and deep learning (DL). The difference between supervised and unsupervised learning was explained. Machine learning "traps" have been highlighted, in particular the risk of overtraining.

Basic problems of ML: classification and regression are presented. Selected methods of SL are presented: regression, decision trees, random forests, support vector machines.

The paper presents dense neural networks and relatively new convolutional networks, which turned out to be particularly effective in image analysis. Attention was paid to the activation function of the rectified linear unit (ReLU), the use of which allowed for a radical increase in the number of hidden layers of networks.

The python language and its libraries used in ML are briefly characterized: numpy, pandas, matplotlib, scikit-learn and keras.

The most popular tool for building ML scripts was mentioned: IPython and jupyter notebook and its cloud versions.

Keywords: machine learning, deep learning, python, scikitlearn, keras

Paweł Andruloniw[1], Piotr Zwierzykowski[2]

W jaki sposób metody uczenia maszynowego mogą pomóc w wykrywaniu degradacji i szacowaniu jakości usług wideo w sieciach telekomunikacyjnych?

1. Wprowadzenie

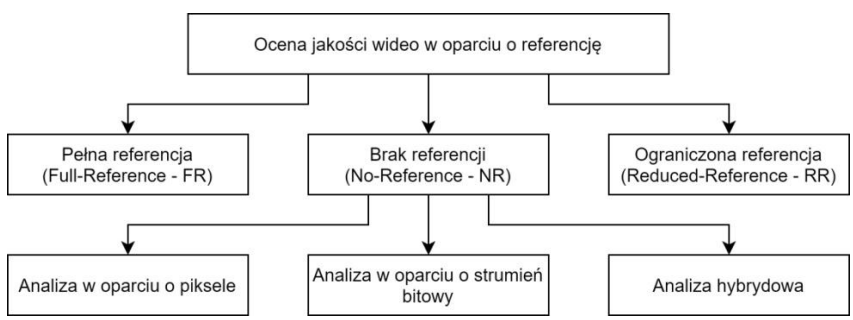
Globalny ruch Internetowy z roku na rok wzrasta, w 2016 roku wynosił on około 60 Tbit/s, natomiast w roku 2020 było to już około 170 Tbit/s. Ze względu na pandemię COVID-19 wzrost światowego ruchu Internetowego był znaczny i między latami 2019 oraz 2020 wyniósł około 50 Tbit/s [1]. Szacuje się, że około 57% światowego ruchu w dół (downstream) jest związane z treścią wideo, a w stosunku do roku poprzedniego wzrost konsumpcji wyniósł ponad 2% [2]. Na urządzeniach mobilnych, które według Cisco stanowią około 25% wszystkich podłączonych urządzeń do Internetu, udział zawartości wideo jest większy i wynosi około 65% [3, 4]. Nie bez znaczenia jest także standardowa telewizja. Liczba podłączonych urządzeń TV to około 13% [4], czyli znacznie mniej niż urządzeń mobilnych, jednakże prognozuje się, że do 2023 dwie trzecie zainstalowanych telewizorów będzie w stanie odtwarzać treść 4K i jest to dwukrotny wzrost w stosunku do 2018 roku [4]. Rosnący udział treści wideo wymaga, aby dostawcy Internetowi wdrożyli mechanizmy oraz polityki zapewniające odpowiedni poziom jakości usług. Dotyczy szczególnie usług telewizyjnych, które świadczone są przez operatorów w oparciu o zasoby własnej sieci. Usługi te muszą konkurować z aplikacjami over-the-top (OTT), takimi jak YouTube czy Netflix. Warto podkreślić, że aplikacje OTT w czasie pierwszych miesięcy pandemii COVID-19 wygenerowały razem około 33,5% całkowitego ruchu Internetowego w obu Amerykach, a w regionie EMEA około 22,7% [3]. W tym czasie operatorzy usług telewizyjnych mogli odnotować nieco zmniejszony ruch, związany głównie z brakiem znaczących wydarzeń sportowych [3]. Sytuacja ta może się jednak w najbliższy czasie zmienić z powodu transmisji wydarzeń związanych z Mistrzostwami Europy w piłce nożnej oraz Igrzyskami Olimpijskimi w Tokio. Przewiduje się, że w drugim oraz trzecim kwartale 2021 roku ruch sieci dosyłowej oraz operatorskiej wzrośnie, szczególnie tam, gdzie usługi telewizyjne są świadczone w technologii unicastowej. Zwiększający się ruch sieciowy powoduje, że operatorzy powinni zweryfikować wykorzystywane metody zapewniające odpowiedni poziom satysfakcji u użytkowników końcowych.

2. Ocena obiektywna i subiektywna

Metody szacowania jakości audiowizualnej oraz jej degradacji mogą być kategoryzowane jako subiektywne, opierające się na percepcji obserwatora, który, oglądając treść multimedialną, może ocenić jej jakość oraz obiektywne, bazujące przykładowo na

[1] pawel.andruloniw@doctorate.put.poznan.pl, Instytut Sieci Teleinformatycznych, Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, https://cat.put.poznan.pl/.
[2] piotr.zwierzykowski@put.poznan.pl, Instytut Sieci Teleinformatycznych, Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, https://cat.put.poznan.pl/.

pomiarach parametrów sygnału lub sieci. Kontrolowana ocena subiektywna często jest wykonywana w pięciostopniowej skali ACR (*absolute category rating* – ACR) wprowadzonej w rekomendacji ITU-T[3] P.910 [8]. Najgorsza jakość jest określana przez ocenę jeden, natomiast najlepsza jako wartość pięć. Obliczenie średniej ze wszystkich ocen obserwatorów skutkuje uzyskaniem oceny MOS (*mean opinion score* – MOS). Ocena w taki sposób dość dobrze oddaje to, jaka jest faktyczna jakość oferowanej treści i czy końcowi użytkownicy są z niej zadowoleni. Istotnym ograniczeniem tej metody jest opóźnienie w uzyskaniu oceny, przez co nie nadaje się do systemów i algorytmów, które działają w czasie rzeczywistym. W takich przypadkach wykorzystywane są techniki określane jako obiektywny lub predykcyjny MOS, pozwalające na numeryczne uzyskanie podobnej oceny do tej uzyskanej subiektywnie [5]. Metody obiektywne zazwyczaj dzielone są na trzy kategorie: *pełna-referencja* (*full-reference* – FR), w której pomiar ewentualnych zniekształceń treści jest oparty na odniesieniu się do źródła i porównaniu sygnału wejściowego z wyjściowym; *zredukowana-referencja* (*reduced-reference* – RR), działająca podobnie jak ocena FR, tyle że pod uwagę nie jest brana całość sygnału, a konkretnie wybrane cechy, oraz *brak-referencji* (*no-reference* – NR), czyli ocena, która nie opiera się na porównaniu z referencją, lecz bazuje na predykcji w oparciu o m.in. cechy sygnału, parametry sieci, czy też zachowania odtwarzacza. Ze względu na odwołanie się do źródła, metody FR i RR dość dobrze oceniają jakość treści wideo lub audio, jednakże mogą być trudne do wykorzystania w systemach monitoringu, gdy niezbędna jest natychmiastowa ocena i może być konieczna adaptacja całego systemu do nagłych zmian warunków. W ten przypadek bardzo dobrze wpisuje się ocena NR, która nie potrzebuje punktu odniesienia, a wykorzystanie efektywnych algorytmów i technik uczenia maszynowego, ze względu na dużą liczbę cech i danych do analizy, może pozwolić na uzyskanie wydajnego i szybkiego sposobu na ocenę jakości, przede wszystkim treści wideo, która jest główną składową ocenianego sygnału.



Rysunek 1. Klasyfikacja oceny jakości wideo dla przypadku braku referencji [opracowanie własne]

Biorąc pod uwagę parametry sieciowe, zdarzenia występujące podczas odtwarzania treści oraz telemetrię dostępną na urządzeniach klienckich, należy wziąć pod uwagę metody oparte o analizę strumienia bitowego oraz metody hybrydowe, które są przykładem metod NR (rys. 1). W tym artykule koncentrujemy się na sieciach operatorskich i dlatego, tylko te metody będą brane pod uwagę w dalszej analizie. Analiza

[3] ITU-T to sektor normalizacji telekomunikacji działający w ramach Międzynarodowego Związku Telekomunikacji (ITU).

w oparciu o strumień bitowy polega na predykcji jakości wideo na podstawie danych powiązanych np. z czasem przepływu pakietów w sieci, wielkością pakietów czy też ich utratą, a w przypadku usług oferujących treść jako unicastowy strumień adaptacyjny ABR (*adaptive bitrate streaming* – ABR), wykorzystujących standard MPEG-DASH (*Dynamic Adaptive Streaming over Http* – DASH) lub HLS (*HTTP Live Streaming* – HLS) mogłyby to być np. wielkości chunków[4] lub przepływność strumienia, jaki jest oferowany. W analizie wspomnianych metryk głównym problemem jest szyfrowanie, które powoduje, że ocena niektórych parametrów w sieci może nie być możliwe lub jest tylko częściowo możliwe dzięki głębokiej inspekcji pakietów DPI (*deep packet inspection* – DPI). Istnieją jednakże algorytmy, oparte głównie o metody uczenia maszynowego, które potrafią efektywnie analizować także zaszyfrowane dane. Analiza hybrydowa pozwala na uwzględnienie zdarzeń, które występują podczas odtwarzania treści. Mogą to być m.in. zacięcia podczas odgrywania treści; zmiany poziomu przepływności bitowej i czas, podczas którego nie był dostępny np. najwyższy profil[5]; opóźnienie w rozpoczęciu odtwarzania oraz inne dane telemetryczne dostępne z poziomu odtwarzacza lub dekodera. W ogólnym przypadku obiektywna ocena jakości sygnału oraz wydajności sieci i aplikacji określana jest jako jakość usług Quality of Service (*Quality of Service* – QoS) i opiera się na analizie danych mierzalnych, takich jak m.in. opóźnienie, zmienność opóźnienia, czy też liczba utraconych pakietów. Sposób, w jaki użytkownik postrzega treść wideo określany jest przez jakość doświadczeń Quality of Experience (*Quality of Experience* – QoE). ITU-T (*International Telecommunication Union Telecommunication Standardization Sector* – ITU-T) określa QoE jako ogólną akceptowalność usługi, ocenioną subiektywnie przez użytkownika końcowego [8]. Poziom QoE może być uzyskany przez subiektywną ocenę wykonaną przez obserwatora otrzymanej treści multimedialnej. Część oceny jest dokonana przez pryzmat behawioralny i obejmuje nawyki, wymagania, osobowość, oczekiwania użytkownika, a nawet samopoczucie. Założeniem Quality of Experience jest ocena subiektywna, natomiast dążenie do ulepszania systemów oraz dostarczania lepszej jakości usług audio i wideo wymusiło powstanie obiektywnej miary QoE, indukując tworzenie coraz to nowych metod i algorytmów, które biorąc pod uwagę określoną liczbę parametrów oraz ich skomplikowanie generują numerycznie, obiektywną ocenę Quality of Experience. Detekcja błędów w sieci, może być wykonana w sposób proaktywny, w trakcie którego analizowane są dane z sieci operatora, a następnie wykorzystane do trenowania modeli, których celem jest wykrycie błędów wpływających na poziom Quality of Experience u klientów końcowych, na podstawie których może być wykonana adaptacja systemu i polepszenie jakości. W reaktywnym podejściu analiza oparta jest o jakość QoE u klienta końcowego i w przypadku spadku wspomnianego poziomu system określa, na którym etapie transportu sygnału wystąpił błąd [9]. Takie szacowanie może być wykorzystane do monitoringu systemów dostarczania treści, adaptacji systemów do aktualnie panujących warunków lub do wczesnego wykrywania degradacji sygnału.

[4] W strumieniowaniu adaptacyjnym treść wideo jest dzielona na segmenty zwane też chunkami. Są to kilkusekundowe części treści wideo, pobierane przez odtwarzacz i przesyłane do bufora w celu późniejszego odtworzenia.

[5] Treść oferowana w strumieniowaniu adaptacyjnym charakteryzuje się dostępnością więcej niż jednego strumienia zwanego profilem. Każdy z nich posiada zazwyczaj inną przepływność oraz może posiadać inną rozdzielczość wideo.

3. Przegląd algorytmów Sztucznej Inteligencji wykorzystania do wykrywania błędów oraz oceny jakości usług wideo

Sieć MLP (*Multilayer Perceptron* – MLP) oraz rekurencyjna sieć LSTM (*Long short-term memory* – LSTM) może być wykorzystana do detekcji błędów w usługach opartych o strumieniowanie adaptacyjne, takich jak Netflix, YouTube, Amazon Prime i innych oraz usługach telekonferencyjnych Zoom oraz Skype. Wykorzystanie danych telemetrycznych zbieranych po stronie klienta pozwoliło na stworzenie systemu, pozwalającemu dostawcy treści na detekcję błędów niezależnie od sieci operatorskich i trasy wiodącej często przez wiele systemów autonomicznych AS, którą pokonują pakiety, aby dotrzeć do odbiorcy. Zestaw danych wykorzystany do detekcji błędów zawierał informacje o czasie odtwarzania treści, długości oraz częstotliwości buforowania, średniej przepływności bitowej, poziomie zadowolenia ocenionym przez zewnętrzny algorytm dostarczony przez firmę Nice People at Work (NPAW) oraz obliczonym poziomie QoE. Uzyskane dane zostały oznaczone jako: brak problemu, problem po stronie klienta oraz problem po stronie operatora sieci. Najlepsze wyniki w detekcji błędów oraz detekcji lokalizacji dla modelu MLP zostały uzyskane z jedną warstwą wejściową sieci neuronowej, trzema ukrytymi i jedną wyjściową. Model LSTM został zaimplementowany z jedną warstwą wejściową sieci neuronowej, czterema ukrytymi oraz warstwą w pełni połączoną jako wyjście. Najlepszą skuteczność w wykrywaniu błędów wynoszącą 97,1% udało się uzyskać za pomocą metody MLP, sieć LSTM była o 0,5% mniej celna. Otrzymane wyniki zostały porównane z wynikami dla metody wektorów nośnych oraz drzewa decyzyjnego. Dla tych metod udało się uzyskać skuteczność odpowiednio 89,7% oraz 90,9% [9].

Dimopoulos et al. skupił się na analizie degradacji poziomów Quality of Experience na podstawie zaszyfrowanego ruchu, związanego z mobilną aplikacją YouTube. Głównymi czynnikami branymi pod uwagę w ocenie były: liczba zacięć w czasie odtwarzania, zmiany rozdzielczości wideo oraz częstotliwość występowania tych zmian. Detekcja takich zdarzeń została oparta o takie cechy, jak: czas RTT (*Round trip time* – RTT), iloczyn przepustowości i opóźnienia BDP (*Bandwidth-delay product*), liczba bajtów w czasie, procent utraty pakietów, procent retransmisji pakietów oraz obliczony rozmiar i czas przybycia chunka. Do wykrywania wysyceń bufora w czasie odtwarzania wykorzystano algorytm losowego lasu decyzyjnego oraz 10-krotną walidację krzyżową. Wykrycie zacięcia jest ściśle powiązane ze zmianą rozmiaru chunków. W takim przypadku odtwarzacz pobiera mniejsze części, które mogą być szybko pobrane i załadowane do bufora. Wraz z poziomem napełnienia bufora rozmiar chunków wzrastał. W celu wykorzystania uczenia nadzorowanego utworzono etykiety oznaczające poziom występowania ponownego buforowania w czasie odtwarzania. Do uczenia wykorzystano ruch nieszyfrowany i dla tego przypadku średnia dokładność wynosiła 93,5%. Zastosowanie tego modelu do danych szyfrowanych pozwoliło na uzyskanie dokładności na poziomie 91,8%. W podobny sposób uzyskano predykcję odtwarzanej rozdzielczości wideo. Średnia dokładność w ocenie odtwarzanej rozdzielczości była na poziomie 84,5%. Było to 2,5% więcej niż dla ruchu szyfrowanego. Analiza częstotliwości zmian rozdzielczości podczas odtwarzania pojedynczego wideo określona jako wskaźnik PQSR (*Presentation Quality Switch Rate* – PQSR) została przeprowadzona metodą analizy szeregów czasowych. Okazało się, że metoda CUSUM (*Cumulative Sum Control Chart* – CUSUM) to dokładniejszy sposób na wykrycie częstotliwości zmian

niż metody uczenia maszynowego. Dokładność wykrycia sesji wideo bez wahań oglądanej rozdzielczości było na poziomie 78%. Niestety, autorzy wspomnianej pracy nie podali wyników dla metod uczenia maszynowego [10].

Analiza ruchu szyfrowanego została także przeprowadzona przez Belmoukadam et al. W tym przypadku na podstawie estymacji długości chunków MPEG-DASH, analizy przepływności sieci oraz czasu między przybyciem kolejnych pakietów udało się określić rozdzielczość oglądanego przez użytkownika filmu. Zastosowanie modelu mieszanki Gaussa pozwoliło na klasyfikację pakietów, w zależności od tego, czy zawierały one informacje dotyczące dźwięku, czy też obrazu. W przypadku audio były to pakiety zawierające chunki mniejsze niż 750 KB i czas pobierania mniejszy niż dwie sekundy. Dla wideo były to pozostałe pakiety z wyłączeniem tych związanych z wysłaniem do serwera CDN (*Content Delivery Network* – CDN) żądania o kolejną porcję danych oraz tych dotyczących nawiązania sesji TCP. Model oparty o metodę lasu losowego był w stanie określić jakość wideo z dokładnością około 85% dla miary F1. Liniowa analiza dyskryminacyjna była o około 5% mniej dokładna niż wcześniejsza metoda. W tym przypadku najgorzej poradziła sobie metoda regresji logistycznej z wynikiem około 68%. Wytrenowany model pozwolił także na ocenę jakości, bazując na części danych. Po 24 sekundach odtwarzania, klasyfikacja została dokonana prawidłowo w 80% przypadków, po 72 sekundach wynik był o 2% lepszy, dążąc do 85% po 120 sekundach. Adaptacja podobnego modelu do innych usług wideo, także usługi telewizyjnej, w której strumieniowanie oparte jest o protokół MPEG-DASH, pozwoliłoby operatorom na zapewnienie odpowiedniej przepływności w sieci, a także sterowanie priorytetyzacją ruchu w taki sposób, aby zawsze uzyskiwać jak najlepszą jakość u klienta końcowego [11].

System nazwany YouQ także skupia się na analizie ruchu szyfrowanego. W tym przypadku ocena QoE jest podzielona na trzy klasy: wysoka (najlepsza jakość), średnia oraz niska (najniższa jakość) i była określona na podstawie liczby zacięć, średniego czasu trwania zacięcia oraz czasu spędzonego na danej rozdzielczości wideo. Do wytrenowania modeli wykorzystano zestaw 33 cech powiązanych z przepływem ruchu szyfrowanego. Najważniejszymi z nich były m.in. średni rozmiar pakietu, liczba transferowanych bajtów w ciągu 5 sekund, przepływność w ciągu pierwszych kilku sekund od rozpoczęcia sesji, średni czas pomiędzy przybyciem kolejnych pakietów, czy też mediana czasu przybycia kolejnych pakietów dla 5-sekundowych interwałów czasowych obejmujących całkowitą sesję. Wykorzystanie metody lasu losowego pozwoliło na osiągnięcie około 80% skuteczności w prawidłowym oznaczaniu klasy QoE na podstawie zaszyfrowanego ruchu. Dodatkowo zauważono, że krótsze czasy pomiędzy przybyciem kolejnych pakietów oznacza zazwyczaj wyższy poziom QoE [12].

Na podstawie zaszyfrowanego ruchu istnieje możliwość oceny metryk związanych bezpośrednio z oceną jakości wideo, takich jak czas potrzebny na rozpoczęcie odtwarzania dla strumieniowania adaptacyjnego oraz ocena występowania zdarzeń związanych z wysyceniem bufora tj. pojedynczego zacięcia, wielu zacięć oraz długiego zacięcia trwającego dłużej niż określony poziom. Rozpatrując cechy związane z prędkością pobierania, takie jak m.in.: prędkość średnia, maksymalna, odchylenie standardowe, liczbę przesłanych danych, czy też czas największego ruchu oszacowano, wykorzystując model oparty o las losowy, opóźnienie w rozpoczęciu odtwarzania z dokładnością 80% oraz błędem bezwzględnym mniejszym niż jedna sekunda. Wykorzystując ten sam

zestaw cech, wytrenowano model oparty o algorytm wstecznej propagacji błędu, który dla 80% filmów był w stanie ocenić, jaka jest aktualna przepływność strumienia wideo z błędem bezwzględnym poniżej 300 Kbps. Algorytm lasu losowego poradził sobie także z estymacją zdarzeń związanych z wysyceniem bufora. Z 88,6% dokładnością wykryto co najmniej jedno zacięcie materiału wideo. W 92,2% przypadków wykryto więcej niż jedno zacięcie, natomiast 91,7% dokładności charakteryzowało próbę wykrycia zacięcia dłuższego niż określony poziom określony w sekundach. Ocena opisanych zdarzeń możliwa jest na poziomie bramy SGW w sieci szkieletowej ISP [13].

Szacowanie jakości wideo dla aplikacji YouTube, bazującej na strumieniowaniu adaptacyjnym na podstawie ruchu w sieci Wi-Fi dla standardów 802,11 g/n/ac/ax dla częstotliwości 2.4 GHz oraz 5 GHz przeprowadził Morshedi et al. [14]. Pod uwagę wzięto występowanie takich zdarzeń, jak opóźnienie w rozpoczęciu odtwarzania, czyli czas potrzebny do napełnienia bufora, rozdzielczość odtwarzanej treści wideo, liczbę zacięć związanych z wysyceniem bufora oraz czas trwania zacięć. Cztery zdarzenia pozwoliły na obliczenie MOS, w skali od 1 (najgorsza jakość) do 5 (najlepsza jakość) za pomocą wzoru (1):

$$MOS = 4.23 - 0.0672L_{ti} - 0.742L_{fr} - 0.106L_{tn} - R \qquad (1)$$

gdzie: MOS – obiektywna ocena jakości, L_{ti} – czas rozpoczęcie odtwarzania, L_{fr} – liczba zacięć w trakcie odtwarzania, L_{tn} – czas trwania zacięć, R – współczynnik wpływu dla zmian rozdzielczości

L_{ti}, L_{fr}, L_{tn} oraz R oznaczały odpowiednio czas rozpoczęcia odtwarzania, liczbę zacięć w trakcie odtwarzania, czas trwania zacięć oraz współczynnik wpływu na MOS dla wybranych rozdzielczości. Przykładowo dla spadku rozdzielczości z poziomu 1080p do 144p wskaźnik ten wynosił 2,07, natomiast dla spadku do 720p wartość to 0,21. Obliczony MOS został zaokrąglony do najbliższej liczby całkowitej, a następnie wykorzystany jako etykiety dla algorytmów uczenia nadzorowanego, które do uczenia wykorzystały zestaw cech związanych bezpośrednio z siecią Wi-Fi, zebranych głównie z urządzeń dostępowych AP (*Access Point* – AP). Liczba cech była zmienna i przykładowo dla sieci 802.11n działającej na częstotliwości 2.4 Ghz zebrano 240 cech, a dla 802.11ax i częstotliwości 5 GHz było to już 299 cech. W celu zbadania możliwości szacowania MOS na podstawie parametrów sieci 802.11 wykorzystano 7 algorytmów uczenia maszynowego: ZeroR, J48 (algorytm C4.5), LMT (*Logistic Model Tree* – LMT), NBTree (*Naive Bayes Tree* – NBTree), RIDOR (*Ripple Down Rule learner* – RIDOR) oraz perceptron wielowarstwowy MLP (*Multilayer Perceptron* – MLP). Najlepiej poradził sobie algorytm LMT nie tylko ze względu na celność przewidywań, która wynosiła od 93 do 99% w zależności od sieci, ale także ze względu na niskie wykorzystanie CPU w czasie szacowania oceny. Taka właściwość pozwala na zaimplementowanie modelu na urządzeniu AP i wykonanie oceny bezpośrednio na urządzeniu. System monitorujący takie wartości pozwoliłby operatorom na proaktywne wykrywanie i usuwanie degradacji oraz usterek u klientów końcowych [14].

W usługach wideo na życzenie spotykana jest prośba o ocenę obejrzanej treści. Często jest ona w skali od 1 do 5, gdzie 1 określa najsłabszy odbiór, a 5 najlepszy. Ocena taka może zostać wykorzystana jako etykieta do zbioru danych, na których operuje algorytm uczenia maszynowego. Połączenie takiej oceny oraz metryk zebranych z poziomu sieci i aplikacji pozwoliło na stworzenie modelu, który na podstawie oceny pojedynczych użytkowników szacował ogólną kondycję systemu [15]. Informacje,

takie jak całkowity czas oglądania, liczba zacięć, liczba porzuconych ramek, całkowity czas potrzebny na ponowne buforowanie oraz subiektywna ocena treści przez użytkownika, pozwoliły na określenie wydajności systemu wideo opartego o strumieniowanie adaptacyjne i oglądanie treści wideo z poziomu przeglądarki oraz aplikacji mobilnej. Zebrane dane zostały wykorzystane do wytrenowania modeli opartych o algorytm metody wektorów nośnych (*Support Vector Machine* – SVM), k-najbliższych sąsiadów (*K-Nearest Neighbors* – KNN) oraz sztuczną sieć neuronową (*Artificial Neural Network* – ANN), które potrafią estymować ocenę wydaną przez użytkownika. Udało się także powiązać opóźnienie aplikacji oraz przepływność z oceną całościową systemu wideo (2):

$$Q_s(S_{gp}, S_l) = 2.205 + 1.01 \cdot S_{gp} + 0.6451 \cdot S_l + 0.7613 \cdot S_{gp}^2 - 0.1645 \cdot S_{gp} \cdot S_l - 0.8037 \cdot S_{gp}^3 - 0.2947 \cdot S_{gp}^2 \cdot S_l \quad (2)$$

gdzie: $Q_s(S_{gp}, S_l)$ – całościowa ocena systemu wideo, S_{gp} – szybkość bitowa mierzona na poziomie aplikacji w Mbit/s (goodput), S_l – opóźnienie usługi w milisekundach

Transmisja multicastowa obecna w sieciach telewizyjnych opartych na protokole IP (*Internet Protocol Television* – IPTV), może być też wykorzystana do transmisji na żywo np. zdalnych lekcji czy koncertów. Dodatkowo w ten sposób przesyłany może być sygnał od nadawcy do operatora świadczącego usługi telewizyjne zarówno IPTV, jak i w technologii unicast. Do oceny QoE można wykorzystać w tym przypadku metryki zarówno sieciowe, takie jak: opóźnienie, zmienność opóźnienia pakietów (jitter) oraz liczbę utraconych pakietów, parametry wideo: parametry kodera, liczbę klatek na sekundę, rozdzielczość czy też przepływność strumienia, a także dane dotyczące urządzenia, z którego korzysta użytkownik. Na podstawie takich danych wytrenowano model oparty na metodzie LASSO (*Least Absolute Shrinkage and Selection Operator* – LASSO), który potrafił ocenić jakość MOS. Zauważono, że opóźnienie nie wpływa znacząco na ocenę subiektywną. Największy wpływ na ocenę miał jitter oraz utrata pakietów, szczególnie gdy pakiety były tracone w dużych ilościach w danej chwili [16]. Wytrenowany model został dodatkowo wdrożony do mechanizmu, który pozwalał na adaptację parametrów systemu wideo, w zależności od obliczonych wartości określających jakość wideo.

Szacowanie jakości wideo możliwe jest także w sieciach programowalnych (*Software Defined Network* – SDN). Na podstawie podobnych cech jak poprzednio, tj. liczby utraconych pakietów, przepływności pakietów, czasu kolejkowania pakietów, czasu ponownego buforowania, przepływności strumienia wideo, początkowego opóźnienia, czy też liczby zacięć wytrenowano model oparty o regresję liniową, który oceniał jakość MOS dla strumieniowania adaptacyjnego w technologii MPEG-DASH. Przy ocenie niższej niż 4 założono, że jakość nie jest akceptowalna. Na tej podstawie system zmieniał swoje parametry w taki sposób, aby podnieść ocenę MOS. Zauważono, że w tym przypadku znaczny wpływ na ocenę był związany z utratą lub opóźnieniem pakietów, a wpływ wystąpienia zacięcia jest inny w zależności od odgrywanej treści, przykładowo dla usługi strumieniującej szkolenie online, użytkownicy gorzej oceniali sesję z zacięciami pod koniec filmu niż na jego początku [17].

4. Dyskusja i wnioski

Wykorzystując metody uczenia maszynowego, możliwa jest ocena jakości treści wideo w sieciach telekomunikacyjnych, zarówno po stronie operatora, jak i po stronie klienta. Badania pokazują, że zarówno w sieci Wi-Fi, jak i Ethernet można z powodzeniem szacować jakość oglądanej przez widzów treści. Istotne jest to, że taka ocena może być wykonana na podstawie szyfrowanego ruchu i w sposób, który nie naraża prywatności oglądających. Dla operatorów świadczących usługi Internetu oraz telewizji ocena jakości wideo może być bardzo istotna m.in. ze względu na możliwość określenia wskaźnika migracji klientów (*churn*). Niezadowolony z usługi klient jest bardziej chętny do rezygnacji z usługi. Integracja wybranych algorytmów i modeli z aktualnie działającymi systemami telewizyjnymi może pozwolić na stworzenie systemów, które są w stanie w czasie rzeczywistym adaptować się do aktualnie panujących warunków i świadczyć usługę na lepszym poziomie, jednocześnie podnosząc poziom zadowolenia klientów. Na podstawie zebranych ocen można wyznaczyć parametr PDU (3) (*Percentage of Dissatisfied Users* – PDU), który całościowo może określić liczbę użytkowników mniej zadowolonych z usługi [18].

$$PDU = \frac{\sum(OS<th)}{N} \cdot 100 \tag{3}$$

gdzie: PDU – procent niezadowolonych klientów, OS – ocena uzyskana od użytkownika oglądającego treść wideo, th – poziom, poniżej którego określono ocenę jako brak satysfakcji, N – liczba zebranych ocen użytkowników

Przykładowo dla ocen: 5, 4, 3, 2, 4, 2, 4, 5, zakładając, że całkowity brak satysfakcji to ocena 2 i mniej możemy określić, że 25% użytkowników nie jest zadowolonych z oglądanej treści. Szybkie wykrywanie degradacji i niwelowanie jej także powinno znacznie zwiększyć zadowolenie użytkowników. Według raportu Akamai 5-sekundowe opóźnienie w rozpoczęciu odtwarzania może skutkować zaprzestaniem oglądania w prawie 25% przypadków, prawie połowa oglądających przestanie oglądać, jeżeli opóźnienie to wzrośnie do 10 sekund. Podobnie jest z buforowaniem treści w czasie oglądania, jeżeli wideo zatrzyma się na 2 sekundy, 25% oglądających chętniej przerwie oglądanie, czas przerwy rosnący do 5 sekund może spowodować przerwanie oglądania w 50% przypadków [19]. Ciągłe ulepszanie systemów monitorujących oraz wykorzystanie mechanizmów uczenia maszynowego powinno pomóc w uniknięciu lub przynajmniej zmniejszeniu liczby zdarzeń, które negatywnie mogą wpływać na jakość postrzegania treści wideo przez klientów. Takie podejście do rozwijania usług powinno zapewnić większe zadowolenie z otrzymanych treści.

Podziękowania

Autorzy dziękują Ministerstwu Edukacji i Nauki za wsparcie finansowe (program Doktorat Wdrożeniowy, nr DWD/4/24/2020).

The authors thank the Polish Ministry of Education and Science for financial support (Applied Doctorate Program, No. DWD/4/24/2020).

Literatura

1. TeleGeography, *Global Internet Map 2021*, February 2021, Dostęp online: https://global-internet-map-2021.telegeography.com (dostęp 29 maja 2021).
2. Sandvine, *The Global Internet Phenomena Report COVID-19 Spotlight. May 2020*. Dostęp online: https://www.sandvine.com/covid-internet-spotlight-report (dostęp 29 maja 2021).
3. Sandvine, *The Mobile Internet Phenomena Report*, February 2020, Dostęp online: https://www.sandvine.com/download-report-mobile-internet-phenomena-report-2020-sandvine (dostęp 29 maja 2021).
4. Cisco, *Cisco Annual Internet Report (2018-2023), March 2020*, Dostęp online: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html (dostęp 29 maja 2021).
5. Akhtar Z., Falk T.H., *Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey*, in IEEE Access, vol. 5, pp. 21090-21117, 2017, doi: 10.1109/ACCESS.2017.2750918.
6. Akhtar Z., Siddique K., Rattani A., Lutfi S.L., Falk T.H., *Why is Multimedia Quality of Experience Assessment a Challenging Problem?*, in IEEE Access, vol. 7, s. 117897-117915, 2019, doi: 10.1109/ACCESS.2019.2936470.
7. ITU-T, *Quality of experience requirements for IPTV services*, Recommendation ITU-T G.1080.
8. ITU-T, *Subjective video quality assessment methods for multimedia applications*, Recommendation ITU-T P.910.
9. Dinaki H.E., Shirmohammadi S., Janulewicz E., Côté D., *Deep Learning-Based Fault Localization in Video Networks Using Only Client-Side QoE*, in IEEE Transactions on Artificial Intelligence, vol. 1, no. 2, s. 130-138, Oct. 2020, doi: 10.1109/TAI.2020.3041816.
10. Dimopoulos G., Leontiadis I., Barlet-Ros P., Papagiannaki K., *Measuring Video QoE from Encrypted Traffic*, in Proceedings of the 2016 Internet Measurement Conference (IMC '16). Association for Computing Machinery, USA, , New York 2016, s. 513-526, doi: 10.1145/2987443.2987459.
11. Belmoukadam O., Barakat C., *From Encrypted Video Traces to Viewport Classification*, 2020 16th International Conference on Network and Service Management (CNSM), 2020, s. 1-9, doi: 10.23919/CNSM50824.2020.9269086.
12. Orsolic I., Pevec D., Suznjevic M. et al., *A machine learning approach to classifying YouTube QoE based on encrypted network traffic*, Multimed Tools Appl 76, 2017, s. 22267-22301, doi: 10.1007/s11042-017-4728-4.
13. Tang S., Qin X., Wei G., *Network-Based Video Quality Assessment for Encrypted HTTP Adaptive Streaming*, in IEEE Access, vol. 6, 2018, s. 56246-56257, doi: 10.1109/ACCESS.2018.2872932.
14. Morshedi M., Noll J., *Estimating PQoS of Video Streaming on Wi-Fi Networks Using Machine Learning*, Sensors 2021, 21, 621, doi: 10.3390/s21020621.
15. Bulkan U., Dagiuklas T., *Predicting quality of experience for online video service provisioning*, Multimed Tools Appl 78, 2019, s. 18787-18811, doi: 10.1007/s11042-019-7164-9.
16. Taha M., Canovas A., Lloret J. et al., *A QoE adaptive management system for high definition video streaming over wireless networks*, Telecommun Syst 77, 2021, s. 63-81, doi:10.1007/s11235-020-00741-2.
17. Letaifa A.B., *Real Time ML-Based QoE Adaptive Approach in SDN Context for HTTP Video Services*, Wirel. Pers. Commun. 103, 3 (December 2018), s. 2633-2656, doi: 10.1007/s11277-018-5952-6.
18. Mocanu D.C., Pokhrel J., Garella P., Seppänen J., Liotou J., Narwaria E.M., *No-reference video quality measurement: added value of machine learning*, Journal of Electronic Imaging, 2015, 24(6), [061208], doi: 10.1117/1.JEI.24.6.061208.
19. Akamai, *How Akamai Defines and Measures Online Video Quality*, 2016, dostęp online: https://www.akamai.com/us/en/multimedia/documents/content/white-paper/how-akamai-defines-and-measures-online-video-quality-white-paper.pdf (dostęp 14 lipca 2021).

W jaki sposób metody uczenia maszynowego mogą pomóc w wykrywaniu degradacji i szacowaniu jakości usług wideo w sieciach telekomunikacyjnych?

Streszczenie

Szacuje się, że około 57,5% ruchu w sieci Internet związane jest ze strumieniowaniem treści wideo. Dynamiczny rozwój szeroko pojętych usług wideo wymusza na dostawcach treści oraz operatorach sieci ciągłe ulepszania narzędzi do detekcji problemów oraz monitoringu jakości wspomnianych usług. Problem ten dotyczy zarówno usług wideo OTT (*over-the-top*), takich jak YouTube czy Netflix, które w 2020 roku w czasie pandemii COVID-19 wygenerowały ponad 25% globalnego ruchu internetowego, jak i nowoczesnych usług telewizyjnych, często także opartych o strumieniowanie adaptacyjne. Monitorowanie degradacji usług, czy też predykcja wystąpień błędów w czasie oglądania treści wideo jest niełatwym zadaniem ze względu na ilość danych oraz metryk, które muszą być wzięte pod uwagę, aby zapewnić odpowiednią jakość usług. W celu uzyskania obiektywnej oceny mogą zostać wykorzystane metody uczenia maszynowego, które, opierając się na danych zebranych z sieci operatora lub danych telemetrycznych dostępnych na urządzeniu końcowym, mogą pomóc w wyznaczeniu cech wpływających na ostateczną ocenę postrzeganej jakości. Na podstawie przeglądu stanu wiedzy określono, że znaczny wpływ na poziom jakości w usługach opartych o strumieniowanie adaptacyjne może mieć wysycenie bufora związane z dostępną przepływnością w sieci, przepływność bitowa strumienia oraz częstotliwość zmian tej przepływności. Wystąpienie opisanych zdarzeń może zostać określone na podstawie analizy ruchu sieciowego. Dowiedziono, m.in. że wykorzystanie metod opartych o sieci neuronowe pozwala na skuteczne wykrywanie miejsca, w których nastąpiła degradacja usług. Ukazano, że modele wytrenowane w oparciu o cechy pozyskane na podstawie ruchu nieszyfrowanego dobrze sobie radzą w predykcji zmian oraz błędów na podstawie danych z ruchu szyfrowanego, a uczenie maszynowe może zostać wykorzystane do wykrycia wysyceń bufora na poziomie danych z warstwy sieciowej oraz transportowej.

Słowa kluczowe: sztuczna inteligencja, uczenie maszynowe, over-the-top, jakość usług, jakość doświadczeń

How can machine learning methods help detect degradation and estimate the quality of video service in telecommunications networks?

Abstract

It is estimated that approximately 57.5% of Internet traffic is related to streaming video content. The dynamic development of broadly understood video services forces content providers and network operators to constantly improve tools for detecting problems and monitoring the quality of these services. This problem concerns both OTT (over-the-top) video services such as YouTube or Netflix, which in 2020 generated over 25% of global internet traffic during the COVID-19 pandemic, and modern TV services, often also based on adaptive streaming technology. Monitoring service degradation or error prediction while watching video content is a difficult task due to the amount of data and metrics that must be taken into account to ensure the right quality of service. To obtain an objective assessment, machine learning methods can be used. Based on data collected from the operator's network or telemetry data available on the end device, mentioned methods can help in determining the characteristics that significantly affect the final assessment of the perceived quality. Based on the review of the state of the art, it was determined that the level of quality in services based on adaptive streaming may be significantly influenced by the buffer saturation related to the available data rate in the network, the bitrate of the stream, and the frequency of bitrate changes. The occurrence of the described events can be determined on the basis of network traffic analysis. It has been proven, inter alia, that the use of methods based on neural networks allows for effective detection of places where service degradation has occurred. It has been shown that models trained on the characteristics obtained from unencrypted traffic perform well in predicting changes and errors based on encrypted traffic data. Moreover, machine learning methods can be used to detect buffer saturation at the level of data from the network and transport layers.

Keywords: artificial intelligence, machine learning, over-the-top, quality of service, quality of experience

Krzysztof Rutkiewicz[1], Julia Płonka[2]

System zarządzania przedsiębiorstwem funkcjonującym na rynku dystrybucji helu ciekłego i gazowego

1. Wprowadzenie

Celem badań zaprezentowanych w tym rozdziale jest ocena systemu zarządzania przedsiębiorstwem, działającym na polskim i międzynarodowym rynku dystrybucji helu ciekłego i gazowego. Przeprowadzone badania koncentrowały się na szczegółowej analizie rynkowej, strategicznej, personalnej, finansowej i procesowej. Wybór badanego przedsiębiorstwa wynikał głównie z posiadania przez firmę szeroko rozbudowanego systemu zarządzania, który dodatkowo skupia się na ciągłym doskonaleniu. Składa się on z mniejszych podsystemów, których zadaniem jest m.in. budowanie, usprawnienie, unowocześnienie i rozwijanie poszczególnych działów struktury organizacyjnej oraz sektorów działań w zakresie procesów w nich zachodzących.

W przypadku analizy rynku, zakres badań obejmował określenie grup odbiorców usług przedsiębiorstwa, dostawców ciekłego helu oraz krajowych i zagranicznych konkurentów. Analiza strategiczna uwzględniała przeprowadzenie analizy ryzyka i szans oraz określenie misji, wizji, strategii i polityki jakości firmy. Ważną częścią badań była diagnoza metod zarządzania personelem, które stanowiły determinantę efektywności funkcjonowania całego przedsiębiorstwa. W przypadku analizy finansowej badaniu poddano wyniki sprzedażowe pochodzące z zamówień publicznych z lat 2010-2017. Zakres ostatniego segmentu dotyczył analizy wewnętrznych procesów zachodzących w przedsiębiorstwie, w tym ich politykę zarządzania i kontrolowania.

Weryfikowaną hipotezą jest twierdzenie, że prawidłowe i efektywne funkcjonowanie systemu zarządzania przedsiębiorstwem na rynku dystrybucji helu ciekłego i gazowego jest determinowane przez szereg parametrów, uwzględniających m.in. konkurencyjność przedsiębiorstwa na rynku, optymalną wartość wskaźników finansowych, poziom zadowolenia klientów, efektywność wykonywanej pracy, nieustanne doskonalenie wszystkich działań i procesów w organizacji oraz ciągłą adaptację do zachodzących w otoczeniu zmian.

W celu dokonania analizy systemu zarządzania przedsiębiorstwem posłużono się triangulacją [1] polegającą na gromadzeniu informacji teoretycznych i empirycznych z różnych metod badawczych, porównywaniu i uogólnieniu końcowych wyników. W tym kontekście zastosowano metody: wywiadu indywidualnego [2, 3] z właścicielem i pracownikami poszczególnych działów firmy, jawnej obserwacji bezpośredniej [4] procesów zachodzących w organizacji oraz krytyczno-poznawczej analizy dostępnych materiałów pierwotnych (zaczerpniętych z firmy) i wtórnych (ze źródeł literaturowych

[1] krzysztof.rutkiewicz@upwr.edu.pl, Katedra Ekonomii Stosowanej, Wydział Przyrodniczo-Technologiczny, Uniwersytet Przyrodniczy we Wrocławiu, https://upwr.edu.pl/.
[2] juliakrol33@gmail.com, Katedra Ekonomii Stosowanej, Wydział Przyrodniczo-Technologiczny, Uniwersytet Przyrodniczy we Wrocławiu, https://upwr.edu.pl/.

i baz danych elektronicznych). Zbiór formalnych danych na temat działalności badanego przedsiębiorstwa uzyskano z Centralnej Ewidencji i Informacji o Działalności Gospodarczej [5]. Część informacji dotyczących procesów administracyjnych, strategicznych i procesów technicznych związanych z helem otrzymano z księgi jakości, instrukcji, kart charakterystyki helu ciekłego i gazowego oraz z oficjalnej strony internetowej firmy. Zakres finansowej analizy działalności przedsiębiorstwa obejmował dane ze sprawozdań z krajowych przetargów publicznych przeprowadzanych w latach 2010-2017.

2. Ogólna teoria systemów i analiza systemu zarządzania przedsiębiorstwem – przegląd literatury

Ogólna teoria systemów jest uniwersalną teorią, która ma zastosowanie do każdego systemu występującego w przyrodzie, w społeczeństwie i w wielu dyscyplinach nauki, w których zjawiska są traktowane holistycznie [6]. Podejście systemowe, zainicjowane przez Ludwiga von Bertalanffy'ego [7] i Norberta Wienera [8], twórcę cybernetyki, stało się popularne w latach 50. XX wieku. Jego istotą jest traktowanie przedsiębiorstw jako systemów otwartych, składających się z zestawów elementów, które tworzą wyróżniającą się całość w otoczeniu [9, 10]. Jedna z najpopularniejszych definicji traktuje system jako zbiór elementów (rzeczy, obiektów, komponentów) posiadających właściwości (atrybuty), a elementy te są ze sobą powiązane [11]. Teoria systemów opiera się zatem na przesunięciu uwagi z części na całość [12, 13] i zakłada dialog między holizmem a redukcjonizmem.

Postrzegając organizację jako system, można wyróżnić cztery elementy:

- komponenty wejścia (nakłady) – w postaci materiałów, zasobów ludzkich, finansowych i informacyjnych zaczerpniętych przez organizację ze środowiska [14, 15];
- procesy transformacji – obejmujące procesy techniczne i kierownicze, dzięki którym nakłady są przekształcane w wyniki [16];
- wyniki – czyli produkty i usługi, zyski lub straty, informacje i zachowania pracowników [17];
- sprzężenie zwrotne – uzyskiwanie informacji o stanie środowiska, w tym o potencjalnych zakłóceniach, co pozwala na zmiany procesów transformacji [18, 19].

Według H.J. Leavitta [20] organizacja jest uporządkowanym systemem złożonym z czterech podsystemów:

- celów i zadań realizowanych w organizacji;
- ludzi wraz z ich indywidualnymi i zbiorowymi dążeniami oraz wzorcami zachowań;
- wyposażenia materialnego i technologicznego oraz określonych zasad posługiwania się nimi;
- struktury formalnej, czyli przyjętych zasad podziału zadań i odpowiedzialności [21].

Biorąc pod uwagę realne podejście systemowe G.M. Golinelli [22] i S. Barile [23] zaproponowali dwa skonsolidowane modele organizacyjne i menedżerskie:

- podsystem (*sub-system*) – koncentrujący się na analizie relacji zachodzących wśród składników wewnętrznych oraz
- system ogólny wyższego szczebla (*supra-system*) – który dotyczy połączeń między przedsiębiorstwami oraz podmiotami występującymi w ich kontekście.

J.W. Gościński [24] i A.K. Koźmiński [25], określając rolę systemu zarządzania w organizacji, stwierdzili, że pełni on nadrzędną funkcję wobec pozostałych podsystemów i odpowiada za ich działania. Samoregulacja systemu zarządzania przejawia

się w koncentracji zarządu przedsiębiorstwa na zagadnieniach prospektywnych, tworzeniu i przeprojektowywaniu podsystemów [26]. Podstawowymi formami zewnętrznego działania systemu zarządzania są m.in.: emisja, gromadzenie i przetwarzanie informacji. W szczególności należy podkreślić realizację całego kompleksu działań kierowniczych, polegających na podejmowaniu decyzji planistycznych i organizacyjnych oraz na motywowaniu i kontrolowaniu ich przebiegu i rezultatu [27].

Proces samoregulacji zmierzający do utrzymania stanu równowagi systemu nazywa się procesem homeostazy [28]. Homeostazę można określić jako utrzymanie jednej lub wielu zmiennych na tym samym poziomie, pomimo zmian zachodzących w otoczeniu. Pozwala ona przedsiębiorstwom pozostać w stanie równowagi dynamicznej [29].

Kluczowe czynniki kształtowania organizacji są równocześnie fundamentalnymi determinantami jej systemu zarządzania. Zalicza się do nich: dynamikę i poziom konkurencyjności rynków i produktów, wielkość przedsiębiorstwa, oczekiwania i władzę kluczowych interesariuszy [30], potencjał członków organizacji, zaawansowanie i rozpowszechnienie kluczowych technologii, uwarunkowania kulturowe organizacji i środowiska [31, 32]. Przedsiębiorcy muszą planować i skutecznie wdrażać dostosowania strukturalne w celu zagwarantowania trwałości systemu zarządzania organizacji, formułując nowe scenariusze biznesowe. Implikuje to pozycjonowanie, transformację i ponowne redefiniowanie struktury organizacyjnej, w celu zapewnienia zrównoważonego rozwoju firmy w perspektywie długoterminowej [33, 34].

3. Najważniejsze właściwości helu i jego zastosowanie

Hel jest drugim w kolejności po wodorze najbardziej rozpowszechnionym pierwiastkiem chemicznym we wszechświecie, choć występuje na Ziemi jedynie w śladowych ilościach [35]. Jest to bezbarwny, bezwonny, bez smaku, nietoksyczny gaz obojętny, który należy do grupy gazów szlachetnych [36].

W 1868 r. Pierre Janssen odkrył hel w widmie słońca podczas obserwacji zaćmienia metodą spektroskopową. W 1895 r. szkocki chemik William Ramsay jako pierwszy odkrył hel na Ziemi podczas eksperymentów z minerałem zawierającym uran. Po zbadaniu próbek metodą spektroskopową przez Jenssena i Lockyera, zidentyfikowano hel identyczny do odkrytego na słońcu [37]. W 1903 r. po raz pierwszy znaleziono hel w złożach gazu ziemnego na terenie kopalni w Kansas w USA [38].

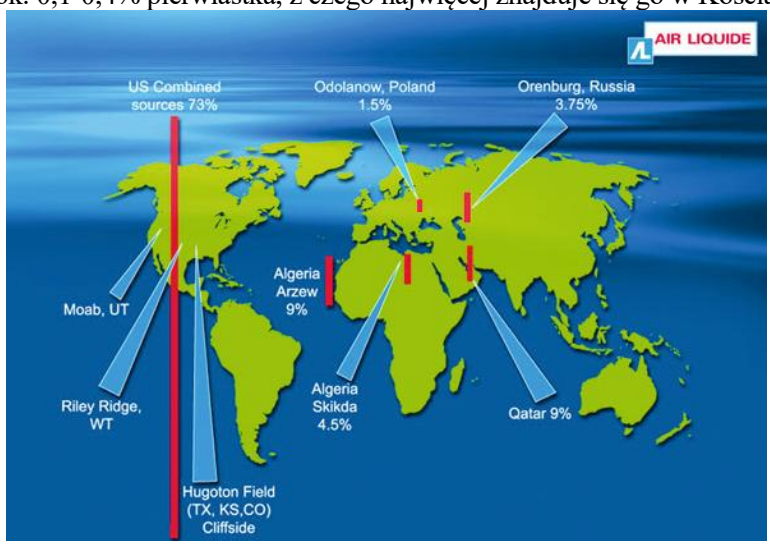
Globalne wydobycie helu odbywa się jedynie w kilku krajach na Ziemi. Wynika to m.in. z ograniczonego dostępu do źródeł, które posiadają wystarczający procent pierwiastka oraz z opłacalności jego wydobycia. Często polega to na zastosowaniu kosztownego procesu niskotemperaturowego skraplania, w wyniku którego następuje wydzielenie z cieczy składającej się z węglowodorów strumienia gazowego helu surowego, a następnie jego oczyszczenie w instalacjach kriogenicznych lub przy użyciu metody PSA (z ang. *pressure swing adsorption*) [39].

Hel pozyskiwany jest głównie w 5 krajach (rys. 1), z czego czołowym liderem (ok. 73% globalnego wydobycia gazu) są Stany Zjednoczone, gdzie roczne wydobycie helu wynosi ok. 75 mln m^3. Największy obszar tych źródeł znajduje się w stanach Kansas, Oklahoma i Teksas, gdzie stężenie pierwiastka wynosi 0,3-2,7% [39].

W czołówce globalnej produkcji helu znajdują się również Algieria z 13,5% udziałem (wydobycie 40 mln m^3/rok) oraz Katar posiadający 9% udział w produkcji (wydobycie

15 mln m^3/rok). Na 4. pozycji jest miasto Orenburg znajdujące się w zachodniej części Rosji, w którym rocznie pozyskuje się do 8,8 mln m^3 helu [41].

W przypadku Europy jedynym krajem pozyskującym hel z gazu ziemnego jest Polska, która posiada 1,5% udziału w światowej produkcji oraz roczne wydobycie na poziomie 3 mln m^3. Proces ten odbywa się w zakładzie Polskiego Górnictwa Naftowego i Gazowego (PGNiG S.A.) zlokalizowanym w Odolanowie [42]. Krajowe zasoby złóż zawierają ok. 0,1-0,4% pierwiastka, z czego najwięcej znajduje się go w Kościanie [43].



Rysunek 1. Miejsca wydobywania helu na Ziemi. Źródło: [40]

Poza wymienionymi obszarami globalnego wydobycia helu z czasem odkrywane są kolejne złoża zawierające hel. Do najgłośniejszych odkryć ostatnich lat należy Rift Valley w Tanzanii, których wielkość złóż oszacowano na ponad 1,5 mld m^3, co jest największym tego typu odkryciem na świecie [44]. Każda pozyskana ilość helu oraz każde nowo odkryte jego źródło ma niebagatelne znaczenie na światowym rynku, zarówno w sferze kreowania cen, jak i szerokich możliwości zastosowań.

Ze względu na niezwykłe właściwości, hel jest powszechnie stosowany w różnorodnych dziedzinach, począwszy od komercyjnych gałęzi przemysłu, a skończywszy na badaniach naukowych. Początki przemysłu helowego sięgają okresu dwudziestolecia międzywojennego, kiedy to hel ze względu na swoją lekkość i bezwładność był przedmiotem zainteresowań sił zbrojnych USA [45]. Obecnie pierwiastek ten wykorzystywany jest głównie w 7 obszarach: kriogenika (30%), gaz podnoszący (17%), półprzewodniki, nadprzewodniki, światłowody (14%), spawanie (9%), inżynieria i nauka (8%), wykrywanie nieszczelności (6%), chromatografia gazowa (6%).

Warto zwrócić uwagę na wykorzystanie helu w międzynarodowych systemach obronnych obejmujących m.in. jednostki obserwacyjne, balony naukowe, testowanie silników rakietowych oraz systemów naprowadzania rakiet powietrze-powietrze. Co więcej, gaz ten wykorzystywany jest w programach kosmicznych przez duże agencje, w tym NASA i Arianespace. Są one czołowym użytkownikiem helu do sprężania paliw wodorowych, produkcji i użytkowania rakiet, maszyn oraz statków kosmicznych [46].

4. Rynek dystrybucji helu ciekłego i gazowego

Rynek handlu helem jest bardzo specyficzny, ponieważ towar ten nie jest dobrem sprzedawanym i kupowanym na publicznych giełdach. Jego ceny podlegają znacznie większym wahaniom niż ceny ropy naftowej. Z tego względu przedsiębiorców dystrybuujących hel powinna cechować wysoka elastyczność zarządzania w sytuacji niespodziewanych zmian.

Kluczową kwestią dotyczącą charakterystyki rynku dystrybucji helu ciekłego i gazowego, jest analiza czynników, które go napędzają. Uwzględniając określone cechy surowców mineralnych, wyróżnia się cztery czynniki, które wpływają na podaż, popyt i cenę helu. Należą do nich: niepewność geologiczna, wielkość zapotrzebowania, producenci i dostawcy helu oraz rynek produkcji gazu ziemnego [47].

Niepewność geologiczna stanowi czynnik, na który człowiek nie ma znaczącego wpływu. Wiąże się to z faktem, iż hel jest nieodnawialnym zasobem o ograniczonych źródłach na Ziemi. Amerykańska Agencja Naukowo-Badawcza i Departament Spraw Wewnętrznych USA oszacowały wspólnie światowe rezerwy helu w 2019 r. na 51,9 mld m^3. Przy założeniu obecnego popytu na poziomie stałym, zapasy helu na Ziemi wystarczą na około 230 lat [48]. Zwiększające się z każdym rokiem zapotrzebowanie na hel może przyczynić się do krytycznego spadku jego podaży nawet w ciągu kilkudziesięciu lat. Czynnik ten sprawia, iż w każdej chwili może dojść do nagłych zmian cen oraz do ograniczonej dostępności helu. Przerywane dostawy są niezwykle szkodliwe w wybranych branżach przemysłu oraz w medycynie, gdzie w wielu aspektach nie ma żadnych zamienników dla helu [49]. Zarządzanie jego zapasami i dystrybucją staje się dużym wyzwaniem dla przedsiębiorstw. Umiejętność dopasowania się firm do nieprzewidywalnych sytuacji na rynku jest kluczem w ich utrzymaniu i rozwoju.

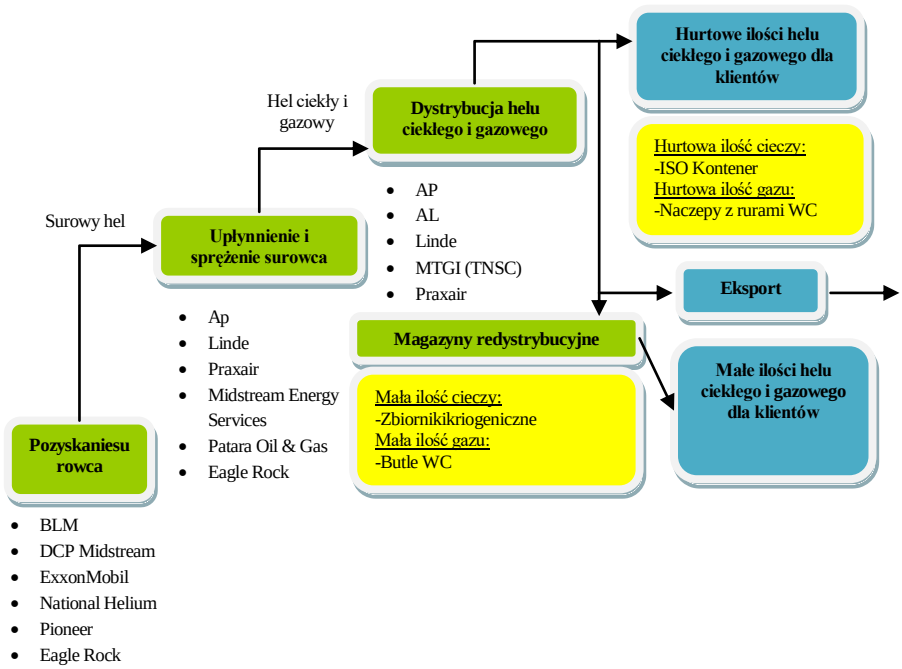
Drugim czynnikiem wpływającym na rynek dystrybucji helu jest wielkość jego zapotrzebowania. Im większy popyt, tym prężniej rozwijający się rynek. Zapotrzebowanie na hel, podobnie jak w przypadku większości innych surowców, jest przede wszystkim popytem pochodnym. Oznacza to, że wielu konsumentów helu nie wykorzystuje go jako produktu finalnego, ale jako wkład do produkcji innych towarów i usług. Przykładem może być hel do schładzania układu rezonansu magnetycznego lub wiązek światłowodów w telekomunikacji.

Wśród międzynarodowych dostawców i producentów helu istnieje pewny potencjał do konsolidacji, której celem jest wewnętrzna spójność branży. Rynek dystrybucji ma wiele etapów, począwszy od wydobycia i przechowywania surowego helu do rafinacji, przez jego upłynnienie i sprężenie, po transport oraz końcowe wykorzystanie. Procesy te wymagają pracy wielu przedsiębiorstw umożliwiających efektywną komunikację między poszczególnymi fazami rynku. Schemat procesu dystrybucji na przykładzie Stanów Zjednoczonych zaprezentowano na rysunku 2.

Łańcuch dostaw rozpoczyna się od wydobycia surowego helu ze złóż gazu ziemnego. Przykładowymi agencjami i przedsiębiorstwami w USA zajmującymi się tym etapem są m.in. BLM, Pioneer, Eagle Rock i inni. Surowy hel przechodzi dalszy etap oczyszczenia, upłynnienia lub sprężenia. Następnie zostaje przetransportowany do dużych klientów, terminali eksportowych oraz wtórnych magazynów redystrybucyjnych w celu przepakowania i dostarczenia helu do średnich i małych klientów.

Czwartym czynnikiem wpływającym na rynek dystrybucji helu ciekłego i gazowego jest rynek produkcji gazu ziemnego, z którego bezpośrednio uzyskuje się surowy hel.

Wynika to z faktu, iż proces wydobycia tego pierwiastka z wybranych źródeł surowca energetycznego jest pierwszym etapem kształtowania jego początkowych cen. Kwestia ta jest ściśle związana z kosztami specjalnie zaprojektowanych i niezwykle drogich instalacji (koszt instalacji membranowych – ok. 50 mln zł), rafinacji uzyskanego gazu i magazynowania zapasów [47].

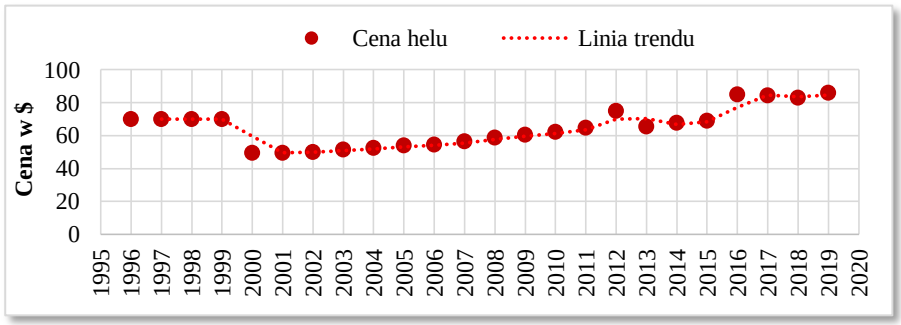


Rysunek 2. Łańcuch dostaw helu ciekłego i gazowego. Źródło: Opracowanie własne na podstawie [50]

5. Ceny helu i ich prognozy w latach 1996-2019

Rynek dystrybucji helu składa się z globalnej konkurencji prywatnych firm, w których dominuje aż sześciu potężnych liderów: Air Products z Pensylwanii, AirLiquide z Paryża, Linde Group i Messer z Niemiec, Praxair z Connecticut oraz Matheson z New Jersey. Dodatkowym uczestnikiem rynku, którego zadaniem jest m.in. sprawowanie kontroli nad globalnym poziomem cen jest rząd Stanów Zjednoczonych. Interwencja federalnej agencji Bureau of Land Management stwarza niepewność dla wszystkich tych firm, które muszą reagować na wszelkie zmiany oraz trendy w wydobyciu gazu ziemnego. Ponadto za jej pośrednictwem tłumione są zawyżone ceny helu po kryzysowych sytuacjach rynku [46].

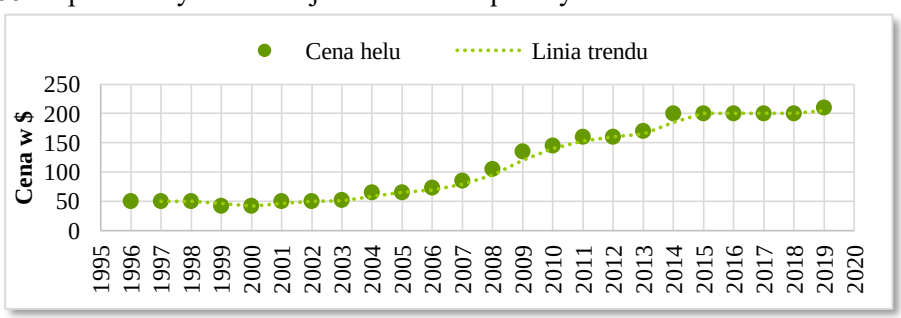
Dane przedstawione na rysunkach 3. i 4. ukazują znaczne różnice w ustanowionych cenach przez prywatnych sprzedawców helu, a rząd USA. Podane ceny w walucie dolara ($) przypisane są na określoną ilość 1000 ft^3, czyli 1000 stóp sześciennych.



Rysunek 3. Ceny rządowe helu za 1000 ft^3. Źródło: Opracowanie własne na podstawie [51]

W przypadku rysunku 3., ceny helu przez pierwsze cztery lata monitorowania utrzymywały się na poziomie 70 $, co mogło wówczas wynikać z posiadania największych światowych rezerw zasobów oraz jednocześnie słabszej konkurencji na rynku. W 2000 r. nastąpiło ich nagłe obniżenie do 50 $, a następnie z każdym mijającym rokiem, powolne zwiększanie. Od 2016 r. ceny ponownie się ustabilizowały, jednak na poziomie już ok. 85 $, co stanowi najwyższą do tej pory cenę ustaloną przez rząd Stanów Zjednoczonych.

Na rysunku 4., który przedstawia średnie ceny helu ustalone przez prywatne firmy produkcyjne i dystrybucyjne, dane od 1996 r. do 2019 r. wskazują wyraźnie widoczną tendencję wzrostową. Pierwszy etap ich podniesienia z 50 na 60 $ za 1000 ft^3 nastąpił w 2004 r. podczas tymczasowej niestabilności podaży.



Rysunek 4. Ceny prywatnych sprzedaży helu za 1000 ft^3. Źródło: Opracowanie własne na podstawie [51]

Obecne ceny helu sięgają ponad 200 $, co stanowi aż 400% początkowej ceny helu z 1996 r. Co więcej, porównując poziom cen ustalonych przez rząd oraz prywatne firmy, można zaobserwować, iż z każdym rokiem różnice ich poziomów znacząco się pogłębiają. Zdecydowanie drożej wychodzi zakup helu za pośrednictwem prywatnych sprzedawców, u których obecne ceny helu przewyższają rządowe o prawie 150%.

6. Działalność badanego przedsiębiorstwa – studium przypadku

Badane przedsiębiorstwo rozpoczęło funkcjonowanie w 2003 r. w miejscowości Odolanów w województwie wielkopolskim. Firma należy do grupy dostawców helu ciekłego i sprężonego na polskim i zagranicznym rynku. W 2004 r. podjęto pierwsze znaczące inwestycje, w postaci zakupu samochodu dostawczego oraz drogich specjali-

stycznych zbiorników do transportu ciekłego helu. Biorąc pod uwagę dynamiczny rozwój przedsiębiorstwa, w tym ciągłe pozyskiwanie nowych pracowników, klientów, dostawców, sprzętów i zbiorników, postanowiono przenieść siedzibę firmy do miejscowości Czarnylas w powiecie ostrowskim. Czarnylas znajduje się w odległości 45 km od wrocławskiej obwodnicy, a także 10 km od krajowej drogi nr 11 łączącej Poznań i Bytom. Udogodnieniem jest również oddalone o 90 km lotnisko we Wrocławiu. Obszar działalności znajduje się na szlaku, którym najczęściej przewożone są rezonanse magnetyczne MRI, które wymagają uzupełnienia poziomu helu w trakcie transportu.

Obecnie przedsiębiorstwo zajmuje się hurtową sprzedażą wyrobów chemicznych i usług z nimi związanych. Firma specjalizuje się w dostawach helu ciekłego i gazowego na terenie całej Europy oraz w różne zakątki świata. Oferuje usługi uzupełniania poziomu helu w sprzętach diagnostyki obrazowej (MRI) znajdujących się na terenie Polski oraz dodatkowo umożliwia przechowywanie rezonansów magnetycznych i ich serwisowanie na terenie siedziby firmy.

Rysunek 5. Siedziba firmy
Źródło: Opracowanie własne

Rysunek 6. Samochód dostawczy wraz ze zbiornikiem do transportu helu ciekłego
Źródło: Archiwum firmy

Rysunek 7. Zestaw zbiorników kriogenicznych do transportu helu ciekłego
Źródło: Archiwum firmy

Rysunek 8. Oferowane rodzaje zbiorników kriogenicznych
Źródło: Archiwum firmy

Rysunek 9. Butle gazowe do przechowywania i transportu helu sprężonego
Źródło: Archiwum firmy

Rysunek 10. Proces uzupełniania poziomu helu w MRI
Źródło: Archiwum firmy

Najczęściej sprzedawanym produktem z oferty jest ciekły hel, dlatego zarządzanie jego zapasami oraz dystrybucją jest jednym z najważniejszych zadań przedsiębiorstwa. Wychodząc naprzeciw oczekiwaniom klientów, firma oferuje 3 rodzaje zbiorników o pojemności: 100, 250 i 500 litrów. Zważywszy na różną szerokość wejść do pomieszczeń, zmodyfikowano szerokość roboczą zbiorników tak, aby mogły dotrzeć w każde miejsce (rys. 7 i 8). Każdy ze zbiorników zaprojektowany jest w charakterze naczynia Dewara, który jest jedynym sposobem na przechowywanie i transportowanie ciekłego helu [52].

Kolejną usługą firmy jest napełnianie systemów rezonansów magnetycznych znajdujących się w placówkach prywatnej służby zdrowia oraz w publicznych szpitalach

(rys. 10). Przedsiębiorstwo, dzięki wysoko wykwalifikowanemu personelowi, osiąga wysoką efektywność napełniania helu zarówno w urządzeniach wieloletnich, jak i w najnowszych modelach znanych producentów sprzętu MRI.

Ostatnią pozycją z listy oferowanych usług jest możliwość magazynowania rezonansów magnetycznych na terenie siedziby firmy, wraz z ich serwisowaniem. Od 2010 r. przedsiębiorstwo magazynowało ponad 70 rezonansów, w tym głównie marki wiodących firm w branży, takich jak Siemens czy Philips. Przedsiębiorstwo posiada doskonałe zaplecze technologiczne, w tym urządzenia i części do wszystkich marek MRI. Firma gwarantuje wysoką jakość wszystkich oferowanych produktów i usług. Z każdym rokiem wypracowane doświadczenie, zdobyta wiedza oraz specjalistyczne zaplecze techniczne pozwala budować silną współpracę z klientami oraz utrzymywać wysoką pozycję na rynku.

7. Misja, wizja i strategia organizacji

Misja, wizja i strategia to trzy filary każdego biznesu, których nie powinno zabraknąć w żadnym przedsiębiorstwie ukierunkowanym na osiągnięcie sukcesu. Badane przedsiębiorstwo dokładnie precyzuje kierunek działania oraz najważniejsze cele krótko- i długoterminowe. Działalność firmy dotycząca wewnętrznych procesów oraz otoczenia opiera się na szczegółowych planach, dostosowanych do aktualnej sytuacji rynkowej. Plany te, przekształcane w pojęcie misji i wizji przedsiębiorstwa, stanowią wyraz aspiracji zarządu oraz zespołu pracowników. Biorąc pod uwagę rozwój przedsiębiorstwa i zaspokojenie oczekiwań klientów, badana firma określiła misję swojej działalności: *Jesteśmy profesjonalnym dostawcą helu ciekłego i gazowego oraz zaufanym usługodawcą przechowywania rezonansów magnetycznych MRI. Jesteśmy dla Ciebie – wszędzie*.

Treść misji przedsiębiorstwa opiera się na trzech głównych przesłaniach, obejmujących profesjonalizm prowadzonej działalności, zaufanie i oddanie dla klienta. Specjaliści do spraw sprzedaży i obsługi służą pomocą i radą w każdej sytuacji, nie pozostają nigdy obojętni na problemy odbiorców, a każdy przypadek i prośbę rozpatrują indywidualnie.

Badane przedsiębiorstwo stworzyło również wizję swojej przyszłości. Została ona jasno zakomunikowana i przyjęta w całej społeczności organizacji. Jest to obraz przyszłości, który obejmuje stabilność przedsiębiorstwa cieszącego się szeroko rozumianym szacunkiem i zaufaniem w branży oraz sukcesywnie zwiększającą się liczbą powracających klientów. Dodatkowo, wyobrażenie to rozwinięte jest o rozwój firmy w postaci zatrudniania nowych pracowników, poszerzania parku technologicznego, środków transportu oraz rozbudowania specjalistycznego laboratorium badawczego. Wizja ta wiąże się z optymistycznym nastawieniem na sukces, którego osiągnięcie jest jednocześnie realne i wiarygodne.

W oparciu o cel istnienia przedsiębiorstwa oraz jego skrupulatnie wykreowaną przyszłość, została opracowana i wdrażana strategia działania, będąca odpowiedzią na pytanie „jak najskuteczniej zrealizować wizję i misję firmy?". W tym celu rozpatrzono szereg czynników determinujących efektywność budowania strategii. Przeprowadzono analizę oraz ocenę otoczenia przedsiębiorstwa, w tym wszystkie elementy, które mogą wpływać na funkcjonowanie firmy. Dotyczyły one zmian zachodzących na rynku w ciągu kilku ostatnich lat oraz rosnących oczekiwań klientów. Kluczową kwestią była

również dokładna analiza potencjału własnej organizacji w wymiarze jakościowym i ilościowym.

W celu dopełnienia planu skonstruowania dokładnej i skutecznej strategii, firma określiła swoje mocne i słabe strony, do których należą:

- Mocne strony – mała rodzinna firma; wysokie zaangażowanie właściciela oraz całej kadry zarządzającej we wszystkie czynności i procesy; doskonale przeszkolona kadra pracownicza; szybki przepływ informacji pomiędzy poszczególnymi szczeblami w hierarchii struktury organizacyjnej; duża elastyczność, mobilność i otwarcie na różne potrzeby klienta; gotowość do działania 7 dni w tygodniu, 24 godziny na dobę; światowy zasięg oferowanych produktów i usług; doskonała lokalizacja zarówno w kraju, jak i w Europie; stała współpraca z licznymi klientami z kraju i Europy; pogodna i przyjazna atmosfera pomiędzy wszystkimi pracownikami; wysoka mobilizacja związana z konkurencją w postaci dużych korporacji; forma prawna przedsiębiorstwa – Sp. z o.o.; dobrze rozwinięte zaplecze technologiczne; własne środki transportu.
- Słabe strony – okresowe niedobory surowca na lokalnym rynku; trudności w znalezieniu na rynku odpowiednio wykwalifikowanej kadry; trudno dostępne na rynku urządzenia (zbiorniki do ciekłego helu); różnice w cenie surowca uzależnione od jego podaży.

Strategia działania przedsiębiorstwa obejmuje rozwój i doskonalenie wszystkich gałęzi działalności. Aby sprostać coraz większym oczekiwaniom klientów, postanowiono stworzyć organizację produkcji, która zapewni nieprzerwane posiadanie helu ciekłego i gazowego oraz elastyczny i nowoczesny system dostaw. Cel ten realizowany jest każdego dnia poprzez: ciągłe poznawanie potrzeb i oczekiwań klientów; dopasowywanie i planowanie dostaw pod kątem indywidualnych potrzeb klienta; prawidłową realizację wszelkich usług w zakresie terminowości, ilości dostarczonego helu, szybkości i jakości serwisu; analizę dostępności surowca na światowym rynku i racjonalne planowanie produkcji, w tym rezerw magazynowych; minimalizowanie strat w procesach technologicznych poprzez korzystanie z nowoczesnych maszyn i urządzeń; doskonalenie kadry pracowniczej za pomocą regularnych szkoleń z zakresu systemu zarządzania, nauki języków obcych, pierwszej pomocy, itp.; doskonalenie systemu zarządzania jakością poprzez prowadzenie audytów wewnętrznych i zewnętrznych oraz aktualizowanie księgi jakości; systematyczną analizę i ocenę efektywności systemu zarządzania.

8. Struktura organizacyjna i zarządzanie personelem

Przez kilka pierwszych lat przedsiębiorstwo funkcjonowało jako jednoosobowa działalność gospodarcza. Właściciel był zobowiązany do wykonywania wszystkich prac związanych z administracją, serwisem oraz dostawą ciekłego helu. Z biegiem czasu narastająca liczba zleceń wywołała potrzebę zatrudnienia dodatkowych pracowników odpowiedzialnych za transport, sprawy techniczne i realizację zamówień. W 2019 r. w organizacji pracowało ponad 20 osób w ramach struktury organizacyjnej obejmującej cztery piony:

- Dział administracyjny – zajmujący się prowadzeniem biura, sprawozdawczością, rozwiązywaniem bieżących spraw, fakturowaniem, księgowością oraz prowadze-

niem księgi kadr i płac. Dział ten odpowiada także za funkcjonowanie systemu zarządzania jakością.

- Dział obsługi klienta i Dział sprzedaży – są odpowiedzialne za opracowywanie i aktualizowanie strategii sprzedażowej, kalkulowanie oferowanych cen, analizowanie sytuacji na rynku, analizowanie działań konkurentów oraz częste negocjacje z klientami i dostawcami ciekłego helu.
- Dział techniczny – zaopatruje przedsiębiorstwo w odpowiednią wielkość zapasów helu w magazynach, przygotowuje produkt dla klientów, napełniając butle i zbiorniki kriogeniczne sprężonym i ciekłym helem oraz sprawuje nadzór nad stanem technicznym i serwis rezonansów magnetycznych powierzonych firmie przez zleceniodawców.

Stosowane w przedsiębiorstwie metody zarządzania personelem opierają się na określonym procesie, którego etapy rozpoczynają się od samego planowania zasobów ludzkich, a kończą na wynagrodzeniu za wykonaną pracę. Zarządzanie kadrami rozpoczyna się w momencie dobierania pracowników, czyli ich rekrutacji i selekcji. Szczegółowy proces zatrudniania nowych osób do firmy obejmuje:

- analizę stanowiska pod kątem zapotrzebowania;
- decyzję o naborze;
- rekrutację;
- selekcję dokumentów zgłoszeniowych;
- rozmowę rekrutacyjną, po której następuje rezygnacja kandydata lub
- przygotowanie oferty pracy, a następnie
- zatrudnienie i adaptacja pracownika.

W przypadku doskonalenia pracowników, który jest kolejnym z etapów procesu zarządzania personelem, wprowadzono w przedsiębiorstwie systematyczne szkolenia. Dzielą się one na wewnętrzne i zewnętrzne. Szkolenia wewnętrzne mają na celu stworzenie lepszej organizacji pracy, w tym jej usprawnienie tak, aby pracownik poczuł się bardziej komfortowo i pewnie. Wśród dużej ilości szkoleń pracowników należy wyróżnić: wymagania ISO 9001:2016 – system zarządzania jakością; efektywne zarządzanie zespołem pracowniczym; zarządzanie czasem własnym i zespołu; zarządzanie zmianami; komunikacja w zespole; charyzmatyczny lider; ochrona przeciwpożarowa i pierwsza pomoc.

Głównym celem szkoleń zewnętrznych jest podnoszenie i rozwijanie kwalifikacji zawodowych. Pracownicy techniczni uczestniczą m.in. w kursach dotyczących uprawnień energetycznych, spawalniczych, obsługi wózków widłowych oraz suwnic. Prowadzone są również szkolenia dla inżynierów serwisu oraz dostawców ciekłego helu, w tym na temat transportu ADR oraz systemu różnych rezonansów magnetycznych, ich budowy i procesu napełniania.

9. Najważniejsi konkurenci badanego przedsiębiorstwa

Badane przedsiębiorstwo funkcjonuje na rynku w branży, która jest ściśle ukierunkowana na określone dziedziny nauki i sfery życia. Klienci, którzy są zainteresowani działalnością organizacji, stanowią specyficzne grupy konsumentów zorientowanych głównie na konkretne zastosowanie oferowanej usługi lub produktu. Najczęściej sprzedawany produkt w firmie – ciekły hel, nie stanowi towaru, który można znaleźć

w zwyczajnych sklepach. Jego oferta wymaga wyspecjalizowanych przedsiębiorstw, które w swojej działalności kierują się określonymi rodzajami zapotrzebowania.

W Polsce, jak i za granicą, można znaleźć niewiele firm oferujących hel ciekły i gazowy, co wynika z unikalności tej branży i w porównaniu do innych dziedzin działalności gospodarczej jest mało spotykane. Biorąc pod uwagę zasięg oferowanych usług, do najważniejszych konkurentów badanego przedsiębiorstwa, należy zaliczyć następujące spółki:

- Polskie Górnictwo Naftowe i Gazownictwo S.A. – to największa spółka w kraju, która zajmuje się wydobyciem ropy naftowej i gazu ziemnego. Zaistniała na rynku już w 1982 r. Prowadzi działalność na szeroką skalę w Polsce (7 oddziałów), jak i za granicą. Filia PGNiG w Odolanowie jest jedynym miejscem w Europie, w którym oddziela się hel z gazu ziemnego posiadającego odpowiednie właściwości. Z tego względu PGNiG zajmuje bardzo silną pozycję na europejskim rynku. Posiadając bezpośredni dostęp do jedynego na kontynencie źródła zasobów, firma ta ma możliwość samodzielnego kreowania cen [53].
- Linde Gaz Polska – to polski oddział powołany do życia w 1993 r. przez niemiecki koncern Linde Group. Główna siedziba oddziału znajduje się w Krakowie wraz z siedmioma zakładami w innych miejscowościach Polski. W swojej ofercie firma posiada m.in. sprężone gazy medyczne i techniczne, mieszanki tych gazów, suchy lód oraz czynniki chłodnicze. Działa na szeroką skalę na terenie Europy [54].
- Air Products – to producent i międzynarodowy dostawca gazów technicznych, którego siedziba znajduje się w stanie Pensylwania w USA od 1940 r. Prowadzi działalność w 50 krajach, w tym w stolicy Polski (Warszawa). W swojej ofercie oprócz rozbudowanego sektora usług, przeglądów oraz sprzedawanych urządzeń, firma ta rozprowadza szereg gazów – od tlenu, argonu i helu – po gazy spawalnicze [55].
- AirLiquide – to francuska firma chemiczna założona w 1902 r. Główna siedziba przedsiębiorstwa znajduje się w Paryżu, a jej poszczególne oddziały mieszczą się w wielu krajach na świecie, w tym także w Polsce. W swojej ofercie posiada m.in. gazy techniczne, spożywcze oraz specjalne [40].

Analizując rynek największych konkurentów badanego przedsiębiorstwa, można zauważyć, iż są to głównie duże międzynarodowe koncerny, które działają w branży od dziesiątków lat. Badana firma, chociaż oferuje swoje usługi na terenie wszystkich kontynentów, jest stosunkowo mała. Tak silna konkurencja motywuje pracowników do ciągłego rozwoju i doskonalenia oferowanych usług.

Rysunek 11. Logo firmy PGNiG
Źródło: [53]

Rysunek 12. Logo koncernu Linde Group
Źródło: [54]

Rysunek 13. Logo firmy Air Products
Źródło: [55]

Rysunek 14. Logo firmy AirLiquide
Źródło: [40]

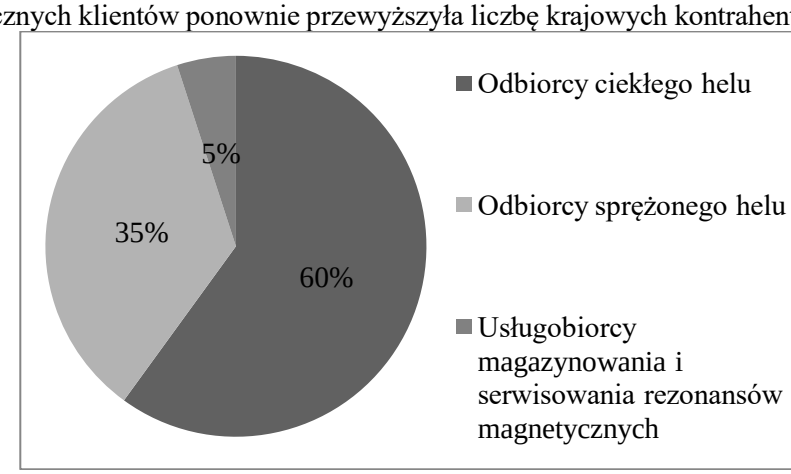
Przedsiębiorstwo często wybierane jest przez klientów, którzy nie potrzebują w jednym czasie hurtowych ilości produktów proponowanych przez duże koncerny. Niejednokrotnie firma wygrywa również dzięki pozyskanej renomie, lojalności i zaufaniu stałych klientów. Za pomocą elastycznej organizacji, indywidualnego podejścia, kompleksowości i wysokiej jakości świadczonych usług, badane przedsiębiorstwo obsługuje klientów, którzy cenią działalność małych i średnich firm.

10. Kontrahenci współpracujący z badanym przedsiębiorstwem

Wszystkich klientów zainteresowanych ofertą usług i produktów przedsiębiorstwa można podzielić na trzy grupy, do których należą:

- odbiorcy kupujący hel w postaci ciekłej – stanowią najliczniejszą grupę (generującą ok. 60% całościowej sprzedaży produktów i usług);
- odbiorcy sprężonego helu gazowego, którego udział stanowi ok. 35% w całościowej sprzedaży produktów i usług;
- klienci zainteresowani usługą magazynowania i serwisowania rezonansów magnetycznych na terenie siedziby firmy (5% udział w sprzedaży towarów i usług).

W latach 2003-2018 liczba polskich, jak i zagranicznych klientów badanego przedsiębiorstwa nieustannie rosła. Od 2003 r. do 2007 r. można zaobserwować przewagę polskich kontrahentów. W latach 2008-2012 to zagraniczni klienci zaczęli stanowić większą grupę zainteresowanych kupnem ciekłego helu. Następnie przez kolejne 2 lata ich liczba wobec polskich odbiorców delikatnie spadła, co wynikało z nawiązania większych kontaktów z polskimi firmami oraz szpitalami. Od 2015 do 2018 r. liczba zagranicznych klientów ponownie przewyższyła liczbę krajowych kontrahentów.



Rysunek 15. Procentowy rozkład sprzedaży towarów i usług firmy podzielony ze względu na udział trzech grup klientów zainteresowanych ofertą przedsiębiorstwa. Źródło: Opracowanie własne

Grupa odbiorców ciekłego helu związana jest najczęściej z dziedzinami medycyny i dyscyplinami obejmującymi nauki ścisłe. Oferta kompleksowej usługi dostarczania ciekłego helu oraz jego napełniania kierowana jest najczęściej do placówek prywatnej służby zdrowia i publicznych szpitali. Posiadają one rezonanse magnetyczne, w których systematycznie należy uzupełniać poziom helu. Do głównych stałych polskich klientów kupujących hel ciekły należą zatem instytucje medyczne i naukowo-badawcze, które:

- prowadzą badania za pomocą rezonansów magnetycznych i tomografii komputerowej;
- wykorzystują hel w laboratoriach termodynamicznych, elektrycznych i do pomiarów magnetycznych;
- zajmują się analizą chromatograficzną i badaniami kriogenicznymi w niskich temperaturach.

Z kolei do najważniejszych zagranicznych odbiorców helu ciekłego należą: dystrybutorzy sprzętu medycznego w zakresie obrazowania magnetycznego dedykowanego dla szpitali i prywatnych klinik medycznych zlokalizowanych na terenie Europy Środkowej i Wschodniej.

W przypadku handlu sprężonym gazem do głównych klientów badanej firmy należą przedsiębiorstwa prowadzące działalność w branży balonowej w kilku krajach Europy Zachodniej.

Grupę kontrahentów przechowujących rezonanse magnetyczne w przedsiębiorstwie w zdecydowanej większości stanowią klienci zagraniczni (z Niemiec, Danii, Austrii i Francji).

Oprócz europejskiego zasięgu oferowanych usług, organizowane są także dostawy produktów do tak dalekich krajów, jak Stany Zjednoczone, Izrael, Armenia, Kenia, Etiopia, Nairobi, czy Wybrzeże Kości Słoniowej. Wskazuje to na osiągnięcie międzynarodowej renomy i biznesowego zaufania w branży.

11. Procesy zachodzące w badanym przedsiębiorstwie

Jednym z fundamentów efektywnego systemu zarządzania przedsiębiorstwem jest właściwe zarządzanie wszystkimi procesami zachodzącymi w ramach prowadzonej działalności. Dzięki logicznemu i nowoczesnemu podejściu do zarządzania procedurami, znacznie łatwiej jest realizować konkretne zadania, monitorować ich przebieg oraz dostosowywać strategie działania do zachodzących zmian.

Do głównych i najważniejszych procesów, których pełną charakterystykę można znaleźć w dostępnej dla pracowników księdze jakości należą:

- proces sprzedaży i obsługi klienta;
- proces zakupu surowca;
- proces napełniania zbiorników do ciekłego helu;
- proces dostawy ciekłego helu LHe;
- proces magazynowania rezonansów magnetycznych MRI;
- proces produkcji helu sprężonego GHe;
- proces zarządzania zasobami;
- proces analizy danych i doskonalenia.

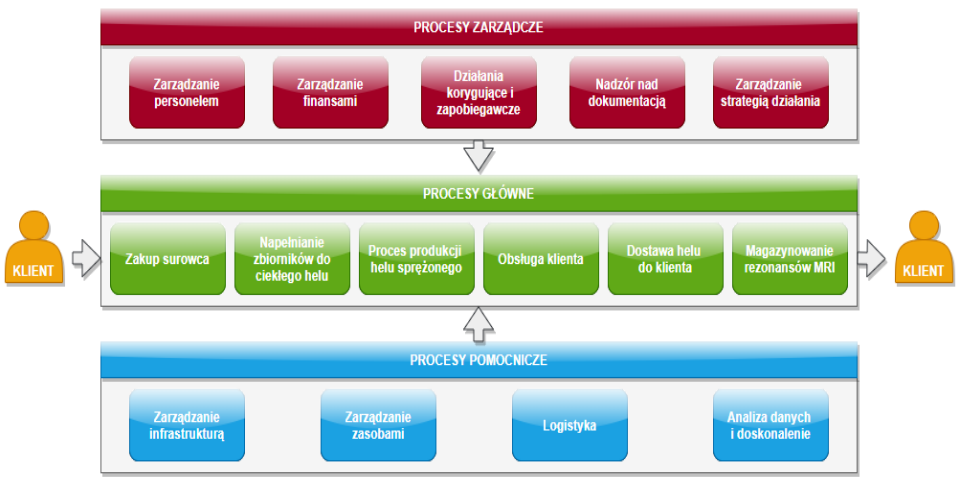
W ramach zachodzących w przedsiębiorstwie procesów można dokonać ich podziału na: procesy główne, zarządcze i pomocnicze. Odpowiednie ich skategoryzowanie jest niezbędne dla realizacji wytyczonych celów oraz organizacji poszczególnych działań.

Identyfikacja procesów umożliwia przede wszystkim standaryzację ich procedur, wdrażanie osobnych systemów ich monitorowania i zarządzania, a także wykazanie pewnych obszarów, w których występują rozbieżności wymagające usprawnień.

W przypadku procesów głównych, ich zakres obejmuje wszystkie działania związane bezpośrednio z charakterem i rodzajem działalności firmy. Są to procesy, które w głównej mierze stanowią podstawowe czynności przyczyniające się do osiągania wyników finansowych. Wynikają one wprost z celu istnienia firmy oraz tworzą tzw. wartość dodaną.

Procesy pomocnicze to wszystkie działania, których przeznaczeniem jest wspieranie w realizacji oraz funkcjonowaniu wszystkich głównych procesów. Ich rola opiera się na rozwoju i osiąganiu wysokiego poziomu jakości realizowanych przez przedsiębiorstwo działań.

Procesy zakwalifikowane do grupy zarządczej dotyczą w głównej mierze kierowania procesami głównymi i pomocniczymi. Oddziałują jednokierunkowo, a w ich zakres wchodzi kontrola i pomiar optymalizacji poszczególnych działań, ustalenie zgodności z określonymi wymogami i celami, a także szczegółowa weryfikacja statusu realizacji procedur korygujących lub zapobiegawczych w sytuacji zachodzących niezgodności. Procesy te związane są z celami przedsiębiorstwa, jego misją i strategią oraz pozycją na rynku.



Rysunek 16. Mapa najważniejszych procesów badanego przedsiębiorstwa. Źródło: Opracowanie własne

Graficznym sposobem przedstawienia poszczególnych procesów zidentyfikowanych przez kadrę zarządzającą, jest mapa procesowa (rys. 16). Informuje ona o dokładnym podziale najważniejszych procesów zachodzących w firmie wraz z ich wzajemnymi powiązaniami.

Biorąc pod uwagę analizę dostępnych materiałów oraz obserwację sposobu zarządzania zachodzącymi procesami w przedsiębiorstwie, stwierdzono, iż badana firma prowadzi względem nich bardzo dobrą politykę. Kadra zarządzająca dba o każdy szczegół związany z działalnością firmy, dlatego też skrupulatnie zajęła się organizacją wszelkich czynności prowadzonych na różnych stanowiskach pracy. W tym celu opracowano dokładnie zasady postępowania, instrukcje oraz wszystkie istotne informacje

przydatne do realizacji zachodzących w firmie procesów. Podzielono je na odpowiednie grupy, co umożliwiło standaryzację procedur, wdrożenie osobnych systemów ich monitorowania, czy też poszerzenie gałęzi określonego doskonalenia do poszczególnych operacji. Przedsiębiorstwo posiada dobrze zorganizowane zarządzanie procesami, którego odbiciem jest efektywne funkcjonowanie firmy na każdym szczeblu organizacyjnym.

12. Wyniki sprzedaży badanego przedsiębiorstwa

Dopełnieniem analizy systemu zarządzania przedsiębiorstwem jest badanie wyników sprzedaży oferowanych przez firmę produktów i usług. Pozyskane informacje dotyczące wielkości sprzedaży i dochodów z nimi związanych, pochodziły z powszechnie dostępnych danych na temat zamówień publicznych.

W przypadku zamówień publicznych, w których przedsiębiorstwo uczestniczyło w latach 2010-2017, firma zwyciężyła w 7 przetargach. Wygrane zamówienia (2 przetargi w 2010 r., 2 w 2014 r., 1 w 2016 r. i 2 w 2017 r.) najczęściej realizowane były w województwie podkarpackim. Łączna liczba organizacji udzielających przetargów wynosiła 5, natomiast instytucją, która najczęściej zamawiała produkty lub usługi przedsiębiorstwa, była publiczna uczelnia wyższa [56].

Szacunkowa wartość wszystkich udzielonych zamówień publicznych badanej firmy wskazuje, że ogłoszeniodawcą zleceń o najwyższej wartości są jednostki naukowe prowadzące badania z użyciem ciekłego helu w zakresie fizyki doświadczalnej.

Najczęściej wygrywane przetargi badanego przedsiębiorstwa dotyczyły dostaw ciekłego i sprężonego helu (43,4% przychodów). Jednakże to usługi związane głównie z napełnianiem aparatury badawczej odpowiednim poziomem helu, przyczyniły się do największego przychodu pochodzącego z publicznych zamówień (56,6% przychodów).

Zamówienia publiczne realizowane przez przedsiębiorstwo w latach 2010-2017 to zaledwie mały ułamek pozyskiwanych przychodów z oferowanych produktów i usług. Głównym sektorem rynku przynoszącym największe dochody jest rynek zamówień prywatnych pochodzących od firm oraz państwowych instytucji i organizacji, które w celu rozpoczęcia współpracy, zwracają się bezpośrednio do przedsiębiorstwa. W związku z ochroną danych, w niniejszej analizie sprzedażowej nie przedstawiono ich wyników. Jednakże, aby rzetelnie móc dokonać jej końcowej oceny, wzięto pod uwagę pozyskane w trakcie rozmów informacje o danych sprzedażowych z prywatnych zamówień.

Przeprowadzona analiza sprzedażowa dotycząca oferowanych przez przedsiębiorstwo produktów i usług w ramach przetargów, umożliwiła ogólne zapoznanie się z osiąganymi przychodami w badanych latach. Uzyskane wyniki sprzedażowe z samych przetargów świadczą o dużym zainteresowaniu potencjalnych odbiorców firmy. Biorąc również pod uwagę fakt, iż jest to mały procent pozostałych dochodów, które pochodzą z prywatnych zamówień można stwierdzić, iż pomimo posiadania dużej i silnej konkurencji w branży gazów technicznych, badane przedsiębiorstwo prężnie działa na rynku.

13. Podsumowanie

Efektywne zarządzanie firmą to niezwykle złożony proces. W przypadku rynku dystrybucji helu ciekłego i gazowego kwestia ta wynika przede wszystkim z tego, że oprócz konieczności posiadania umiejętności, doświadczenia i rozległej wiedzy

z zakresu szeroko rozumianego zarządzania przedsiębiorstwem, wymogiem jest również doskonała wiedza związana z helem oraz wysoka specyfika rynku dystrybucji helu.

Badane przedsiębiorstwo, funkcjonujące na polskim i zagranicznym rynku dystrybucji helu ciekłego i gazowego, stanowi przykład firmy silnie nastawionej na rozwój i potrzeby zarówno klientów, jak i pracowników. Wyniki badań skupiających się na analizie rynkowej, strategicznej, personalnej, finansowej i procesowej, wskazały wiele mocnych stron przedsiębiorstwa, które determinują jego dynamiczny rozwój. Każdy z pięciu podsystemów zarządzania, osiągnął korzystny wynik, czego rezultatem jest pozytywna ocena końcowa systemu zarządzania badanego przedsiębiorstwa. Zgodnie z weryfikowaną hipotezą, na efektywność systemu zarządzania w badanym przedsiębiorstwie, wpływają takie czynniki, jak:

- wysoka konkurencyjność firmy, która wynika z pozyskania bardzo dobrej renomy oraz biznesowego zaufania wśród uczestników rynku;
- wysoki poziom zadowolenia klientów, czego dowodem jest stale rosnąca grupa zarówno nowych, jak i powracających klientów z różnych zakątków świata;
- wysoka efektywność pracy wynikająca z odpowiednio opracowanej strategii działania obejmującej każdy najmniejszy aspekt funkcjonowania przedsiębiorstwa, w tym również silnie motywującą politykę personalną;
- nieustanne doskonalenie i rozwijanie każdego działania i procesu realizowanego w firmie, w tym umiejętności i wiedzy pracowników;
- stała adaptacja strategii działania do zachodzących w otoczeniu zmian.

Literatura

1. Stańczyk S., *Triangulacja – łączenie metod badawczych i urzetelnianie badań,* [w:] Czakon W. (red.), *Podstawy metodologii badań w naukach o zarządzaniu*, Oficyna a Wolters Kluwer business, Warszawa 2011, s. 80.
2. Czarniawska B., *Interviews and organizational narratives,* [w:] Gubrium J.F., Holstein J.A. (red.), *Handbook of interview research. Context and method*, Sage Publications, Thousand Oaks 2002, s. 735.
3. Stachak S., *Podstawy metodologii nauk ekonomicznych*, Wydawnictwo Książka i Wiedza, Warszawa 2006, s. 179.
4. Kostera M., *Antropologia organizacji. Metodologia badań terenowych*, PWN, Warszawa 2003, s. 101.
5. *Centralna Ewidencja i Informacja o Działalności Gospodarczej,* 2021, https://prod.ceidg.gov.pl/ceidg.cms.engine/ (dostęp: 15 maja 2021).
6. Mele C., Pels J., Polese F., *A brief review of systems theories and their managerial applications*, Service Science, nr 2 (1-2), 2010, s. 126-135.
7. Von Bertalanfyy L., *General system theory: Foundations, development, applications*, George Braziller Inc., New York 1968.
8. Wiener N., *Cybernetics: Or control and communications in the animal and the machine,* Hermann & C Editeurs, The Technology Press Cambridge Mass., John Wiley & Sons Inc., Paris 1948.
9. Ng I.C.L., Maull R., Yip N., *Outcome-based Contracts as a driver for Systems thinking and Service-Dominant Logic in Service Science: Evidence from the Defence industry*, European Management Journal, vol. 27, 2009, s. 377-387.
10. Gadomska K., *Rozwój nauki o organizacji i zarządzaniu*, [w:] Dobrodziej B. (red.), *Podstawy organizacji i zarządzania*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin 2008, s. 53.

11. Kast F.E., Rosenzweig J.E., *General systems theory: Applications for organizations and management,* Academy of Management Journal, vol.15, nr 4, 1972, s. 447-465.
12. Jackson M., *Systems Thinking: Creative Holism for Managers*, John Wiley & Sons, Ltd., Chichester 2003.
13. Weinberg G.M., *An Introduction to General Systems Thinking*, Dorset House Publishing Company, New York 2001.
14. Katz D., Kahn R.L., *The Social Psychology of Organizations*, Wiley, New York 1978.
15. Burns T., Stalker G.M., *The management of innovation*, Tavistock, London 1961.
16. Emery F.E., Trist E.L., *Socio-Technical Systems. Management sciences, models and technique*, Pergamon C.W and others Churchman, London 1960.
17. Clark A., *Associative engines,* MIT Press, Boston 1993.
18. Beer S., *Brain of the Firm*, The Penguin Press, London 1972.
19. Griffin R., *Podstawy zarządzania organizacjami,* PWN, Warszawa 2004, s. 55.
20. Leavitt H.J., *Applied organisational change in industry: Structural, technological and humanistic approaches*,[w:] March J.G. (red.), *Handbook of organization*, Rand McNally and Company, Chicago-Illinois 1965.
21. Sokołowska S., *Organizacja i zarządzanie: Ujęcie teoretyczne*,Wydawnictwo Uniwersytetu Opolskiego, Opole 2009, s. 75.
22. Golinelli G.M., *L'approcciosistemico al governodell'impresa. L'impresasistemavitale*, CEDAM, Padova 2005.
23. Barile S., *L'impresa come sistema – Contributisull'ApproccioSistemico Vitale*, Giappichelli, Torino 2008.
24. Gościński J.W., *Elementy cybernetyki w zarządzaniu*, Państwowe Wydawnictwo Naukowe, Warszawa 1968.
25. Koźmiński A.K., *Analiza systemowa organizacji*,Państwowe Wydawnictwo Ekonomiczne, Warszawa 1979.
26. Beer S., *Preface*, [w:] Maturana H.R., Varela F.J. (red.), *Autopoietic systems*, University of Illinois, Illinois 1975.
27. Dźwigoł H., *Zarządzanie przedsiębiorstwem w warunkach XXI wieku,*Wydawnictwo Politechniki Śląskiej, Gliwice 2013, s. 18.
28. Hannan M.T., Freeman J., *The population ecology of organizations*, American Journal of Sociology, 82(5), 1977, s. 929-964.
29. Jajuga T., Jajuga K., *Elementy teorii systemów i analizy systemowej*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław 1993, s. 58.
30. Christopher W.F., *Holistic Management: Managing What Matters for Company Success*, Wiley-Interscience, Hoboken 2007.
31. Baker M.J. (red.), *The Marketing Book*, Butterworth-Heinemann, Oxford 1994, s. 139-192.
32. Bełz G., Skalik J., *Kształtowanie i doskonalenie systemu zarządzania w przedsiębiorstwach*,Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2011, s. 25.
33. Vicari S., *Risorseaziendali e funzionamentod'impresa,* FinanzaMarketing e Produzione, 3, 1992, s. 127-151.
34. Mele C., Pels J., Polese F., *A brief review of systems theories and their managerial applications*, Service Science, 2 (1-2), 2010, s. 126-135.
35. Wheeler M., *Helium: The disappearing Element*, Springer, New York 2015, s. 2-8.
36. Niechciał J., *Niezwykłe właściwości helu w kriogenicznych temperaturach,* Wszechświat, vol. 114, nr 8-9, 2013, s. 269.
37. Heather H., *Helium*, The Rosen Publishing Group, New York 2007, s. 6-9.
38. McKinney D.B., *Exploring the Elements: Helium*, Enslow Publishing LLC, New York 2019, s. 25-31.

39. Szwast M., Zalewski M., Nikpour R., Sobczak A., *Pozyskiwanie helu z gazu ziemnego za pomocą technik membranowych*, Inżynieria i Aparatura Chemiczna, vol. 53, nr 4, 2014, s. 304-305.
40. Air Liquide, www.airliquide.comdostęp [20.06.2021].
41. U.S. Department of the Interior Bureau of Land Management, www.blm.govdostęp [20.06.2021].
42. Perez A.A., *The Mineral Industry of Poland*, [w:] U.S. Geological Survey. Area Reports – International – Europe and Central Eurasia, vol.III, United States Government Printing Office, Washington 2017, s. 351.
43. Parkiet, www.parkiet.com dostęp [20.06.2021]
44. Polskie Centrum Studiów Afrykanistycznych, www.pcsa.org.pl dostęp [20.06.2021].
45. Nuttall W., Clarke R., Głowacki B., *Stop squandering helium*, Nature, vol. 485, 2012, s. 573-575.
46. Nuttall W., Clarke R., Głowacki B., *The Future of Helium as a Natural Resource*, Routledge, London 2012, s. 15-47.
47. National Research Council, *The Impact of Selling the Federal Helium Reserve*, National Academies Press, Washington 2000, s. 49-50.
48. Danabalan D., *Helium: Exploration Methodology for a Strategic Resource*, Durham Universtity, Durham 2017, s. 1-17.
49. Stokes M., *Our research in on ice due to shortage of helium*, The Independent, 04.01.2013. https://www.independent.co.uk/news/science/our-research-ice-due-shortage-helium-8439110.html.
50. Campbell J.R., *Determination of Fair Market Value Pricing of Crude Helium*, U.S. Department of Interior's Bureau of Land Management and Office of Minerals Evaluation, Lexington-Massachusetts 2013, s. 1-6.
51. U.S. Geological Survey, www.usgs.govdostęp [20.06.2021].
52. Van Sciver S.W. *Helium Cryogenics*, Springer-Verlag, New York 2012, s. 5-6.
53. Polskie Górnictwo Naftowe i Gazownictwo, www.pgnig.pl, dostęp [20.06.2021].
54. Linde Gaz Polska, www.linde-gaz.pl, dostęp [20.06.2021].
55. Air Products, www.airproducts.com.pl, dostęp [20.06.2021].
56. Serwis eGospodarka.pl, www.przetargi.egospodarka.pl, dostęp [20.06.2021].

System zarządzania przedsiębiorstwem funkcjonującym na rynku dystrybucji helu ciekłego i gazowego

Streszczenie

Rynek dystrybucji helu ciekłego i gazowego jest bardzo specyficzny i wyjątkowy ze względu na unikalność surowca, jego właściwości, zastosowania oraz rzadkość występowania, a jego temat stanowi przedmiot interdyscyplinarnych badań naukowych. Celem pracy jest ocena systemu zarządzania przedsiębiorstwa, działającego na polskim i międzynarodowym rynku dystrybucji helu ciekłego i gazowego. Przeprowadzone badania koncentrowały się na szczegółowej analizie rynkowej, strategicznej, personalnej, finansowej i procesowej. Zastosowano metodę triangulacji, w tym: 1) wywiad indywidualny z właścicielem i pracownikami poszczególnych działów firmy, 2) jawną obserwację bezpośrednią procesów zachodzących w organizacji, 3) analizę krytyczno-poznawczą materiałów pierwotnych (zaczerpniętych z firmy) oraz materiałów wtórnych (z przeglądu literatury i elektronicznych baz danych). Zbiór formalnych danych na temat działalności badanego przedsiębiorstwa uzyskano z Centralnej Ewidencji i Informacji o Działalności Gospodarczej. Informacje dotyczące procesów administracyjnych, strategicznych i procesów technicznych związanych z helem otrzymano z księgi jakości, instrukcji, kart charakterystyki helu ciekłego i gazowego oraz z oficjalnej strony internetowej firmy. Zakres finansowej analizy działalności przedsiębiorstwa obejmował dane ze sprawozdań z krajowych przetargów publicznych przeprowadzanych w latach 2010-2017. Wyniki badań wskazały wiele mocnych stron przedsiębiorstwa, które determinują jego dynamiczny rozwój. Każdy z pięciu podsystemów zarządzania (rynkowy, strategiczny, personalny, finansowy i procesowy), osiągnął korzystny wynik, czego rezultatem jest pozytywna ocena końcowa całego systemu zarządzania przedsię-

biorstwem. Zgodnie z weryfikowaną hipotezą, na efektywność systemu zarządzania w badanej organizacji, wpływają takie czynniki, jak: 1) wysoka konkurencyjność firmy, która wynika z pozyskania bardzo dobrej renomy oraz biznesowego zaufania wśród uczestników rynku (dostawców, odbiorców, kooperantów); 2) wysoki poziom zadowolenia klientów, czego dowodem jest stale rosnąca grupa zarówno nowych, jak i powracających klientów z różnych zakątków świata; 3) wysoka efektywność pracy wynikająca z odpowiednio opracowanej strategii działania obejmującej każdy najmniejszy aspekt funkcjonowania przedsiębiorstwa, w tym również silnie motywującą politykę personalną; 4) nieustanne doskonalenie i rozwijanie każdego działania i procesu realizowanego w firmie, w tym również umiejętności i wiedzy pracowników; 5) stałe dostosowywanie strategii działania do zachodzących w otoczeniu zmian.
Słowa kluczowe: system zarządzania, ciekły i gazowy (sprężony) hel, marketing i dystrybucja helu

A management system of a company operating on the liquid and gaseous helium distribution market

Abstract

Liquid and gaseous helium distribution market is very specific and unique due to the extraordinariness of the raw material, its properties, applications and rarity, and its topic encompasses a great deal of interdisciplinary research. This paper focuses on the assessment of the enterprise management system operating on the Polish and international market of liquid and gaseous helium distribution. The conducted research focused on a detailed market, strategic, personnel, financial and process analysis. A triangulation method was used including: 1) an individual interview with the owner and employees; 2) a direct observation of the processes taking place in the organization and 3) a critical-cognitive analysis of available primary materials (derived from the company) and secondary materials (literature review and information from electronic databases). The data on the activity of the enterprise under investigation obtained from the Polish Central Register and Information on Economic Activity. Information on the administrative, strategic and technical processes related to helium obtained from a quality manual, instructions and safety data sheets for liquid and gaseous helium. The financial analysis of the company's operations included data from reports on the national public tenders. The results of the study showed many strengths of the company. All subsystems (market-driven, strategic, personnel, financial and process) achieved a favorable effect, which resulted in a positive final evaluation of the entire management system of the enterprise under investigation. The effectiveness of the company management system operating on helium distribution market is influenced by the following factors: 1) high competitiveness of the enterprise associated with a very good reputation among market participants (suppliers, recipients, cooperators); 2) high level of customer satisfaction manifested in constantly growing group of both new and returning customers from all over the world; 3) high work efficiency resulting from a properly developed action strategy covering every smallest aspect of the company's operation, including a highly motivating personnel policy; 4) continuous improvement and development of each activity and processes carried out in the company, including the skills and knowledge of all employees; 5) constant adaptation of the strategy to the changes in the environment.
Keywords: management system, liquid and gaseous (compressed) helium, marketing and distribution of helium

Andrzej Kępys[1], Ernest Syska[2]

VEO Weather Routing – algorytm wyznaczania tras w oparciu o dane pogodowe, mapowe i żeglowne jachtów

1. Wstęp

Wyznaczanie trasy jachtu żaglowego w oparciu o dane pogodowe jest od dawna zagadnieniem istotnym w praktyce żeglowania. Dotychczas zadanie to realizowano przede wszystkim w oparciu o planowanie „w głowie" kapitana jachtu, a więc o dane pogodowe i wiedzę/intuicję żeglarza odnośnie sposobu wykorzystania wiatru i prądów morskich przez konkretny model statku. Wersja zmatematyzowana, algorytmiczna wyznaczania tras morskich była znana bardziej w świecie regatowym, gdzie każdy błąd to gorszy wynik w wyścigu. W regatach i w zastosowaniach turystycznych dominowały dotychczas algorytmy izochronowe. Algorytmy te są wykonywalne na komputerach o niskiej mocy, ale posiadają szereg ograniczeń funkcjonalnych. Mianowicie nie omijają złożonej linii brzegowej – a jeśli posiadają obejścia tego problemu to tracą na jakości wyznaczonych tras lub całkowicie przestają działać.

W najbliższej przyszłości nastąpi popularyzacja wykorzystania algorytmów wyznaczania tras morskich w standardowych działaniach nawigacyjnych kapitana jachtu. Podobnie jak to miało miejsce w Google Maps i nawigacji po lądzie, podstawą matematyczną tego procesu będą algorytmy Dijkstry. Algorytm Dijkstry musi jednak zostać zaadaptowany do zagadnień nawigacji morskiej. Algorytmem takim jest VEO Weather Routing prezentowany w niniejszym artykule. Logika procedury Dijkstry-ego w systemie VEO pozwala znacznie lepiej omijać lądy oraz strefy wyłączone z ruchu wodnego zyskując przy tym bardzo wysoką dokładność w zakresie rozmiaru kwadratu geograficznego, do którego zmierza pojazd. Odbywa się to kosztem mocy obliczeniowej komputerów, w tym przypadku komputerów wysokiej mocy graficznego przetwarzania danych (GPU).

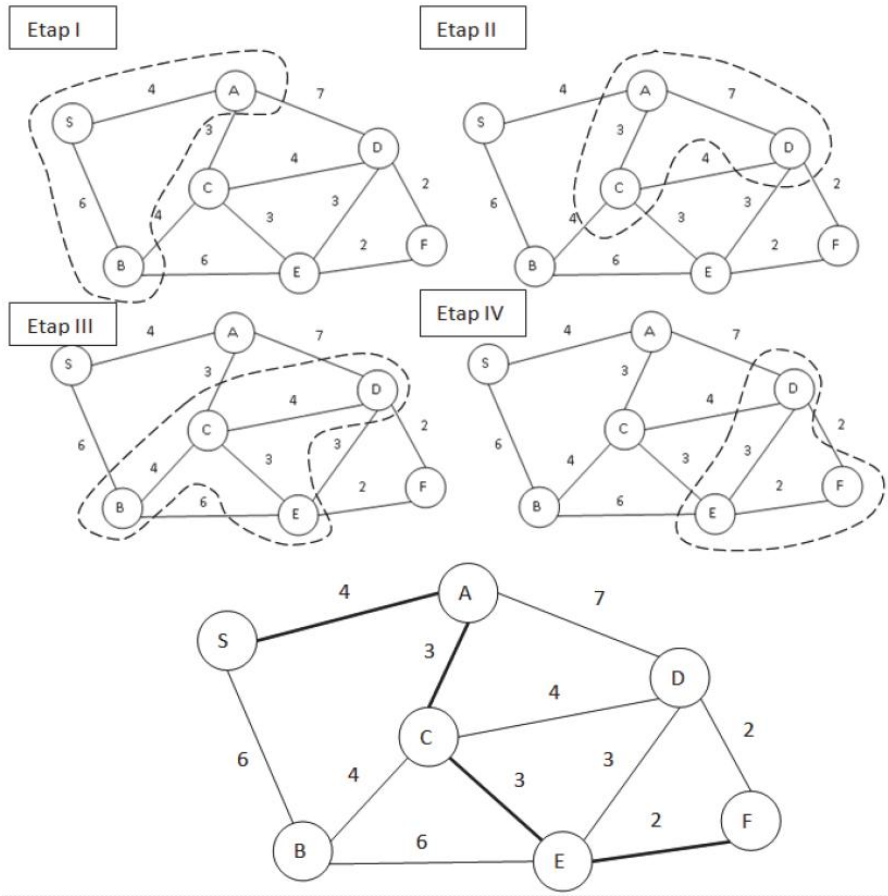
2. Podstawy teoretyczne

Głównym narzędziem wykorzystywanym w wyznaczaniu tras jest opracowany w 1959 r. przez Edsger'a Dijkstrę tzw. algorytm Dijkstry [1]. Służy on do wyznaczania najkrótszej ścieżki w grafie obejmującej wszystkie węzły – począwszy od węzła pierwszego (źródła) lub wyznaczeniu najkrótszej ścieżki między dwoma węzłami.

Algorytm Dijkstry dokonuje podziału zbioru wierzchołków grafu na trzy zbiory: V, Q, S. Pierwszy z nich V jest zbiorem wszystkich węzłów w grafie. Q jest zbiorem wszystkich wierzchołków o wspólnej krawędzi z węzłem źródłowym. Do zbioru S trafia węzeł ze zbioru Q o najkrótszej ścieżce od środka po czym przekształca się on w węzeł środkowy, a pierwotny węzeł źródłowy przestaje być uwzględniany w dalszym działaniu algorytmu. Działanie algorytmu kończy się, w momencie kiedy zbiór Q jest pusty bądź też kiedy dotrze do wcześniej zdefiniowanego węzła.

Algorytm Dijkstry dla wyznaczenia trasy między dwoma węzłami w grafie można opisać na następującym przykładzie – rysunek 1.

[1] andrzej.kepys@veo.glass, VEO Sp z o.o., https://veo.glass.
[2] ernest.syska@veo.glass, VEO Sp z o.o., https://veo.glass.



Rysunek 1. Procedura algorytmu Dijkstry [1]

Wierzchołek S stanowi węzeł źródłowy natomiast wierzchołek F końcowy. Każda krawędź w grafie ma nadaną wagę (nieujemną). Działanie algorytmu Dijkstry polega na znalezieniu wierzchołka połączonego z wierzchołkiem źródłowym wspólną krawędzią o najniższej wadze, co na rysunku stanowi połączenie węzła źródłowego S z węzłem A. Następnie węzeł połączony z węzłem źródłowym przejmuje jego rolę (A). Taka iteracja zachodzi do momentu dotarcia do punktu końcowego (S->A->C->E->F). Na rysunku przedstawiona została (pogrubioną linią) ścieżka o najniższej sumie wag (4 + 3 + 3 + 2 = 12). Ich suma wyniosła 12 i przedstawia optymalną trasę (minimalny koszt dotarcia) między węzłami S i F.

Algorytm Dijkstry jest powszechnie stosowany w programach do wyznaczania tras dla jednego pojazdu. Na bazie tego algorytmu pracuje m.in. system wyznaczania tras w Google Maps [2].

3. Adaptacja morska algorytmu Dijkstry

Aby wykonać adaptację algorytmu Dijkstry do nawigacji morskiej, konieczne jest uwzględnienie następujących aspektów:

1. Zapewnić sposób ważenia odcinków trasy adekwatnych do specyfiki nawigacji morskiej, a więc:

   a) Uwzględnić, że pewne krawędzie będą szybsze od innych dzięki optymalnemu dopasowaniu cech żeglownych jachtu i wiatru, który wieje w danym kwadracie geograficznym w danym czasie;
   b) Uwzględnić, że docieranie jachtu do granicy kwadratu geograficznego wymaga podania nowych danych pogodowych z kolejnego okresu czasu.
2. Zapewnić mapę zawierającą obszary, po których można pływać bezpiecznie bez wpłynięcia na ląd/płyciznę.
3. Zapewnić mapę zawierającą obszary, po których nie można pływać bezpiecznie:
   a) żeglowanie jest ryzykowne (np. płycizny, prognozowana w danym oknie czasowym burza);
   b) żeglowanie jest niekomfortowe (wysoka fala, silny rozkołys);
   c) żeglowanie jest zabronione (łowiska, strefy wojskowe, inne).
4. Uwzględnić wpływ prądów morskich pokrywających się z pewnymi krawędziami rejsu i dodającymi jachtu prędkości.
5. Uwzględnić znos jachtu, a więc konieczność ustawienia innego kąta żeglowania korygującą znos jachtu przez wiatr i falę.
6. Zapewnić tak szczegółową kwadratyzację mapy, aby możliwe było precyzyjne omijanie niewielkich przeszkód wodnych – np. obiektów o szerokości 100 metrów.
7. Zapewnić procedurę wyznaczania trasy w oparciu o podejście interacyjne – najpierw zgrubna topologia punktów brzegowych, później korekta o szczegółowe uwarunkowania mapowe i pogodowe.

Uwzględnienie części tych wytycznych opisano w dziale Implementacja. Na kolejnych rozwojach algorytmu wszystkie z nich zostaną w pełni zaimplementowane.

4. Implementacja

Rozwiązanie zostało zaimplementowane jako zestaw tabel oraz procedur składowanych bazy PostgreSQL z rozszerzeniem przestrzennym PostGIS oraz maszyną routingu PG_Routing.

Ze względów wydajnościowych działanie algorytmu podzielone zostało na dwie fazy:

- W fazie wstępnej przygotowujemy wszystkie dane globalne oraz niezmienne w czasie, wykonywane są również bardziej czasochłonne operacje, takie jak obliczanie zawartości lądu w oczku siatki, przygotowanie polar plotów jachtów, wyznaczenie krawędzi dla routingu wstępnego czy wyznaczenie wierzchołków grafu topologicznego używanego w późniejszych krokach do wyznaczania finalnej trasy;
- W fazie właściwej wykonywane są już tylko operacje niezbędne do wyznaczenia trasy i wyłącznie na fragmencie przestrzeni, w którym trasa będzie wyznaczana. Pozwala to na znaczne przyspieszenie działania algorytmu i nie generuje niepotrzebnych operacji na fragmentach przestrzeni, które nie będą odwiedzane.

W trakcie działania serwera wykonywane są również operacje cykliczne, takie jak pobieranie danych pogodowych i aktualizacja ich w bazie danych.

Proces obliczeniowy składa się z dwóch faz: fazy przygotowania środowiska oraz z fazy wyznaczenia trasy.

5. Faza przygotowania środowiska

5.1. Przygotowanie tabeli z lądami

Tabela inicjalna z lądami tworzona jest na podstawie danych Natural Earth [3], które są importowane do bazy danych za pomocą QGIS [4].

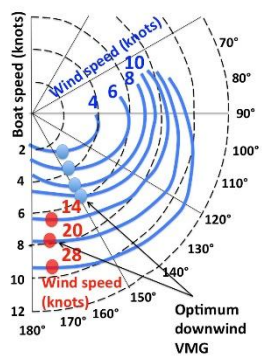
5.2. Przygotowanie tabeli z obszarami wykluczonymi

Obszary wykluczone tworzone są na podstawie danych projektu OpenStreetMap [5] przez społeczność zgromadzoną wokół projektu OpenSeaMap [6]. Ekstrakt danych jest pobierany i filtrowany przy pomocy aplikacji osmosis, po czym wynikowy plik importowany do bazy danych procesorem osm2pgsql. Zaimportowane dane przepisywane są do tabeli zgodnych ze schematem rozwiązania.

5.3. Przygotowanie tabeli z prędkościami jachtów w stosunku do wiatru

Polar plot to wykres wiążący cechy żeglowne jachtu z siłą i kierunkiem wiatru, który na niego oddziałuje, pozwalający odczytać jaką prędkość osiągnie jacht w przy danym kierunku i sile wiatru.

Przykładowy polar plot wygląda następująco:



Rysunek 2. Przykład polar plotu [7]

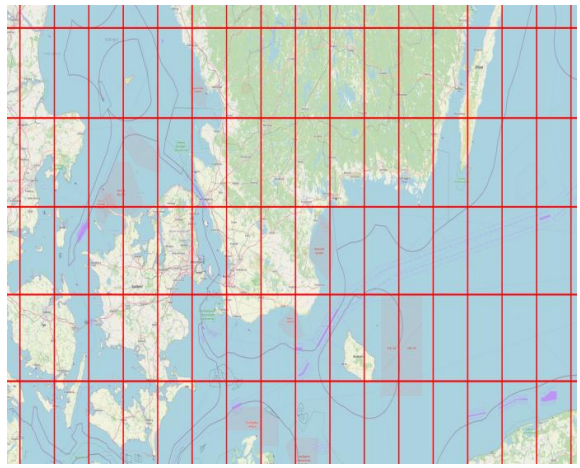
Tabela z prędkościami jachtów w stosunku do wiatru tworzona jest na podstawie danych dostępnych w repozytorium [8]. W ramach projektu utworzone zostało dedykowane narzędzie pozwalające na pobieranie tych danych i import ich do bazy w strukturze zgodnej ze schematem rozwiązania.

5.4. Przygotowanie tabeli z danymi pogodowymi

Na potrzeby rozwiązania utworzona została usługa, która cyklicznie pobiera aktualne dane pogodowe w formacie GRIB ze źródeł NOAA GFS [9], wyodrębnia z nich wartości konieczne do prawidłowego działania algorytmu, po czym importuje je do bazy danych.

5.5. Utworzenie tabeli z kwadratami o wielkości 0,5 stopnia

Na podstawie przeprowadzonych doświadczeń określiliśmy, że kompromisowa wielkość maksymalna kafelka to 0,5 stopnia (w przyszłości, w miarę zwiększania mocy obliczeniowych zmniejszymy ten kafelek co najmniej 10-krotnie). Wielkość ta użyta jest do utworzenia wstępnej siatki podziału.



Rysunek 3. Rozmiar kafli 0,5 stopnia [opracowanie własne]

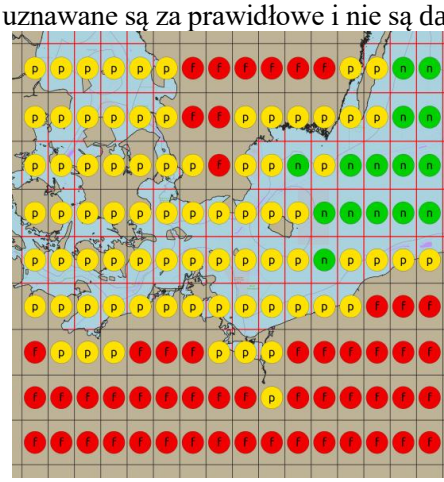
5.6. Wstawienie do tabeli fragmentów lądu przecinających się z kwadratem

Dla każdego oczka siatki wygenerowanego w poprzednim kroku wydzielany jest fragment lądu pokrywający się z nią. Geometria lądu w późniejszych krokach jest używana do określenia zawartości lądu w oczku, a podczas dalszego dzielenia siatki używana jest jako źródło lądu do mniejszych oczek siatki. To podejście znacznie skróciło czas generowania danych początkowych.

5.7. Określenie zawartości lądu w kwadracie

Na podstawie porównania pola powierzchni oczka siatki, pola powierzchni zawartego w niej lądu oraz obszarów wykluczonych określany jest poziom wypełnienia oczka siatki lądem w trzech wartościach:

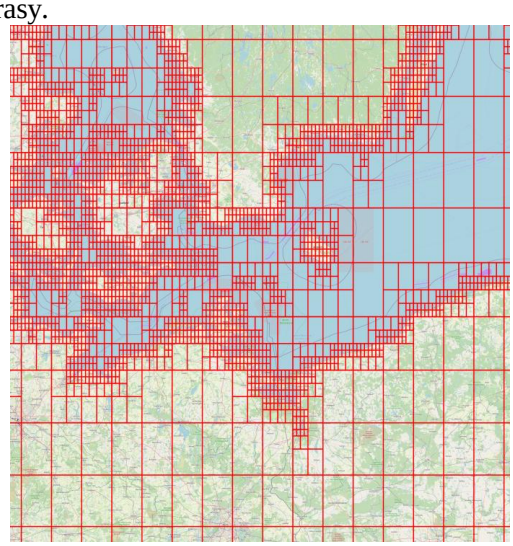
- f – full – te oczka będą pomijane przy dalszym processingu;
- p – partial – te oczka będą nadal dzielone w celu wyodrębnienia z nich części niezawierających lądu;
- n – none – te oczka uznawane są za prawidłowe i nie są dalej dzielone.



Rysunek 4. Identyfikacja kafli z lądem, z lądem i wodą, z samą wodą [opracowanie własne]

5.8. Rekurencyjne dynamiczne dzielenie kwadratów o częściowej zawartości lądu

Kolejny krok ma za zadanie ekstrakcję z oczek siatki częściowo zawierających ląd ich fragmentów, w których ląd nie występuje. Dzięki temu wydzielamy fragmenty przestrzeni, po których jacht może się bezpiecznie poruszać. Wszystkie oczka siatki o częściowej zawartości lądu i obszarów wykluczonych dzielone są dalej, oczka o pełnej zawartości lądu są odrzucane, a oczka bez lądu pozostawiane jako prawidłowe. Proces odbywa się w rekurencyjnej pętli aż do osiągnięcia doświadczalnie określonej kompromisowej wielkości minimalnej oczka siatki. Krok ten zakłada kompromis między czasem generowania zasobu inicjalnego, czasem generowania trasy oraz dokładnością tej trasy i możliwością wyznaczenia jej na obszarach gęsto pokrytych wyspami lub w wąskich cieśninach. W kolejnych fazach rozwoju aplikacji kompromis ten może być przesunięty w kierunku dokładności w miarę uzyskiwania wyższej wydajności na etapie wyznaczania trasy.



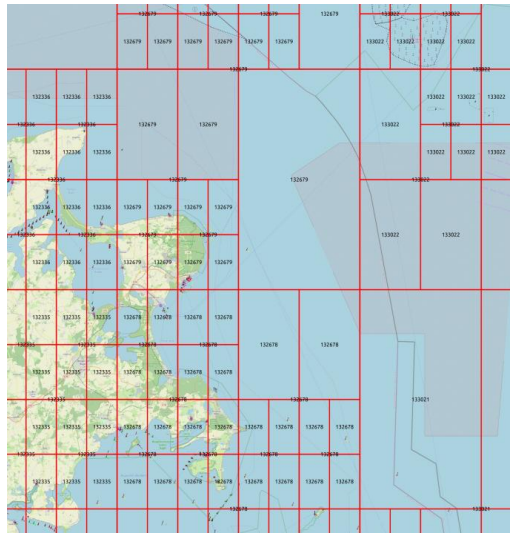
Rysunek 5. Rekurencyjne wydzielenie kwadratów z lądem i wodą na drobniejsze elementy [opracowanie własne]

5.9. Uzupełnienie relacji do tabeli zawierającej dane pogodowe

W kolejnym kroku na podstawie relacji przestrzennej tworzona jest relacja klucza obcego do tabeli zawierającej dane pogodowe. Identyfikator oczka tabeli pogodowej przepisywany jest na wszystkie elementy pochodne (krawędzie grafu topologicznego), co zapobiega wykonywaniu wielokrotnych złączeń przestrzennych i znacząco wpływa na wydajność końcowej procedury wyznaczania trasy.

5.10. Utworzenie tabeli ze zgrubnymi krawędziami

W kolejnym kroku między centroidami oczek siatki niezawierającymi lądu wyznaczany jest graf topologiczny używany w pierwszym kroku procedury wyznaczania trasy do określenia najkrótszej trasy z pominięciem warunków pogodowych.



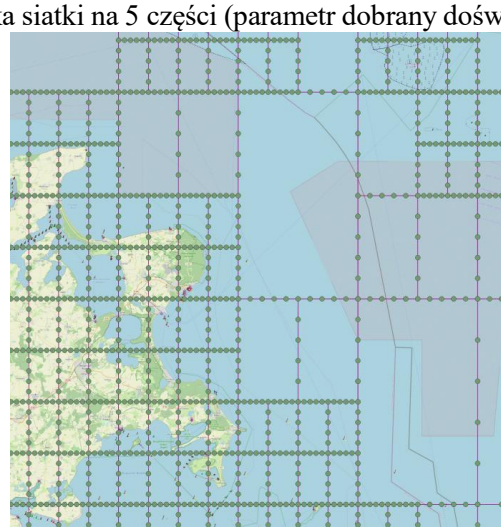
Rysunek 6. Dodanie do kafli powiązania z danymi pogodowymi [opracowanie własne]

5.11. Utworzenie głównej siatki

W kolejnym kroku na podstawie siatki zgrubnej tworzona jest siatka główna przez podział multilinii tworzących oczka siatki na poszczególne linie, które w kolejnych krokach będą dzielone i używane do utworzenia grafu topologicznego używanego do wyznaczania docelowej trasy uwzględniającej warunki pogodowe.

5.12. Utworzenie wierzchołków topologii

Jako ostatni krok fazy przygotowania danych na utworzonej siatce tworzone są punkty używane w kolejnych krokach wyznaczania trasy jako wierzchołki grafu topologicznego, po których odbywał się będzie routing. Punkty tworzone są przez podział każdej krawędzi oczka siatki na 5 części (parametr dobrany doświadczalnie).



Rysunek 7. Topologia punktów brzegowych kafli [opracowanie własne]

6. Faza wyznaczania trasy

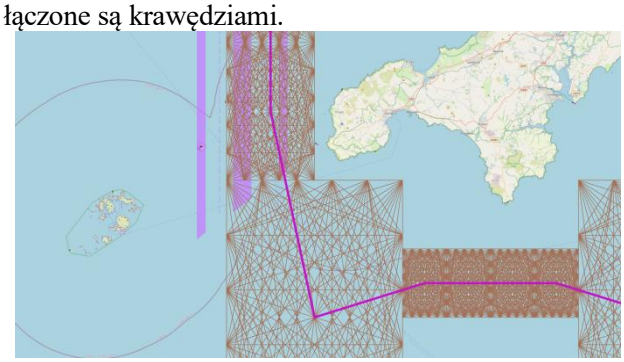
6.1. Wyznaczenie najkrótszej drogi po zgrubnej topologii

Pierwszym krokiem wyznaczania trasy jest wyszukanie oczek siatki przecinających się z punktem początkowym i końcowym oraz wyznaczenie zgrubnej trasy. Przeprowadzone doświadczenia pozwoliły na określenie, że wzrost długości trasy ze względu na warunki pogodowe opłacalny jest jedynie w zakresie 30%, dlatego trasa końcowa musi bezpośrednio korespondować z trasą najkrótszą. Jako koszt przebycia krawędzi zastosowana została wyłącznie jej długość, a trasa wyznaczona jest za pomocą funkcji implementującej algorytm Dijkstra w rozszerzeniu PG_Routing.

Rysunek 8. Uproszczona trasa po zgrubnej topologii punktów brzegowych [opracowanie własne]

6.2. Utworzenie tabeli z krawędziami

Na obecnym etapie wiemy już którędy zgrubnie prowadzi najkrótsza trasa między zadanymi punktami. W kolejnym kroku na terenie oczek siatki przebytych w ramach najkrótszej trasy wyznaczany jest gęsty graf topologiczny, po którym będziemy wyznaczać docelową trasę z uwzględnieniem warunków pogodowych. Wierzchołki topologii na terenie oczek przecinających trasę zgrubną utworzone w fazie przygotowania danych łączone są krawędziami.



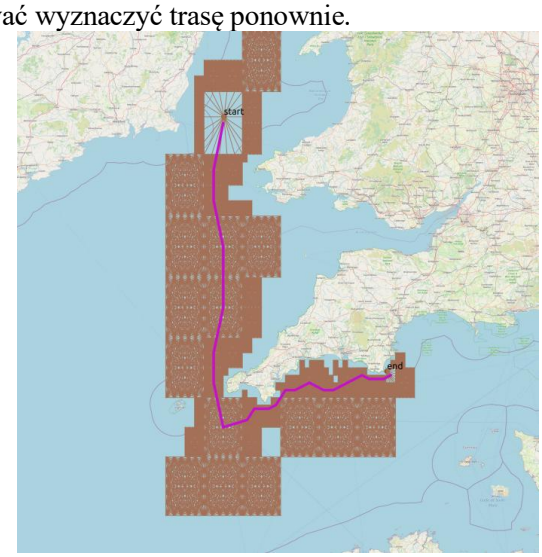
Rysunek 9. Szczegółowa topologia punktów brzegowych odniesiona do trasy zgrubnej [opracowanie własne]

6.3. Dołączenie do topologii punktów początkowego i końcowego trasy

Jako że poprzednie wyznaczenia wykonywane były na podstawie centroidów przebywanych oczek siatki do przeprowadzenia docelowego wyszukania trasy konieczne jest podłączenie właściwych punktów początku i końca trasy do grafu topologicznego. W tym celu krawędzie na terenie oczek siatki przecinających się z punktem początkowym i końcowym są usuwane, a tworzone są nowe krawędzie, które mają swój początek w punkcie początkowym trasy, a końce na uprzednio przygotowanych wierzchołkach topologii.

6.4. Rozszerzenie grafu o kafle sąsiadujące

Aby umożliwić wyznaczenie najszybszej trasy korespondującej z trasą zgrubną musimy rozszerzyć topologię o kafle sąsiadujące z tymi, po których przebiega ta trasa. Procedura rozszerzająca istniejący zestaw krawędzi przygotowana jest do działania rekurencyjnego i może wielokrotnie rozszerzać topologię. Ten punkt jest kolejnym miejscem kompromisu – każde rozszerzenie grafu wydłuża czas wyznaczania końcowej trasy, ale równocześnie zwiększa prawdopodobieństwo odnalezienia trasy znacznie efektywniejszej w większej odległości od trasy zgrubnej. W kolejnych krokach rozwoju algorytmu krok ten można będzie również uzależnić od wyniku wyznaczania – jeśli trasa docelowa nie zostanie odnaleziona można do niego wrócić, rozszerzyć topologię i spróbować wyznaczyć trasę ponownie.



Rysunek 10. Szczegółowa topologia punktów brzegowych odniesiona do trasy zgrubnej [opracowanie własne]

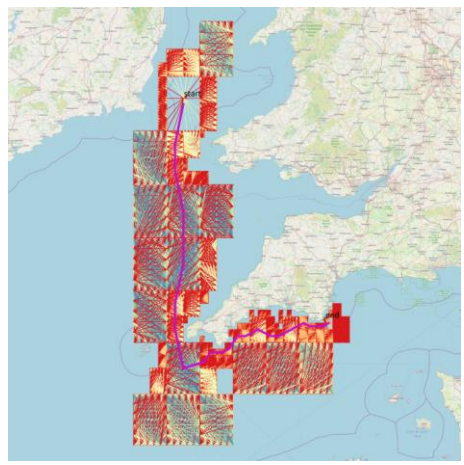
6.5. Dodanie do krawędzi informacji o pływach i wiatrach

Do wyznaczonych krawędzi dodajemy informacje o pływach i wiatrach na podstawie pobranych prognoz pogody. Informacja pogodowa przesunięta jest w czasie na podstawie szacowanego czasu dotarcia do każdej z krawędzi. W kolejnym kroku obliczany jest azymut każdej krawędzi oraz jej kąty w stosunku do wiatru zarówno w przypadku podróżowania zgodnie, jak i przeciwnie do jej kierunku – parametry te będą konieczne do określenia czasu przebycia krawędzi.

6.6. Obliczenie czasu przebycia każdej krawędzi w obu kierunkach

Ostatnią operacją konieczną do przygotowania danych do wyznaczenia trasy jest obliczenie czasu przebycia każdej krawędzi w obu kierunkach. Czasy przebycia krawędzi interpolowane są na podstawie tabeli średnich prędkości jachtów dla zadanego typu jachtu, siły i kierunku wiatru, która została zaimportowana do systemu w fazie przygotowania danych.

Na przedstawionym obrazku kolor krawędzi odzwierciedla czas, w jakim krawędź ta może być przebyta.



Rysunek 11. Wygenerowanie różnych tras i czasów przebycia oznaczonych kolorem [opracowanie własne]

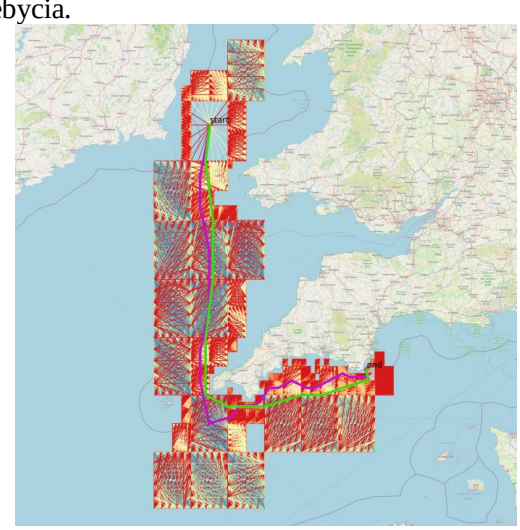
Na tym etapie najlepiej można dostrzec różnice z popularnym na rynku algorytmem izochronowym. Trasy wyznaczone algorytmem Dijkstry znajdują się w kaflach powiązanych z trasą zgrubną, na punktach brzegowych o wysokiej gęstości, podczas gdy w algorytmach izochronowych trasy te znajdują się w lejkach generowanych w pewnym kącie rozwarcia, bez powiązania z kaflami geograficznymi i ich punktami brzegowymi. Przykład lejka izochronowego w programie MAX SEA TIMEZERO:

Rysunek 12. Algorytm izochronowy dla podobnej trasy w programie MAX SEA TIMEZERO [opracowanie własne]

6.7. Wyszukanie najszybszej trasy po krawędziach

Zwieńczeniem procesu przygotowania danych jest wyznaczenie trasy docelowej na przygotowanych krawędziach. Do wyznaczenia trasy – tak samo jak w przypadku trasy zgrubnej – używana jest procedura implementująca algorytm Djikstry z rozszerzenia PG_Routing. W tym przypadku kosztem przebycia krawędzi jest czas jej przebycia obliczony w poprzednim kroku.

Poniższy obrazek przedstawia wyznaczoną trasę zgrubną na tle krawędzi z obliczonym czasem ich przebycia.



Rysunek 13. Wyznaczenie trasy optymalnej (kolor jasnozielony) [opracowanie własne]

Porównanie trasy zgrubnej oraz trasy docelowej na tle danych pogodowych przedstawia się następująco.

Rysunek 14. Trasa zgrubna i trasa zoptymalizowana (jasnozielona) wyznaczona automatycznie [opracowanie własne]

7. Przykłady obliczeń

Poniżej przedstawiono 3 kalkulacje dla przykładowych tras, dla jachtu Salona 44, dla przykładowej pogody dla wybranej daty.

Parametr „Czas rejsu na trasie uproszczonej" reprezentuje w przybliżeniu czas rejsu dla trasy wyznaczonej ręcznie, w oparciu o manualnie wyznaczone waypointy, w przeliczeniu na typową prędkość jachtu turystycznego tej klasy. Parametr „Czas rejsu na trasie zoptymalizowanej" przedstawia efekt działania algorytmu VEO Weather Routing dla kwadratyzacji map rzędu 0,5 stopnia (wraz z rekurencyjną korektą dla zadanego obszaru mapy). Docelowo kwadratyzacja map sięgnie 0,005 stopnia przy co najmniej 20-krotnym przyśpieszeniu obliczeń dzięki zastosowaniu procesorów GPU o bardzo wysokich parametrach przetwarzania. Istnieje dalszy potencjał na zmniejszenie kwadratów mapy podlegających analizie i dalsze przyspieszenie algorytmów (zwiększania mocy obliczeniowej komputera głównego).

Parametr „Czas kalkulacji algorytmu na środowisku klasy średni PC" należy traktować jako informację porównawczą, dającą jedynie wyobrażenie o typowym obciążeniu obliczeniowym z perspektywy przeciętnych komputerów osobistych.

Tabela 1. Porównanie wyników działania algorytmu VEO Weather Routing (źródło – obliczenia własne)

Trasa morska od-do	Czas rejsu na trasie uproszczonej (topologia zgrubna)	Czas rejsu na trasie zoptymalizowanej (topologia szczegółowa)	Czas kalkulacji algorytmu na środowisku klasy średni PC
Gdynia (marina) – Łeba (port) – ok. 80 mil	16,5 h	14,3 h	12,8 s
Ystad (Szwecja) – Rosklide (Dania) – ok. 150 mil	31,2 h	26,1 h	15,4 s
Krk (miasto, wyspa KRK Chorwacja) – Gujak, Kornat (Chorwacja) – ok. 80 mil	16,8 h	13,4 h	12,8 s

Źródło: [opracowanie własne]

8. Podsumowanie

W artykule wykazano przydatność algorytmu Dijkstry w adaptacji morskiej zbudowanej przez autorów o nazwie VEO Weather Routing. Mechanizm ten służy do rozwiązywania problemu wyznaczania trasy morskiej pod kątem danych pogodowych, cech żeglownych jachtu – z automatycznym omijaniem lądów (i stref wyłączonych z żeglugi) o wysokiej efektywności i precyzji. Przedstawiono pełną procedurę obliczeniową w obu fazach działania algorytmu. W efekcie udowodniono istotny uzysk czasowy na trasie wyznaczonej za pomocą zbudowanego mechanizmu względem trasy wyznaczonej ręcznie.

Literatura

1. Dijkstra E.W., *A note on two problems in connexion with graphs*, Numerische mathematik, 1(1), 1959, s. 269-271.
2. Lanning D.R., Harrell G.K., Wang J., *Dijkstra's algorithm and Google maps*, https://www.researchgate.net/publication/296639227_Dijkstra's_algorithm_and_Google_maps [data dostępu 1.11.2021].
3. Natural Earth Data, https://www.naturalearthdata.com/ [data dostępu 1.11.2021].

4. QGIS, https://www.qgis.org/pl/site/ [data dostępu 1.11.2021].
5. OpenStreetMaps, https://www.openstreetmap.org/ [data dostępu 1.11.2021].
6. OpenSeaMaps, https://www.openseamap.org/index.php?id=openseamap&no_cache=1 [data dostępu 1.11.2021].
7. Wikipedia, https://en.wikipedia.org/wiki/Polar_diagram_(sailing) [data dostępu 1.11.2021].
8. JIETER, http://jieter.github.io/orc-data/site/ [data dostępu 1.11.2021].
9. NOAA, https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast [data dostępu 1.11.2021].

VEO Weather Routing – nowatorski algorytm wyznaczania tras w oparciu o dane pogodowe, mapowe i żeglowne jachtów

Streszczenie

W artykule przedstawiono algorytm Dijkstry w adaptacji morskiej o nazwie VEO Weather Routing. Mechanizm ten służy do rozwiązywania problemu wyznaczania trasy morskiej pod kątem danych pogodowych, cech żeglownych jachtu – z automatycznym omijaniem lądów o wysokiej efektywności i precyzji. Przedstawiono pełną procedurę obliczeniową w obu fazach działania algorytmu. W efekcie udowodniono istotny uzysk czasowy na trasie wyznaczonej za pomocą zbudowanego mechanizmu względem trasy wyznaczonej ręcznie.

Słowa kluczowe: pogodowe wyznaczanie tras morskich, nawigacja morska, nawigacja jachtowa

VEO Weather Routing – an innovative algorithm for calculating routes based on weather, map and navigable data of yachts

Abstract

The article presents the Dijkstra algorithm in a marine adaptation called VEO Weather Routing. This mechanism is used to solve the problem of determining the sea route in terms of weather data, navigable characteristics of the yacht - with automatic avoiding land with high efficiency and precision. The full calculation procedure in both phases of the algorithm's operation was presented. As a result, a significant time gain was proved on the route marked with the use of the constructed mechanism in relation to the route marked by hand.

Keywords: weather routing, sea navigation, yacht navigation

Mariusz Bedla[1], Adam Krechowicz[2], Stanisław Deniziak[3], Grzegorz Łukawski[4], Paweł Pięta[5], Daniel Kaczmarski[6], Mateusz Pawełkiewicz[7], Mirosław Borkowski[8]

Zastosowanie metodyki big data w diagnostyce pacjentów z mózgowym porażeniem dziecięcym

1. Wprowadzenie

Niniejsza praca dotyczy zastosowania metodyki Big Data w diagnostyce pacjentów z mózgowym porażeniem dziecięcym, poprzez analizę danych biometrycznych uzyskanych za pomocą urządzeń biomechanicznych Biodex System 4 i Zebris FDM-T. Praca omawia również krótko możliwość użycia w tym kontekście platform stabilometrycznych firmy AC International East.

Mózgowe porażenie dziecięce (w skrócie MPD) to złożone, wielowymiarowe schorzenie, będące najczęstszą przyczyną niepostępującej niepełnosprawności ruchowej i zaburzenia postawy u dzieci [1-3]. Objawy MPD różnią się w zależności od osoby i mogą zmieniać się w czasie [2]. Mózgowemu porażeniu dziecięcemu towarzyszą często liczne choroby współistniejące, takie jak: osłabione zdolności komunikacyjne, zaburzenia mowy, zaburzenia wzroku, obniżona czujność, upośledzenie funkcji poznawczych i behawioralnych, trudności w przyjmowaniu pokarmów i płynów, problemy z układem mięśniowo-szkieletowym, drgawki, czy padaczka [1, 2]. Obecnie nie jest znane lekarstwo na MPD, jednak pewne formy leczenia wspomagającego, takie jak fizjoterapia, pomagają złagodzić objawy choroby i poprawiają zdolności motoryczne pacjentów [3].

Na przestrzeni ostatnich kilku lat były prowadzone liczne badania związane z mózgowym porażeniem dziecięcym. W pracy [1] wykazano istotny związek pomiędzy stopniem upośledzenia zdolności komunikacyjnych a nasileniem MPD. W pracy [4] przeprowadzono sekwencjonowanie całego eksomu dla 183 przypadków MPD, wykazując znaczną niejednorodność genetyczną tego schorzenia. W pracy [5], podczas rehabilitacji metodą CIMT do zbierania danych na temat aktywności kończyn górnych pacjentów z MPD, zostały wykorzystane po raz pierwszy akcelerometry. W pracy [6]

[1] mbedla@tu.kielce.pl, Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, Katedra Systemów Informatycznych, www.tu.kielce.pl.
[2] adeik.k@gmail.com, Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, Katedra Systemów Informatycznych, www.tu.kielce.pl.
[3] s.deniziak@tu.kielce.pl, Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, Katedra Systemów Informatycznych, www.tu.kielce.pl.
[4] g.lukawski@tu.kielce.pl, Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, Katedra Systemów Informatycznych, www.tu.kielce.pl.
[5] p.pieta@tu.kielce.pl, Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, Katedra Systemów Informatycznych, www.tu.kielce.pl.
[6] d.kaczmarski@tu.kielce.pl, Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, Katedra Systemów Informatycznych, www.tu.kielce.pl.
[7] mpawelkiewicz@tu.kielce.pl, Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, Katedra Systemów Informatycznych, www.tu.kielce.pl.
[8] twec-m.borkowski@tu.kielce.pl, Politechnika Świętokrzyska, www.tu.kielce.pl.

zaproponowano schemat umożliwiający opracowanie klinicznych modeli nerwowo-mięśniowo-szkieletowych pacjentów z MPD na podstawie cech wyodrębnionych za pomocą elektromiografii. Nieustannie powiększa się wolumen danych opisujących pacjentów z MPD i rośnie różnorodność tych danych.

Istotnym problemem związanym z mózgowym porażeniem dziecięcym jest rozproszenie i niespójność danych opisujących tę chorobę, jak również brak zintegrowanych systemów informatycznych dedykowanych temu schorzeniu, które ułatwiłyby obserwację pacjentów oraz pomogły w określeniu odpowiedniego planu opieki i dobraniu właściwego procesu leczenia [2]. System informatyczny o nazwie Mobilize Center [7] adresuje problem rozproszenia danych opisujących ruch człowieka, jednak obecnie jego zasoby nie zawierają informacji na temat osób z MPD. Ponadto praca ta wskazuje analitykę Big Data jako potencjalną metodykę, która może zrewolucjonizować badania nad mobilnością człowieka. Metodyka ta została zastosowana w pracy [8], w której baza danych podstawowej opieki zdrowotnej Wielkiej Brytanii (*Clinical Practice Research Datalink*) została poddana analizie algorytmem uczenia maszynowego w celu zidentyfikowania potencjalnych przypadków mózgowego porażenia dziecięcego. Po zweryfikowaniu przez ekspertów danych medycznych dzieci wskazanych przez algorytm, w przypadku 23,3% z nich potwierdzono MPD. W pracy [9] dane z lat 2009-2019, które znajdowały się w elektronicznych kartach pacjentów dzieci z mózgowym porażeniem dziecięcym, zostały poddane klasteryzacji w celu zidentyfikowania grup pacjentów korzystających z podobnych świadczeń zdrowotnych. W innej pracy [10] poddano z kolei analizie dane z lat 2016-2017 pochodzące z ogólnodostępnej bazy National Inpatient Sample, zawierającej informacje na temat hospitalizacji pacjentów w Stanach Zjednoczonych. Metodyka Big Data została tu zastosowana w celu oszacowania śmiertelności i wykorzystania zasobów szpitalnych przez hospitalizowanych pacjentów z MPD, szczególnie tych, którzy cierpieli na niedożywienie.

Rosnąca ilość danych uzyskiwanych na skutek badań biomechanicznych człowieka wiąże się z koniecznością opracowania zaawansowanych metod analizy wielowymiarowej i uczenia maszynowego. W pracy [11] przedstawiono zautomatyzowany system 3D GAIT, który służy do gromadzenia trójwymiarowych danych opisujących chód. Ponadto w pracy tej przedstawiono przegląd badań z zakresu analizy chodu opartej na metodach wielowymiarowych i uczenia maszynowego, które można zastosować do analizy Big Data (ekstrakcja i wybór cech, klasyfikacja, grupowanie, analiza danych topologicznych). W pracy [3] wskazano również liczne przykłady badań w tym zakresie, ukierunkowane na pacjentów z mózgowym porażeniem dziecięcym. W pracy [12] opisano prototyp urządzenia wykonanego w architekturze Wearable Internet of Things (WIoT), służącego do śledzenia aktywności fizycznej człowieka, które może zostać wykorzystane do analizy chodu osób z MPD. W pracy tej pokazano, że uzyskane dane mogą zostać użyte do wytrenowania konwolucyjnej sieci neuronowej. Istotnym kierunkiem badań jest również szukanie korelacji pomiędzy różnymi metrykami określającymi jakość chodu, takimi jak GDI czy GGI, a innymi danymi opisującymi stan zdrowia pacjentów z MPD [13-15].

Ze względu na różnorodność form niepełnosprawności trudno jest jednoznacznie ocenić jakość chodu osób z MPD. Ocena ta jest niezbędna do diagnostyki i weryfikacji skuteczności procesu rehabilitacji. Najbardziej zaawansowane metody bazują na wyliczeniu metryk stanowiących kompilację wybranych cech kinematycznych chodu.

Jednak pomiar takich cech wymaga stosowania zaawansowanych urządzeń do optycznej analizy chodu, które są dostępne jedynie w zaawansowanych laboratoriach analizy ruchu. Dlatego w pracy podjęto próbę aproksymacji wartości tych metryk na podstawie pomiarów biometrycznych wykonanych za pomocą dostępnych w ośrodkach rehabilitacyjnych urządzeń biomechanicznych, takich jak bieżnie, platformy czy dynamometr izokinetyczny. Metodyka badań obejmuje:

1. wykonanie pomiarów biometrycznych dla reprezentatywnej grupy pacjentów z MPD;
2. ocenę stopnia niepełnosprawności ww. pacjentów za pomocą wybranych metryk z wykorzystaniem optycznej analizy chodu;
3. wyszukiwanie modelu dokonującego predykcji ww. metryk, na podstawie wybranych pomiarów biometrycznych, metodami uczenia maszynowego: klasteryzacją i analizą regresji;
4. predykcję metryk dla ww. pacjentów i określenie błędu predykcji.

W rozdziale 2 zostały omówione metryki stosowane w diagnostyce chodu osób z MPD. W rozdziale 3 scharakteryzowano urządzenia diagnostyczne służące do oceny biomechanicznej pacjenta, które zostały wykorzystane w ramach tej pracy. W rozdziale 4 przedstawiono wyniki biometrycznej diagnostyki chodu pacjentów z MPD uzyskane z zastosowaniem danych z urządzeń omówionych w poprzednim rozdziale. Rozdział 5 zawiera wnioski sformułowane na podstawie wykonanych badań.

2. Metryki stosowane w diagnostyce chodu osób z MPD

Najczęściej stosowane metryki w diagnostyce chodu pacjentów z mózgowym porażeniem dziecięcym to GGI i GDI. Metryka GGI (ang. *Gillette Gait Index*, wcześniej *Normalcy Index* – NI) to wskaźnik określający stopień normalności chodu. Jest on wyrażony w postaci jednej liczby, która jest traktowana jako miara odległości między zestawem dyskretnych parametrów chodu danego pacjenta, a średnimi parametrami chodu osoby zdrowej. Zestawy parametrów mogą zostać przedstawione w postaci wektorów w wielowymiarowej przestrzeni, których liczba składowych wyznacza wymiary tej przestrzeni (długość wektorów). W celu uniknięcia błędów wynikających na przykład ze współzależności niektórych zmiennych i stosowania różnych jednostek, do obliczenia odległości między zestawami danych wykorzystuje się metody statystyki wielowymiarowej. GGI jest najczęściej wykorzystywanym wskaźnikiem do analizy chodu dzieci z MPD [16]. Wskaźnik GGI oblicza się osobno dla kończyny dolnej prawej i lewej. Średnia z obu tych wartości stanowi ogólną wartość GGI wzorca chodu pacjenta. GGI jest obliczany na podstawie arbitralnie wybranych szesnastu parametrów kinematycznych chodu [16]. W przypadku dzieci z MPD, obliczenie globalnego wskaźnika odchyleń chodu pozwala na precyzyjne określenie stopnia zaburzeń chodu u każdego dziecka indywidualnie. Dzięki temu możliwe staje się wykonywanie porównań międzyosobniczych, co ze względu na olbrzymie zróżnicowanie objawów MPD nie było dotąd możliwe [16].

Metryka GDI (ang. *Gait Deviation Index*) to wskaźnik odchylenia chodu bazujący na wielkościach kinematycznych. Podobnie jak metryka GGI został on wprowadzony jako miara ogólnej patologii chodu. Definiowany jest jako miara odległości pomiędzy kinematyką patologicznego chodu a normalnego wzorca chodu [17]. W celu wyznaczenia wskaźnika GDI branych jest pod uwagę 9 zmiennych kinematycznych chodu

opisujących: kinematykę stawu biodrowego i miednicy we wszystkich trzech płaszczyznach, kinematykę stawu kolanowego i skokowego w płaszczyźnie strzałkowej, oraz progresję stopy. Wartość GDI wskazuje na odległość chodu badanej osoby od chodu prawidłowego [18]. Wskaźnik GDI rozpatrywany jest niezależnie dla każdej kończyny dolnej. Otrzymany wynik GDI należy interpretować w następujący sposób:

- gdy GDI wynosi 100, to chód badanej osoby jest zbliżony do chodu grupy kontrolnej;
- każda zmiana wskaźnika GDI o 10 poniżej 100 oznacza jedno odchylenie standardowe od średniej grupy kontrolnej [19].

Metryki GGI i GDI posiadają wiele zalet, m.in. umożliwią:

- klasyfikację wzorców chodu;
- ocenę nasilenia odchyleń chodu;
- monitorowanie stanu pacjenta w trakcie leczenia;
- porównywanie różnych metod leczenia [20].

Wadą metryk GGI i GDI jest konieczność wykonania dokładnego pomiaru parametrów kinematycznych chodu. Zwykle dokonuje się to przy pomocy zaawansowanych systemów do analizy ruchu metodą optoelektryczną (np. system BTS Smart[9], Vicon[10]). Zatem ocena chodu pacjenta z wykorzystaniem tych metryk jest możliwa jedynie w wyspecjalizowanych laboratoriach analizy ruchu.

3. Diagnostyczne urządzenia do oceny biomechanicznej pacjenta

3.1. Biodex System 4

Urządzenie Biodex System 4[11] pozwala na przeprowadzanie badań diagnostycznych oraz rehabilitacji zarówno kończyn górnych, jak i dolnych. W zależności od konfiguracji urządzenia pozwala ono na badanie kolana, kostki, biodra, ramienia, łokcia oraz nadgarstka. Z uwagi na możliwość szczegółowego badania stawów kończyny dolnej urządzenie wykorzystywane jest do diagnozy oraz rehabilitacji pacjentów z MPD.

Podczas wykonywania ruchu badanym stawem urządzenie Biodex System 4, dzięki posiadanym czujnikom, umożliwia jednoczesny pomiar następujących wartości:

- pozycja kończyny;
- prędkość poruszania kończyną;
- moment obrotowy.

Wykorzystując te podstawowe parametry oraz dodatkowo dysponując czasem trwania ruchu urządzenie pozwala obliczyć także dodatkowe parametry ruchu kończyny, takie jak np.: zakres ruchów, całkowita praca wykonana przez kończynę, maksymalny moment obrotowy, czas wymagany do osiągnięcia maksymalnego momentu obrotowego itp. Dodatkowo urządzenie pozwala na agregację osobnych wyników dla ruchu prostowania kończyny oraz zginania kończyny.

W ramach niniejszej pracy pacjenci z MPD zostali poddani badaniu na urządzeniu Biodex System 4. Sesja diagnostyczna obejmowała badanie stawu kolanowego pracującego z trzema różnymi prędkościami obrotowymi (60, 90 oraz 180 stopni na sekundę).

[9] https://www.btsbioengineering.com/products/smart-dx/.
[10] https://www.vicon.com/.
[11] https://www.biodex.com/physical-medicine/products/dynamometers/system-4-pro.

3.2. Bieżnia Zebris FDM-T

Bieżnia Zebris FDM-T[12] jest bardzo uniwersalnym urządzeniem rehabilitacyjno-diagnostycznym umożliwiającym różne tryby pracy. Może zostać wykorzystana zarówno do analizy dynamicznej chodu, jak również do analizy statycznej podczas stania pacjenta. Istnieje możliwość zastosowania bieżni na potrzeby diagnostyki oraz do rehabilitacji kończyn dolnych pacjentów z MPD.

Do urządzenia można podłączyć zastaw do elektromiografii (EMG), kamerę do rejestracji treningu, projektor służący do wyświetlania śladów stóp na bieżni, za którymi pacjent powinien podążać lub monitor, na którym wyświetlany jest konfigurowalny tor z przeszkodami, które pacjent powinien omijać.

Wyniki pomiarów mogą być eksportowane w postaci tekstowej, graficznej oraz wideo. Do podstawowych parametrów, które mogą zostać zmierzone, można zaliczyć wartości siły i nacisku stóp oraz parametry czasowo-przestrzenne chodu.

W niniejszej pracy dla urządzenia Zebris FDM-T analizie poddane zostały wyniki z pochodzące z następujących pomiarów: analiza statyczna podczas stania z oczami otwartymi i zamkniętymi oraz analiza chodu dla prędkości preferowanej i maksymalnej.

3.3. Platformy stabilometryczne

Platformy stabilometryczne to zaawansowane urządzenia wykorzystywane podczas rehabilitacji osób cierpiących na zaburzenia zmysłu równowagi. Ich działanie polega przede wszystkim na monitorowaniu postawy badanego pacjenta. W ramach pracy zostały wykorzystane dwa typy platform: AC East Alfa[13] (jednopłytowa) oraz AC East Gamma[14] (dwupłytowa).

3.3.1. AC East Alfa

AC East Alfa to jednopłytowa platforma stabilometryczna umożliwiająca wyłącznie trening oraz ocenę równowagi pacjenta. Dzięki urządzeniu możliwe jest zwiększenie sprawności pacjentów posiadających dysfunkcje mięśniowe powstałe w wyniku wszelkiego rodzaju urazów głowy, po udarach, a także u osób chorych na mózgowe porażenie dziecięce. Trening na platformie Alfa ma na celu stymulację elementów układu ruchu i układu nerwowego odpowiedzialnego m.in. za równowagę.

Urządzenie to współpracuje z oprogramowaniem, które działa w oparciu o czujnik ruchu rozpoznający ruchy poszczególnych części ciała we wszystkich płaszczyznach (XYZ). Dzięki użyciu platformy Alfa możliwa jest:

- ocena parametrów statycznych i dynamicznych związanych z utrzymaniem równowagi na stabilnym podłożu;
- analiza COP (ang. *Center of Pressure*), czyli położenia środka nacisku stóp.

3.3.2. AC East Gamma

AC East Gamma to dwupłytowa platforma dynamograficzna umożliwiająca przeprowadzanie ćwiczeń oraz treningów dla pacjentów posiadających problemy z koordynacją ciała, w szczególności dla osób chorych na mózgowe porażenie dziecięce.

[12] https://www.zebris.de/en/medical/dynamic-gait-analysis-on-the-treadmill.
[13] http://acinternational-east.pl/pl/alfa/.
[14] http://acinternational-east.pl/pl/gamma/.

Ćwiczenia na tym urządzeniu odbywają się z zastosowaniem gier komputerowych, co wpływa pozytywnie na motywację do ich wykonywania. Platforma Gamma umożliwia:

- ocenę równowagi i balansu pacjenta;
- analizę dystrybucji obciążenia w osi pionowej, dynamicznego i statycznego pomiaru obciążenia.

4. Biometryczna diagnostyka chodu pacjentów z MPD

Ponieważ urządzenia diagnostyczne przedstawione w p. 3 nie umożliwiają pozyskania danych wymaganych do obliczenia wartości metryki GGI lub GDI, w ramach pracy podjęto próbę predykcji wartości tych parametrów na podstawie pomiarów biometrycznych wykonywanych na tych urządzeniach. Takie podejście, naszym zdaniem, jest uzasadnione ze względu na to, że stosowane urządzenia dokonują pomiarów kinematycznych i dynamometrycznych kończyn dolnych i jest wysoce prawdopodobne, że istnieje korelacja pomiędzy wynikami tych pomiarów a parametrami stosowanymi w obliczaniu metryk GGI i/lub GDI.

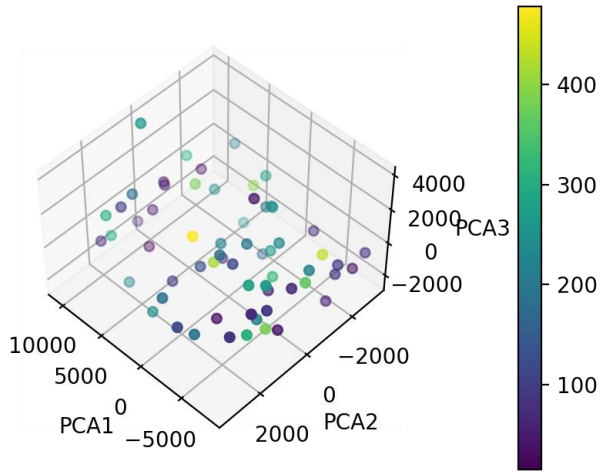
4.1. Diagnostyka za pomocą Biodex System 4

Ocena jakości chodu pacjentów z MPD została przeprowadzona na pozyskanej próbie 31 wyników badań pacjentów reprezentujących różne natężenie schorzeń. Z uwagi na specyfikę badań na urządzeniu, pozwalających na niezależne badanie każdej kończyny, możliwe było potraktowanie kończyny lewej oraz kończyny prawej jako osobnych wyników. Pozwoliło to na poszerzenie dostępnego zbioru do 62 próbek.

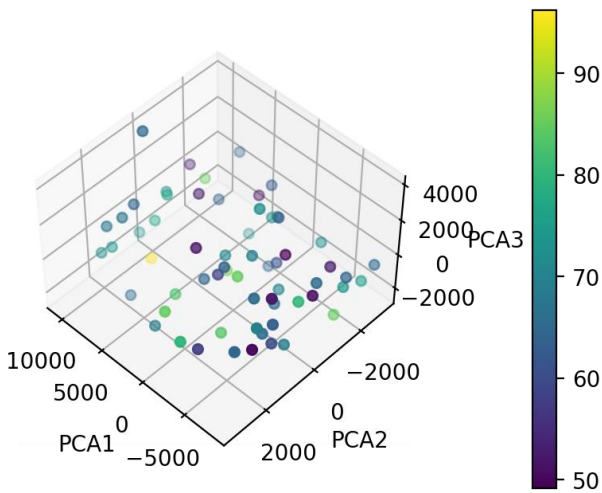
Pozyskane dane zostały poddane analizie głównych składowych. Wizualizacja tej analizy w przestrzeni trójwymiarowej zwizualizowana została na rysunku 2. i 3. Rysunek 2. przedstawia wykonaną analizę w odniesieniu do metryki GGI, natomiast rysunek 3. – analizę w odniesieniu do metryki GDI. Kolor próbki w przestrzeni trójwymiarowej odpowiada wartości odpowiedniej metryki dla danej kończyny pacjenta.

Biorąc pod uwagę metryki GDI oraz GGI analiza rysunku 2. oraz 3. wykazała, że możliwa jest identyfikacja punktów skupień odpowiadających kończynom pacjentów reprezentujących podobną jakość chodu. Pozwala to założyć, że możliwa jest ocena jakości chodu na podstawie danych pochodzących z urządzenia Biodex System 4.

Uwzględniając wartość metryki GDI oraz GGI i na podstawie pozyskanych danych dokonano próby klasyfikacji pacjentów do jednej z trzech grup. Analizę wykonano dla następujących metod klasyfikacji: maszyna wektorów nośnych (SVM), lasy losowe (RFC), klasyfikacja z wykorzystaniem regresji logistycznej (LRC), metoda k-najbliższych sąsiadów (KNN), klasyfikator naiwny Bayesowski Gaussa (GNB) oraz metodę wzmocnionych gradientów (ABC). Ewaluację stworzonych modeli dokonano z wykorzystaniem standardowych miar jakości modeli uczenia maszynowego (dokładności, precyzji, pełności oraz współczynnika f1). Uzyskane wyniki zostały przedstawione w tabeli 2.



Rysunek 2. Wizualizacja analizy głównych składowych danych pozyskanych z urządzenia Biodex System 4 w odniesieniu do metryki GGI [opracowanie własne]



Rysunek 3. Wizualizacja analizy głównych składowych danych pozyskanych z urządzenia Biodex System 4 w odniesieniu do metryki GDI [opracowanie własne]

Tabela 2. Klasyfikacja z wykorzystaniem danych z urządzenia Biodex System 4 Pro

Metoda	Metryka	Dokładność	Precyzja	Pełność	f1
SVM	GGI	30%	64%	30%	41%
SVM	GDI	65%	65%	65%	65%
RFC	GGI	39%	53%	39%	45%
RFC	GDI	78%	78%	78%	78%
LRC	GGI	35%	52%	43%	48%
LRC	GDI	61%	61%	61%	61%
KNN	GGI	34%	66%	35%	46%
KNN	GDI	65%	65%	65%	65%
GNB	GGI	39%	63%	43%	51%
GNB	GDI	74%	74%	74%	74%

QDA	GGI	35%	53%	43%	48%
QDA	GDI	56%	56%	56%	56%
ABC	GGI	17%	46%	26%	33%
ABC	GDI	74%	74%	74%	74%

Źródło: Opracowanie własne

Z przedstawionych w tabeli 2. wartości wynika, że klasyfikacja w oparciu o metrykę GDI pozwala na uzyskanie znaczne lepszych wyników. Podczas przewidywania jakości chodu najlepiej sprawdził się model lasów losowych, uzyskując dokładność na poziomie 78%. Wartość ta oraz zastosowanie klasyfikacji jedynie do trzech grup zostawia jednak wiele miejsca na udoskonalenie.

Pozyskane z urządzenia Biodex System 4 dane wykorzystano następnie do próby predykcji dokładnej wartości metryk GGI oraz GDI z wykorzystaniem metod regresji. W tym celu wykorzystano następujące metody uczenia maszynowego: regresja z wykorzystaniem maszyny wektorów nośnych (SVR), regresja z wykorzystaniem metody k-najbliższych sąsiadów (KNN), regresja z wykorzystaniem drzew decyzyjnych (DTR) oraz regresja z wykorzystaniem metody wzmocnionego gradientu (GBR). Wyniki zostały ocenione z wykorzystaniem wartości średniego błędu kwadratowego (MSE), średniego błędu procentowego (MAPE) oraz współczynnika determinacji (R2). Uzyskane wyniki zebrano w tabeli 3.

Z przedstawionych w tabeli 3. wartości wynika, że regresja wartości metryki GGI nie przyniosła zadowalających efektów. Najmniejszy średni błąd procentowy (58%) udało się uzyskać dla regresji z wykorzystaniem metody wzmocnionego gradientu. Znacząco lepsze wyniki udało się uzyskać podczas regresji wartości metryki GDI osiągając średni błąd procentowy na poziomie 11% dla regresji z wykorzystaniem drzew decyzyjnych. Chociaż taką wartość można uznać za zadowalającą to wartości współczynnika determinacji zostawiają jednak pole do poprawy wytrenowanych modeli. Dalsze ulepszenie modelu regresji pozwoli na uzyskanie narzędzia diagnostycznego pozwalającego na identyfikację nawet niewielkich zmian motorycznych pacjenta, co jest bardzo istotne z punktu widzenia oceny skuteczności stosowanych terapii.

Tabela 3. Regresja z wykorzystaniem danych z urządzenia Biodex System 4 Pro

Metoda	Metryka	MSE	MAPE	R2
SVR	GDI	105,26	14%	-26%
SVR	GGI	13204,57	70%	-14%
KNN	GDI	132,01	16%	-58%
KNN	GGI	19749,40	97%	-71%
DTR	GDI	87,28	11%	-5%
DTR	GGI	20507,14	80%	-77%
GBR	GDI	77,11	12%	7%
GBR	GGI	10136,02	58%	12%

Źródło: Opracowanie własne

4.2. Diagnostyka za pomocą bieżni Zebris FDM-T

Wykorzystując dane pochodzące z bieżni Zebris FDM-T opracowane zostało rozwiązanie mające na celu predykcję wartości GDI dla kończyn dolnych pacjenta: GDI left, GDI right oraz GDI avg. Użyto 50 próbek danych pochodzących z różnych

etapów badań. Uzyskane w ten sposób wartości zostały porównane z rzeczywistymi wartościami GDI wyliczonymi na podstawie danych z urządzenia BTS Bioengineering Smart DX.

Wykorzystując narzędzie Orange[15], utworzono modele przy użyciu wybranych metod predykcji: regresja liniowa (ang. *Linear Regression*), k-najbliższych sąsiadów (ang. *K-nearest Neighbors*), drzewo (ang. *Tree*), las losowy (ang. *Random Forest*), metody wektorów wspierających (ang. *Support Vector Machine*), AdaBoost, stochastyczny spadek wzdłuż gradientu (ang. *Stochastic Gradient Descent*). Początkowo pod uwagę brana była również sieć neuronowa, jednakże uzyskane za jej pomocą wyniki znacznie odbiegały od wyników otrzymanych za pomocą innych metod, co mogło być spowodowane zbyt małą liczbą danych uczących. Ostatecznie ta metoda nie została wzięta pod uwagę.

Następnie poszczególne metody predykcji zostały poddane ocenie, która polegała na wyliczeniu różnicy pomiędzy rzeczywistymi danymi a prognozowanymi wartościami. W ocenie wyników eksperymentów zostały zastosowane następujące wartości miar błędów: błąd średniokwadratowy MSE (ang. *Mean Squared Error*), średni bezwzględny błąd procentowy MAPE (ang. *Mean Absolute Percentage Error*) oraz współczynnik determinacji R^2 (ang. *Coefficient of determination R-squared*).

Przeprowadzone zostały trzy eksperymenty, w których wyniki badań na różne sposoby były przydzielane do dwóch grup: pierwszej użytej do stworzenia modelu oraz drugiej wykorzystanej do oceny modelu. W tabelach 4. 5., 6. znajdują się wyniki średnich wartości MSE, MAPE oraz R^2 wyliczone dla trzech eksperymentów. Rysunki 4., 5. i 6. przedstawiają wizualizację wyników przedstawionych w tych tabelach.

Jeżeli wziąć pod uwagę średnie wartości miar MSE oraz MAPE, to:

- we wszystkich przypadkach najmniejsze średnie wartości uzyskuje GDI avg;
- dla miary MSE największe średnie wartości ma GDI right;
- dla miary MAPE największe średnie wartości uzyskuje GDI right, poza dwoma przypadkami, gdy największe średnie wartości uzyskuje GDI left.

Jeżeli wziąć pod uwagę średnią wartość miary R^2, to:

- najmniejsze wartości uzyskuje GDI left;
- zależnie od wybranej metody największe wartości osiąga GDI right lub GDI avg.

Tabela 4. Średnie wartości MSE wyliczone dla trzech eksperymentów

Model	GDI avg	GDI left	GDI right
Linear Regression	63,06	83,78	99,71
kNN	45,78	80,07	94,86
Tree	95,28	130,22	191,18
Random Forest	56,84	79,69	94,73
SVM	50,41	65,76	89,24
AdaBoost	62,55	86,59	110,85
SGD	47,58	63,27	91,80

Źródło: Opracowanie własne

[15] https://orangedatamining.com/.

Tabela 5. Średnie wartości MAPE wyliczone dla trzech eksperymentów

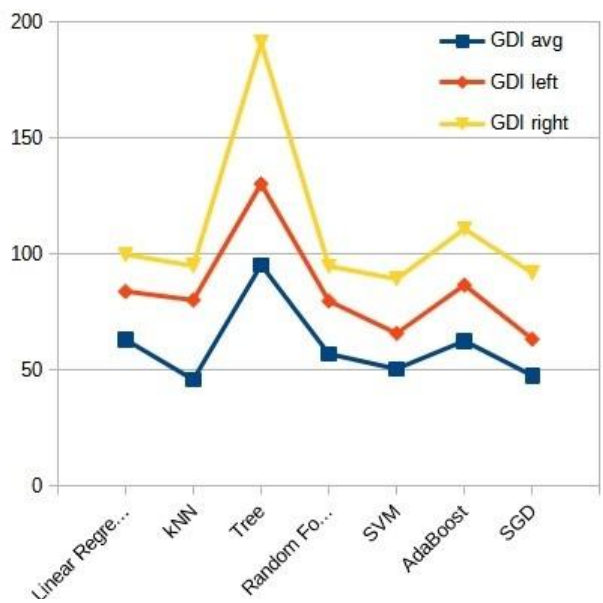
Model	GDI avg	GDI left	GDI right
Linear Regression	9,91%	11,44%	12,13%
kNN	7,92%	11,27%	10,97%
Tree	11,56%	13,51%	14,87%
Random Forest	8,52%	11,11%	10,67%
SVM	8,21%	10,01%	10,48%
AdaBoost	8,38%	10,71%	11,17%
SGD	8,60%	10,01%	11,45%

Źródło: Opracowanie własne

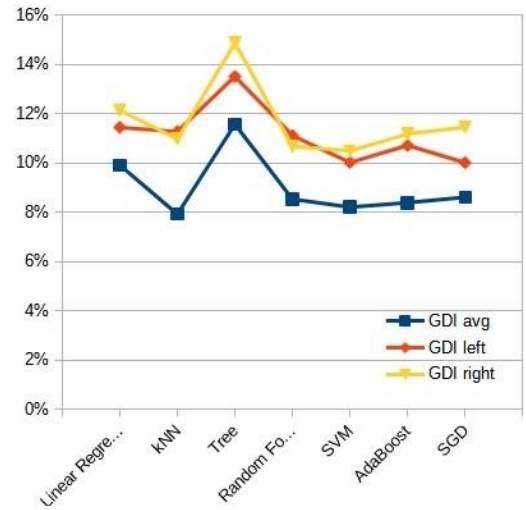
Tabela 6. Średnie wartości R^2 wyliczone dla trzech eksperymentów

Model	GDI avg	GDI left	GDI right
Linear Regression	0,09	-0,12	0,16
kNN	0,33	-0,19	0,21
Tree	-0,38	-0,73	-0,63
Random Forest	0,18	-0,16	0,20
SVM	0,28	0,16	0,26
AdaBoost	0,06	-0,29	0,07
SGD	0,30	0,14	0,23

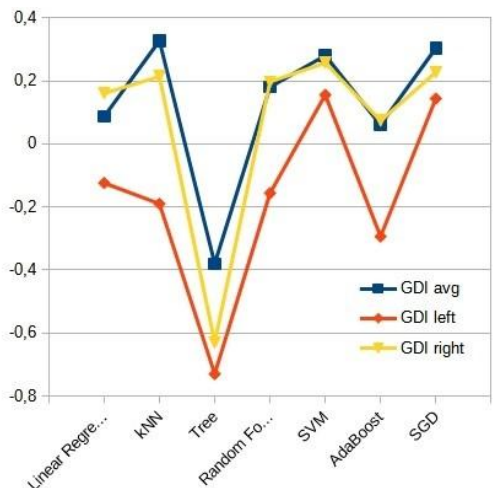
Źródło: Opracowanie własne



Rysunek 4. Wykres średnich wartości MSE wyliczone dla trzech eksperymentów [opracowanie własne]



Rysunek 5. Wykres średnich wartości MAPE wyliczone dla trzech eksperymentów [opracowanie własne]



Rysunek 6. Wykres średnich wartości R^2 wyliczone dla trzech eksperymentów [opracowanie własne]

Na podstawie powyższych danych można wywnioskować, że najtrafniej prognozowana jest wartość GDI avg. Jeżeli dla GDI avg w następujący sposób posortujemy średnie wartości miar uzyskiwane przez poszczególne metody:

- rosnąco: MSE, MAPE;
- malejąco: R^2;

to powstanie zestawienie widoczne w tabeli 7.

Tabela 7. Zestawienie średnich wartości miar

Numer pozycji	MSE (rosnąco)	R^2 (malejąco)	MAPE (rosnąco)
1	kNN	kNN	kNN
2	SGD	SGD	SVM
3	SVM	SVM	AdaBoost

4	Random Forest	Random Forest	Random Forest
5	AdaBoost	Linear Regression	SGD
6	Linear Regression	AdaBoost	Linear Regression
7	Tree	Tree	Tree

Źródło: Opracowanie własne

Można tu zaobserwować następujące prawidłowości:

- metoda kNN jest zawsze na pozycji 1;
- metoda Random Forest jest zawsze na pozycji 4;
- metoda Tree jest zawsze na pozycji 7;
- pozostałe metody są na różnych pozycjach.

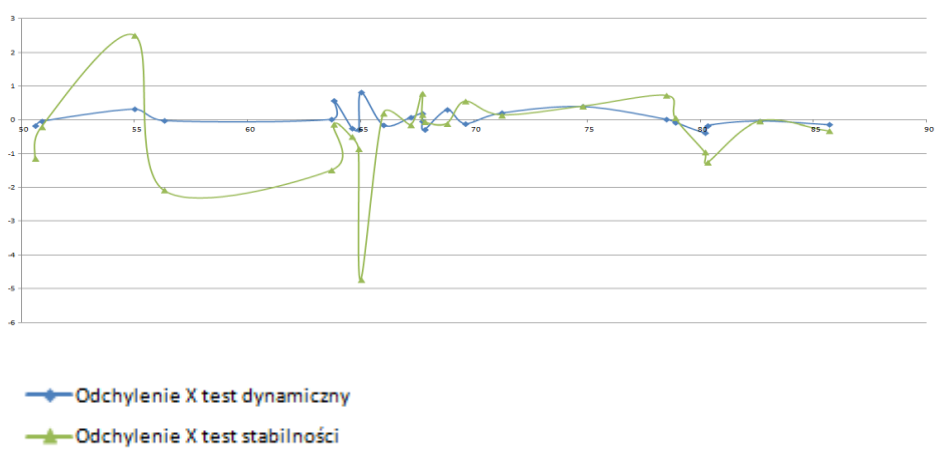
Biorąc pod uwagę powyższe obserwacje, można wywnioskować, że preferowaną metodą jest metoda kNN.

Podsumowując, na podstawie aktualnie dostępnych danych w celu predykcji średniej wartości Gait Deviation Index dla obu kończyn dolnych (GDI avg) proponuje się wykorzystanie metody k-najbliższych sąsiadów (kNN).

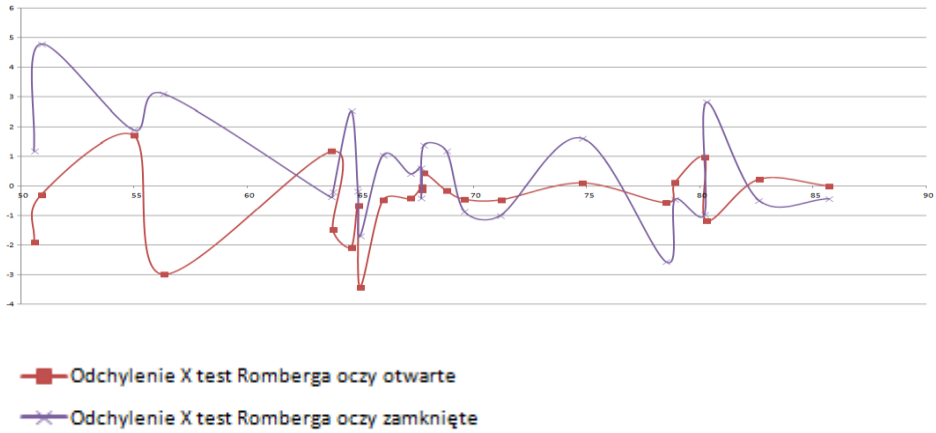
4.3. Diagnostyka za pomocą platform Alfa i Gamma

Platformy stabilometryczne i dynamometryczne dostarczają jedynie danych związanych z utrzymaniem równowagi pacjenta, w związku z tym nie pozwalają na ocenę chodu pacjenta i dlatego możliwość predykcji metryk GDI lub GGI za pomocą danych pochodzących z platform jest mało prawdopodobna. Ale odkrycie jakichkolwiek korelacji pomiędzy wartościami wybranych pomiarów wykonanych na platformach a tymi metrykami stanowiłoby interesującą informację diagnostyczną. Dlatego w ramach tej pracy wykonano badania mające na celu sprawdzenie czy istnieje zależność pomiędzy parametrami opisującymi wyniki badań zaburzeń stanu równowagi pacjenta a wartością metryki GDI.

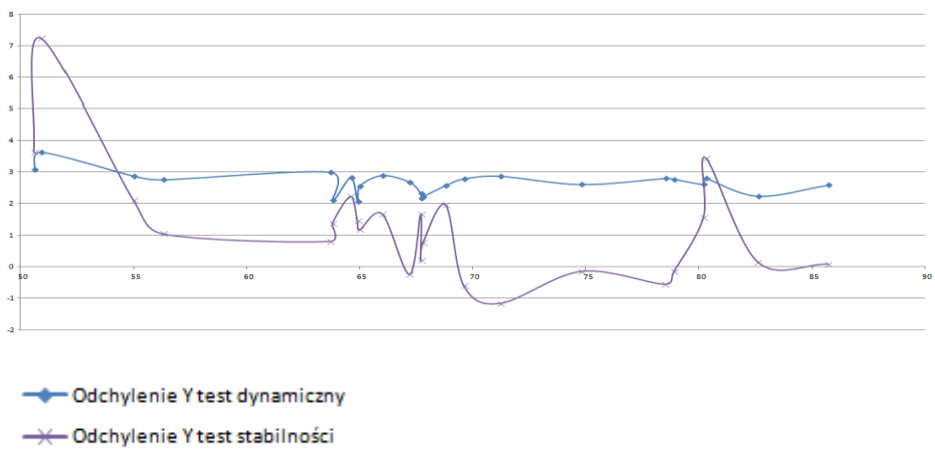
Rysunki 7, 8 i 9 przedstawiają przykładowe wizualizacje uzyskanych wyników dla kilkunastu pacjentów. Oś X na tych rysunkach odpowiada wartości metryki GDI, natomiast na oś Y określa zmierzone parametry.



Rysunek 7. Odchylenie X dla dynamicznego i testu stabilności [opracowanie własne]



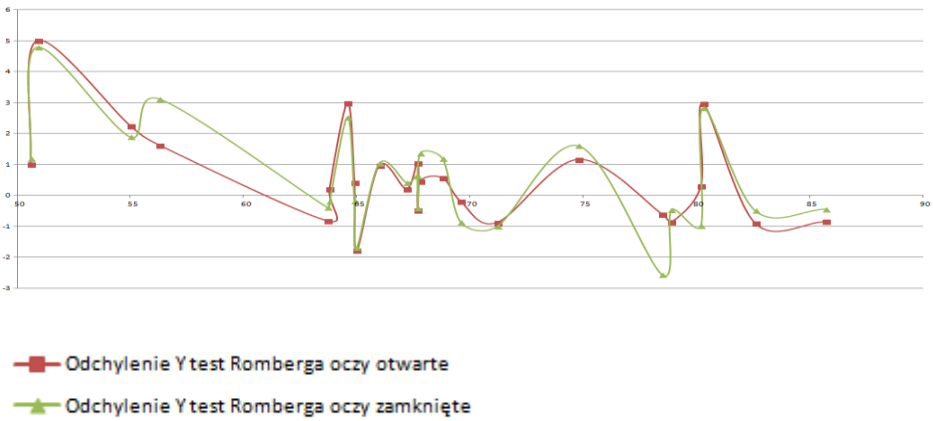
Rysunek 8. Odchylenie X dla testu Romberga [opracowanie własne]



Rysunek 9. Odchylenie Y dla dynamicznego i testu stabilności [opracowanie własne]

Jak można zauważyć nie istnieją proste zależności pomiędzy zmierzonymi parametrami a wartością metryki GDI. Podobne wyniki uzyskano dla innych pomiarów: długość ścieżki, średnią szybkość. Jednak w niektórych przypadkach można zauważyć pewne tendencje zmian, np. w przypadku testu Romberga można zauważyć mniejsze odchylenie Y dla większych wartości GDI.

Ze względu na duży rozrzut danych biometrycznych przy zbliżonych wartościach GDI można stwierdzić, że do uzyskania konstruktywnych wniosków konieczne jest przeprowadzenie złożonych analiz dla dużej grupy pacjentów. Wówczas, stosując metody uczenia maszynowego możliwe będzie znalezienie lub wykluczenie złożonych korelacji pomiędzy GDI a pomiarami biometrycznymi. Występująca w wielu przypadkach dość płaska linia trendu może też świadczyć o braku korelacji pomiędzy metryką GDI a danymi parametrami biometrycznymi mierzonymi za pomocą platform. W takim przypadku można poszukiwać zależności w odniesieniu do innych metryk, np. związanych z oceną stabilności np. Skala Równowagi Berga (BBS).



Rysunek 10. Odchylenie Y dla testu Romberga [opracowanie własne]

5. Wnioski

W pracy przedstawiono badania dotyczące możliwości diagnostyki chodu pacjentów z MPD za pomocą nowoczesnych urządzań do oceny biomechanicznej pacjentów. W tym celu zastosowano metody uczenia maszynowego obejmujące metody big data do stworzenia modeli predykcyjnych pozwalających na oszacowanie wartości metryk stosowanych w diagnostyce chodu. Wykazano, że dla urządzeń Zebris FDM-T oraz Biodex System 4 jest możliwa predykcja metryki GDI z dosyć wysoką dokładnością, pomimo że urządzenia te nie dokonują pomiaru parametrów wymaganych do bezpośredniego wyliczenia wartości tej metryki. Wyniki badań potwierdzają ogromny potencjał metod analityki danych w diagnostyce medycznej. Metody te pozwalają na wydobywanie wiedzy z danych, w tym informacji dotyczących oceny chodu pacjenta. Badania przedstawione w tej pracy pozwoliły na znalezienie korelacji pomiędzy danymi pochodzącymi z różnych urządzeń, dzięki czemu możliwe jest wsparcie diagnostyki chodu pacjentów z MPD bez konieczności korzystania z zaawansowanych pracowni analizy ruchu.

W dalszych badaniach automatyczna diagnostyka będzie stosowana do oceny skuteczności terapii. Wykorzystanie dokładnych pomiarów biometrycznych pozwoli na identyfikację nawet niewielkich zmian motorycznych pacjentów, dzięki temu będzie możliwe wykrywanie zmian niezauważalnych w tradycyjnych badaniach klinicznych. Pozwoli to na szybszą i dokładniejszą ocenę stosowanych terapii i lepsze dostosowanie postępowań terapeutycznych do stanu niepełnosprawności pacjenta.

Uwagi ogólne

Praca została zrealizowana w ramach projektu badawczego POIR.04.01.04-00-0035/19 dofinansowanego ze środków UE.

Literatura

1. Zhang J.Y., Oskoui M., Shevell M., *A population-based study of communication impairment in cerebral palsy*, Journal of child neurology, 30(3), 2015, s. 277-284.
2. Afzali M., Etemad K., Kazemi A., Rabiei R., *Cerebral palsy information system with an approach to information architecture: a systematic review*, BMJ health & care informatics, 26(1), 2019.
3. Zhang J., *Multivariate Analysis and Machine Learning in Cerebral Palsy Research*, Frontiers in Neurology, 8, 2017.
4. McMichael G., Bainbridge M.N., Haan E., Corbett M., Gardner A., Thompson S., Van Bon B.W.M., Van Eyk C.L., Broadbent J., Reynolds C., O'callaghan M.E., *Whole-exome sequencing points to considerable genetic heterogeneity of cerebral palsy*, Molecular psychiatry, 20(2), 2015, s. 176-182.
5. Coker-Bolt P., Downey R.J., Connolly J., Hoover R., Shelton D., Seo N.J., *Exploring the feasibility and use of accelerometers before, during, and after a camp-based CIMT program for children with cerebral palsy*, Journal of pediatric rehabilitation medicine, 10(1), 2017, s. 27-36.
6. Sartori M., Fernandez J.W., Modenese L., Carty C.P., Barber L.A., Oberhofer K., Zhang J., Handsfield G.G., Stott N.S., Besier T.F., Farina D., *Toward modeling locomotion using electromyography-informed 3D models: application to cerebral palsy*, Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 9(2), 2017, e1368.
7. Ku J.P., Hicks J.L., Hastie T., Leskovec J., Ré C., Delp S.L., *The mobilize center: an NIH big data to knowledge center to advance human movement research and improve mobility*, Journal of the American Medical Informatics Association, 22(6), 2015, s. 1120-1125.
8. Fan H., Li L., Wijlaars L., Gilbert R., *A machine-learning approach to identify cerebral palsy cases using primary care database*, International Journal of Population Data Science, 3(4), 2018.
9. Kurowski B.G., Greve K., Bailes A.F., Zahner J., Vargus-Adams J., Mcmahon M.A., Aronow B.J., Mitelpunkt A., *Electronic health record and patterns of care for children with cerebral palsy*, Developmental Medicine & Child Neurology, 2021.
10. Harris C.M., Wright S.M., *Malnutrition in hospitalized adults with cerebral palsy*, Journal of Parenteral and Enteral Nutrition, 2021.
11. Phinyomark A., Petri G., Ibáñez-Marcelo E., Osis S.T., Ferber R., *Analysis of big data in gait biomechanics: current trends and future directions*, Journal of medical and biological engineering, 38(2), 2018, s. 244-260.
12. Kumari P., Cooney N.J., Kim T.S., Minhas A.S., *Gait analysis in Spastic Hemiplegia and Diplegia cerebral palsy using a wearable activity tracking device-a data quality analysis for deep convolutional neural networks*, 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2018, s. 1-4.
13. Malt M.A., Aarli Å., Bogen B., Fevang J.M., *Correlation between the Gait Deviation Index and gross motor function (GMFCS level) in children with cerebral palsy*, Journal of children's orthopaedics, 10(3), 2016, s. 261-266.
14. Matsunaga N., Ito T., Noritake K., Sugiura H., Kamiya Y., Ito Y., Mizusawa J., Sugiura H., *Correlation between the Gait Deviation Index and skeletal muscle mass in children with spastic cerebral palsy*, Journal of physical therapy science, 30(9), 2018, s. 1176-1179.
15. Ito T., Noritake K., Sugiura H., Kamiya Y., Tomita H., Ito Y., Sugiura H., Ochi N., Yoshihashi Y., *Association between gait deviation index and physical function in children with bilateral spastic cerebral palsy: A cross-sectional study*, Journal of clinical medicine, 9(1), 2020, 28.

16. McMulkin M.L., MacWilliams B.A., *Application of the gillette gait index, gait deviation index and gait profile score to multiple clinical pediatric populations*, Gait & posture, 41(2), 2015, s. 608-612.
17. Romei M., Galli M., Motta F., Schwartz M., Crivellini M., *Use of the normalcy index for the evaluation of gait pathology*, Gait & posture, 19(1), 2004, s. 85-90.
18. Schwartz M.H., Rozumalski A., *The Gait Deviation Index: a new comprehensive index of gait pathology*, Gait & posture, 28(3), 2008, s. 351-357.
19. Molloy M., McDowell B.C., Kerr C., Cosgrove A.P., *Further evidence of validity of the Gait Deviation Index*, Gait & posture, 31(4), 2010, s. 479-482.
20. Nowakowska K., Jochymczyk-Woźniak K., *Ocena chodu dzieci z mózgowym porażeniem na podstawie wskaźnika GDI,* Aktualne Problemy Biomechaniki, 2014, s. 127-132.

Zastosowanie metodyki big data w diagnostyce pacjentów z mózgowym porażeniem dziecięcym

Streszczenie
W pracy przedstawiono zastosowanie metod analityki danych, zgodnej z koncepcją big data, wykorzystywanej do wspomagania diagnostyki stopnia niepełnosprawności chorych na mózgowe porażenie dziecięce. Na podstawie danych biometrycznych pobieranych z urządzeń diagnostycznych, metodami big data aproksymowane były wartości metryk stosowanych przez lekarzy-rehabilitantów. Pomiary biometryczne wykonywane były za pomocą bieżni, systemu do oceny dynamometrycznej i izokinetycznej oraz platform stabilometrycznych i dynamograficznych. Przebadano różne techniki analityki danych big data, takie jak: klasteryzacja, klasyfikacja, regresja, szukanie reguł asocjacyjnych, wyszukiwanie charakterystyk, w celu pozyskania wiedzy z danych biometrycznych. Punktem odniesienia do oceny stanu niepełnosprawności były wyniki badań klinicznych wspomagane pomiarami optycznymi. Wyniki te były wyrażone za pomocą metryk, takich jak GGI i GDI. Na podstawie uzyskanych pomiarów dla grupy kilkudziesięciu osób wytrenowano modele uczenia maszynowego, w taki sposób, aby możliwa była predykcja wartości danej metryki na podstawie wybranych danych biometrycznych pozyskanych z danego urządzenia. Uzyskane wyniki badań pokazały, że za pomocą wybranych urządzeń można z dużą dokładnością oceniać stopień niepełnosprawności pacjenta w odniesieniu do dokładnej diagnostyki wykonanej za pomocą zaawansowanych systemów optycznej analizy ruchu.
Słowa kluczowe: mózgowe porażenie dziecięce, diagnostyka medyczna, analityka big data, uczenie maszynowe, biometrics

Diagnosis of patients with cerebral palsy using the big data methodology

Abstract
The paper presents the application of data analysis methods, consistent with the concept of big data, used to support the diagnosis of the degree of disability of patients with cerebral palsy. Based on biometric data collected from diagnostic devices, big data methods were used to approximate the values of metrics used by physicians-rehabilitators. Biometric measurements were performed using a treadmill, a dynamometric and isokinetic evaluation system, and stabilometric and dynamographic platforms. Various techniques of big data analytics were tested, such as: clustering, classification, regression, searching for association rules, searching for characteristics, in order to obtain knowledge from biometric data. The results of clinical trials supported by optical measurements were the reference point for assessing the state of disability. These results were expressed using metrics such as GGI and GDI. Based on the obtained measurements, machine learning models were trained for a group of several dozen people in such a way that it was possible to predict the value of a given metric based on selected biometric data obtained from a given device. The obtained results showed that with the use of selected devices it is possible to accurately assess the degree of patient's disability in relation to accurate diagnostics performed with the use of advanced systems of optical motion analysis.
Keywords: cerebral palsy, medical diagnosis, big data analytics, machine learning, biometrics

Aneta Lubańska[1], Szymon Kiełczawa[2], Michał Mazur[3]

Badanie wpływu wybranych parametrów procesu rozpylania magnetronowego na morfologię powierzchni oraz właściwości optyczne i elektryczne cienkich warstw tlenków wolframu

1. Wstęp

Tlenki wolframu od lat poddawane są analizom w wielu ośrodkach badawczych. Jednym z przedstawicieli tlenków wolframu jest trójtlenek wolframu – WO_3. Jest to materiał półprzewodnikowy typu n, który występuje zarówno w formie krystalicznej, jak i amorficznej. Wykazuje on wiele unikatowych właściwości, takich jak dobrą stabilność chemiczną i mechaniczną, dobre właściwości elektryczne, a ponadto jest przyjazny środowisku [1]. Fantastyczne właściwości zarówno fizyczne, jak i chemiczne powodują, że WO_3 i pozostałe tlenki wolframu znajdują obecnie szereg zastosowań, między innymi jako katalizatory, warstwy sensorowe, w urządzeniach elektrochromowych lub w przyrządach do hydroodsiarczania paliw [2, 3].

Tlenki wolframu mogą być wytwarzane różnymi sposobami, m.in. metodą hydrotermalną, metodą zol-żel, chemicznym osadzaniem na mokro, impulsowym osadzaniem laserowym [4], a także metodą rozpylania magnetronowego. Metoda rozpylania magnetronowego należy do grupy metod fizycznych osadzania kompozytów chemicznych i jest obecnie jedną z najczęściej stosowanych metod wytwarzania struktur cienkowarstwowych bazujących na tlenkach metali [5]. Rozpylanie magnetronowe posiada przewagę nad innymi technikami wytwarzania cienkich warstw w postaci braku konieczności wygrzewania poprocesowego, ale także możliwości osadzania wielowarstw o dobrej adhezji do podłoża. Rozpylanie magnetronowe pozwala na taki dobór parametrów procesu, aby w sposób kontrolowany móc modyfikować parametry wytwarzanych warstw [6, 7]. Ponadto, w procesie rozpylania magnetronowego parametry procesowe zależą od siebie nawzajem i znacznie wpływają na właściwości materiału. Między innymi, ciśnienie podczas trwania procesu wpływa na ruchliwość nośników ładunku powstałych warstw, co ma swoje przełożenie na zastosowanie w urządzeniach elektrochromowych [8].

Jako przykład publikacji analizującej właściwości tlenków wolframu może posłużyć ta autorstwa Xu i innych [9]. Autorzy tej pracy badali tlenki wolframu wytworzone metodą rozpylania magnetronowego w atmosferze mieszaniny argonu i tlenu, do zastosowań jako warstwy sensorowe czujników gazów. Stosunek stężenia atomowej zawartości wolframu do tlenu zdefiniowano w zakresie od 0 do 100%, a więc zmie-

[1] aneta.lubanska@pwr.edu.pl, Katedra Mikroelektroniki i Nanotechnologii, Wydział Elektroniki Mikrosystemów i Fotoniki, Politechnika Wrocławska, https://pwr.edu.pl/.
[2] szymon.kielczawa@pwr.edu.pl, Katedra Mikroelektroniki i Nanotechnologii, Wydział Elektroniki Mikrosystemów i Fotoniki, Politechnika Wrocławska, https://pwr.edu.pl/.
[3] michal.mazur@pwr.edu.pl, Katedra Mikroelektroniki i Nanotechnologii, Wydział Elektroniki Mikrosystemów i Fotoniki, Politechnika Wrocławska, https://pwr.edu.pl/.

niono jej stechiometrię. Poniżej 60% powłoki cienkowarstwowe były metaliczne. Powyżej 60% nastąpiła zmiana charakteru z metalicznego na półprzewodnikowy lub dielektryczny (bliżej 100% wartości). Poskutkowało to także zmianą charakteru warstwy ze substechiometrycznej na w pełni utlenione WO_3. Zatem, zamiana parametrów procesu spowodowała wytworzenie tlenku wolframu o odmiennych właściwościach [9]. Inni naukowcy, Arfaoui i inni [10] analizowali optoelektroniczne właściwości tlenków wolframu, wykonanych parowaniem próżniowym. W pracy tej porównano modyfikację struktury oraz właściwości elektryczne i optyczne, jako następstwo wygrzewania warstw, po procesie nanoszenia. Pokazano, że wygrzewanie struktur cienkowarstwowych powoduje zmianę charakteru warstwy z amorficznej na krystaliczną. Dalszy wzrost temperatury wygrzewania powodował zmianę struktury krystalicznej. Ponadto obróbka poprocesowa powodowała poprawienie właściwości optycznych oraz elektrycznych (zmniejszenie rezystywności o trzy rzędy), co przekłada się na możliwość poprawienia wydajności urządzeń optoelektroniczych [10].

Celem niniejszej pracy było zbadanie wpływu dwóch parametrów procesu rozpylania magnetronowego, tj:

- odległości target-podłoże;
- oraz zmiany proporcji gazu roboczego (argon) i reaktywnego (tlen);

na właściwości powierzchni (morfologia, zwilżalność), elektryczne oraz optyczne cienkowarstwowych tlenków wolframu. Dlatego, wytworzono cienkie warstwy dla dwóch odległości target-podłoże (8 i 16 cm) oraz dwóch proporcji gazu roboczego i reaktywnego (Ar : O_2 1 : 1 i 9 : 1). Jeśli chodzi o dobór odległości umieszczenia podłoża od targetu, to są to wartości graniczne, dla których możliwe było przeprowadzenie procesu rozpylania magnetronowego na stanowisku, którym dysponowali autorzy. Zastosowano target o wymiarach 28,5 mm średnicy oraz 3 mm grubości, stąd odległość mniejsza niż 8 cm podczas wytwarzania warstw skutkowała pojawieniem się niejednorodności grubości cienkich warstw na podłożach, natomiast szybkość osadzania maleje wraz z kwadratem odległości od targetu, co postanowiono zanalizować w niniejszej pracy. Z kolei zastosowane proporcje gazu roboczego i reaktywnego pozwoliły uzyskać zarówno warstwy dotlenione o dużej przezroczystości (Ar : O_2 1 : 1), jak i niedotlenione o małej przezroczystości (Ar : O_2 9 : 1). Wpływ tych parametrów procesu rozpylania magnetronowego na właściwości powierzchni wytworzonych cienkich warstw tlenków wolframu został oceniony na podstawie pomiarów topografii oraz zwilżalności powierzchni. Z kolei właściwości optyczne oraz elektryczne analizowano na podstawie widm transmisji oraz rezystancji.

2. Metoda wytwarzania i metody pomiarowe tlenków wolframu

Tlenki wolframu naniesiono, używając do tego celu metody rozpylania magnetronowego. Każdy z procesów trwał 3600 s, a ciśnienie w komorze roboczej wynosiło 4,5·10-3 mbar. Do procesu użyto wolframowych targetów o dużej klasie czystości (99,95%), średnicy 30 mm oraz grubości 3 mm. Moc wyładowania wynosiła 400 W. Powłoki cienkowarstwowe naniesiono na podłoża krzemowe (Si) i z krzemionki amorficznej (SiO_2) o wymiarach 1 × 1 oraz 2 × 2 cm. Wytworzono 4 próbki, porównując ze sobą dwie odległości targetu od podłoża podczas procesu rozpylania magnetronowego, wynoszące 8 i 16 cm, a także dwie proporcje mieszaniny gazów Ar:O_2 w atmosferze podczas procesu nanoszenia wynoszące 1 : 1 i 9 : 1. Proporcje gazów do-

bierano na podstawie ustalenia przepływu masowego z wykorzystaniem dwukanałowego przepływomierza MKS Instruments PR4000B. Proporcja Ar : O_2 1 : 1 oznaczała zadany przepływ na kanale O_2 równy 20 sccm oraz na kanale Ar 20 sccm natomiast proporcja Ar : O_2 9 : 1 oznaczała odpowiednio 36 : 4 sccm. W tabeli 1. zestawiono proporcje Ar : O_2 oraz odległości umieszczenia podłoży od targetów dla poszczególnych tlenków wolframu.

Tabela 1. Zestawienie zmiennych parametrów procesu wytwarzania wykonanych cienkowarstwowych tlenków wolframu

Nazwa	Proporcja Ar : O_2	Odległość target-podłoże [cm]
P1	1 : 1	16
P2	1 : 1	8
P3	9 : 1	16
P4	9 : 1	8

Źródło: Opracowanie własne

Morfologię powierzchni wytworzonych warstw tlenków wolframu zbadano, używając profilometru optycznego typu CCI Lite firmy Taylor Hobson oraz mikroskopu optycznego Olympus BX51. Profilometr optyczny pozwolił na pomiar grubości naniesionych warstw, a także określenie chropowatości powierzchni. Za pomocą mikroskopu optycznego wykonano zdjęcia powierzchni naniesionych cienkich warstw.

Właściwości optyczne analizowano, bazując na pomiarach współczynnika transmisji światła. W tym celu wykorzystano spektrofotometr Ocean Optics QE65000 wyposażony w źródło światła Mikropack DH-BAL 2000, którego charakterystykę można znaleźć w notach katalogowych producenta [11]. Źródło światła złożone jest z dwóch lamp: deuterowej, zapewniającej odpowiednią intensywność w zakresie światła ultrafioletowego oraz halogenowej zapewniającej odpowiednią intensywność w zakresie światła widzialnego i bliskiej podczerwieni. Pomiarom poddano warstwy naniesione na podłoża z krzemionki amorficznej, której znaną cechą jest współczynnik załamania równy około 1,46 oraz współczynnik ekstynkcji światła bliski 0 w zakresie długości fal od około 200 nm. Pomiar współczynnika transmisji światła wykonano w zakresie od 250 nm do 1000 nm, w którym SiO_2 charakteryzuje się brakiem absorpcji światła. Podczas wykonywania pomiarów niezbędne było skorzystanie z dedykowanego oprogramowania Spectra Suite.

Do pomiaru właściwości elektrycznych użyto sondy czteroostrzowej Jandel 4-point Probe oraz źródła mierzącego Keithley 2600.

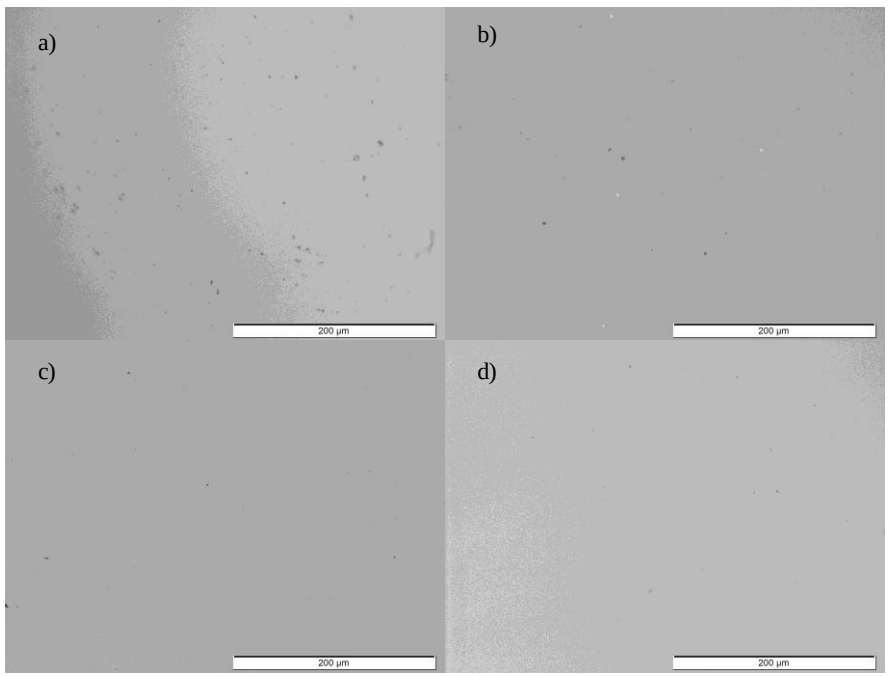
Określenie zwilżalności powierzchni wykonanych tlenków wykonano, używając do tego celu specjalnie dedykowanego stanowiska pozwalającego na pomiar kąta zwilżania – tensjometru optycznego firmy Attension, model Theta Lite. Ocenę zwilżalności przeprowadzono z użyciem trzech cieczy: wody dejonizowanej, alkoholu etylowego oraz glikolu etylenowego. Analizę kształtu kropli badanych cieczy przeprowadzono z wykorzystaniem oprogramowania komputerowego.

3. Wyniki i analiza pomiarów

Niniejszy rozdział zawiera wyniki wykonanych badań wraz z ich interpretacją. Rozdział podzielono na sekcje, w zależności od rodzaju mierzonych wielkości. Opisano morfologię powierzchni naniesionych powłok, analizę ich właściwości optycznych, omówienie właściwości elektrycznych oraz ocenę zwilżania powierzchni przez wybrane ciecze.

3.1. Morfologia powierzchni

Ocenę morfologii powierzchni zaczęto od wykonania zdjęć mikroskopem optycznym. Na rysunku 1. przedstawiono zdjęcia powierzchni warstw tlenków wolframu wykonane mikroskopem optycznym w powiększeniu ×20. Zaobserwowano, że każda wytworzona warstwa była jednolita i ciągła. Pojedyncze defekty widoczne na zdjęciach wynikają z niedoskonałości podłoży.

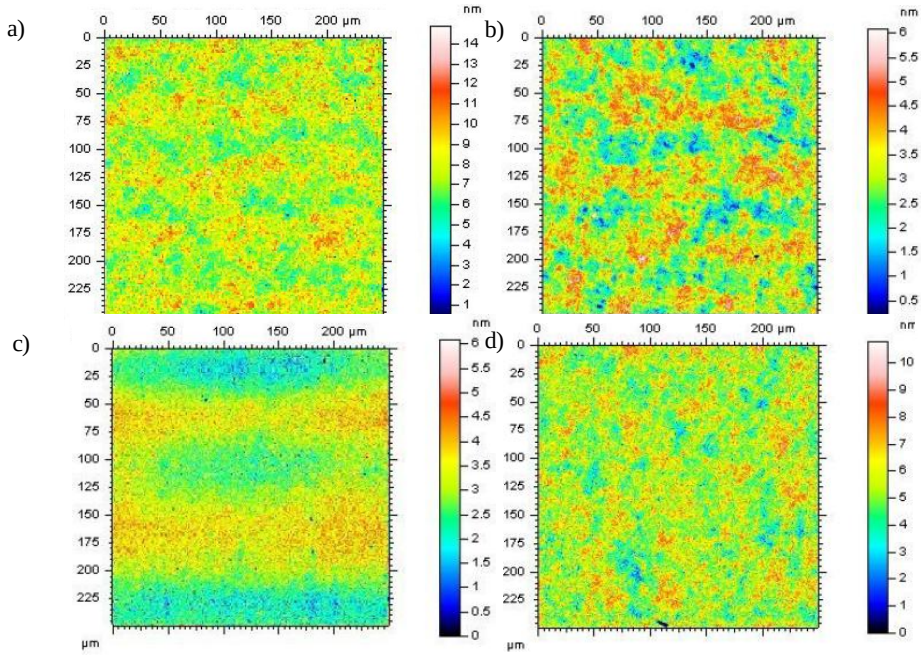


Rysunek 1. Wygląd powierzchni wykonanych struktur cienkowarstwowych tlenków wolframu – zdjęcia wykonane mikroskopem optycznym: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]

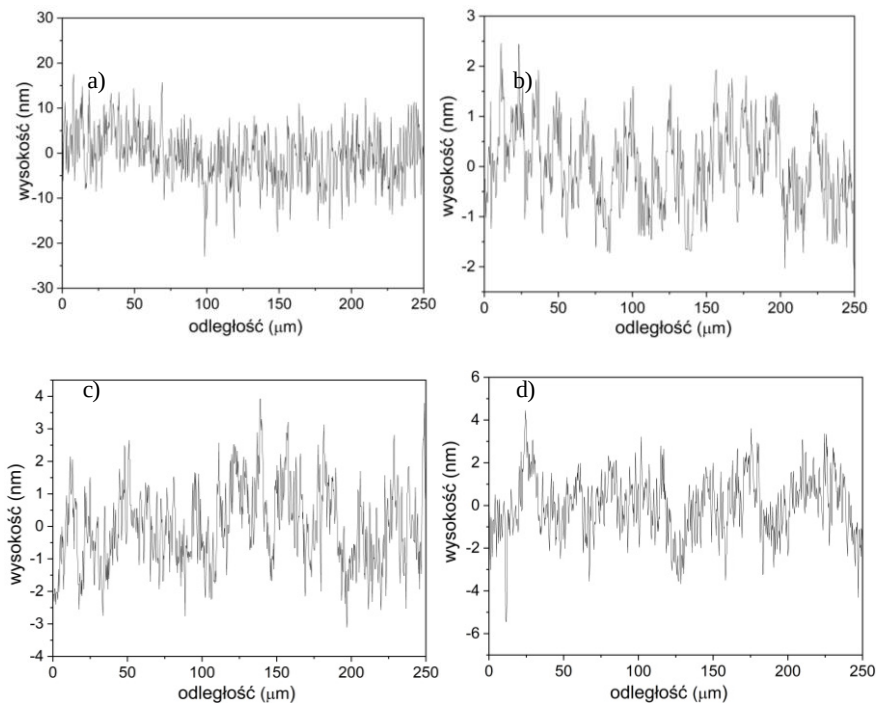
Pomiary właściwości powierzchni wykonano z wykorzystaniem profilometru optycznego, analizując obrazy 2D i 3D, a także wykreślając profile przekrojów powierzchni warstwy. Wykonano również pomiary grubości oraz chropowatości wytworzonych warstw.

Na rysunku 2. przedstawiono obrazy powierzchni wykonane profilometrem optycznym. Najbardziej jednolitą powierzchnią charakteryzują się próbki P1 i P3. Nie zaobserwowano defektów powierzchni.

Na rysunku 3. przedstawiono profile przekrojów powierzchni cienkich warstw wykreślone na podstawie danych zmierzonych profilometrem, w oprogramowaniu OriginPro. Pomiar wykonano na długości 250 µm warstwy. Można zauważyć, że profile są zbieżne ze zdjęciami 2D powierzchni (rys. 2).



Rysunek 2. Obrazy powierzchni wytworzonych struktur cienkowarstwowych tlenków wolframu – zdjęcia wykonane profilometrem optycznym: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]



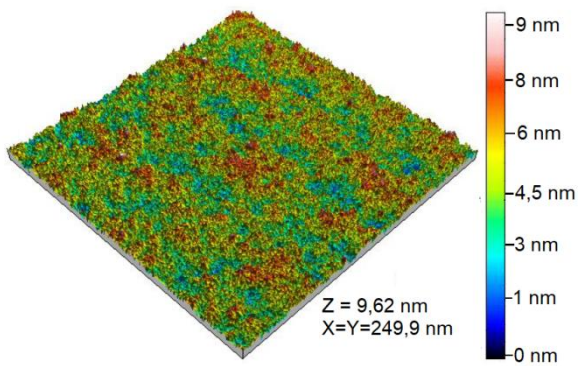
Rysunek 3. Profile przekroju przez cienkie warstwy tlenków wolframu na odległości 250 µm warstwy: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]

Obrazy trójwymiarowe powierzchni wykonanych cienkich warstw tlenków wolframu przedstawiono na rysunku 4. Ponownie, każda z warstw charakteryzuje się nieco innym rozłożeniem barw na obrazie. Różnice w wysokościach badanych fragmentów cienkich warstw mogą wynikać z niedoskonałości ich podłoży. Chropowatość powłok zmierzono z wykorzystaniem profilometru optycznego. Wartości zebrano w tabeli 2., gdzie:

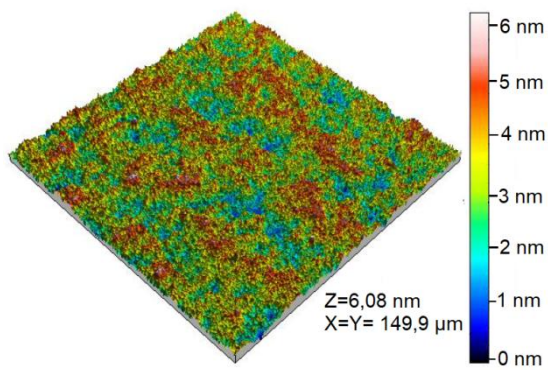
- R_a – odchylenie średnie arytmetyczne profilu chropowatości;
- R_q – odchylenie średnie kwadratowe profilu chropowatości;
- S_a – średnia arytmetyczna wysokość powierzchni;
- S_q – wysokość średniokwadratowa powierzchni.

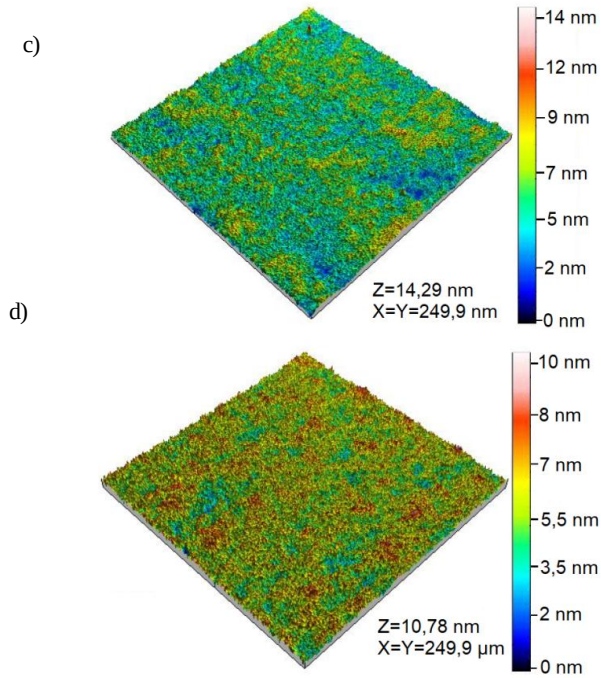
Odchylenie profilu chropowatości powierzchni wyniosło od 0,60 nm do 1,02 nm. Najmniejsze odchylenie profilu chropowatości powierzchni miała warstwa oznaczona jako P2, a największe próbka P4. Zwiększenie zawartości argonu w atmosferze gazowej podczas procesu rozpylania magnetronowego skutkowało wzrostem chropowatości wytworzonych warstw.

a)



b)





Rysunek 4. Powierzchnia wykonanych warstw tlenków wolframu – widok trójwymiarowy: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]

Tabela 2. Wartości średnich odchyleń profilu chropowatości oraz średnich wysokości powierzchni w całym jej przekroju wykonanych cienkowarstwowych tlenków wolframu

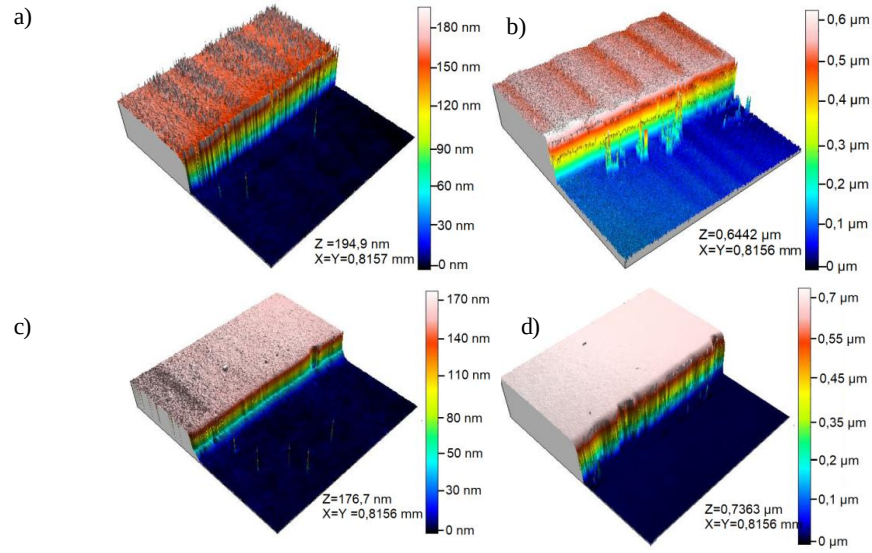
Nazwa	R_a [nm]	R_q [nm]	S_a [nm]	S_q [nm]
P1	0,89	1,08	1,10	1,39
P2	0,60	0,73	0,75	0,93
P3	0,86	1,05	1,03	1,10
P4	1,02	1,25	1,07	1,35

Źródło: Opracowanie własne

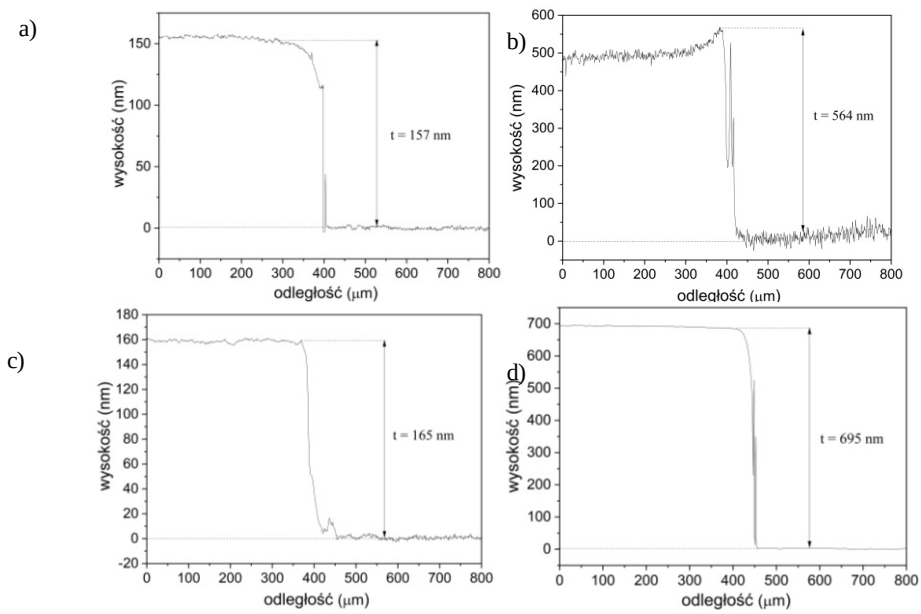
Pomiar grubości wytworzonych cienkich warstw tlenków wolframu wykonano na stopniu-uskoku, między podłożem a naniesioną warstwą. Obrazy trójwymiarowe zestawiono na rysunku 5. Ponadto na rysunku 6. zamieszczono wykreślone profile z zaznaczoną grubością poszczególnych warstw. Pokazano, że większe grubości uzyskano dla warstw wykonanych z większą zawartością argonu w atmosferze. Zmniejszenie odległości targetu od podłoża również skutkuje zwiększeniem grubości nanoszonych tlenków, przy zachowanym identycznym czasie trwania procesu. Na podstawie zmierzonych grubości oraz czasu trwania procesu rozpylania magnetronowego, obliczono szybkości nanoszenia powłok cienkowarstwowych. Wielkości te zestawiono w tabeli 3. Szybkość nanoszenia wzrasta wraz ze zmniejszeniem odległości target-podłoże oraz wzrostem zawartości argonu w atmosferze podczas procesu nanoszenia warstw, metodą rozpylania magnetronowego. Zgodnie z zależnością teoretyczną (2) dwukrotne zwiększenie odległości powinno skutkować zmniejszeniem grubości warstw czterokrotnie. Zgodnie z wynikami zamieszczonymi w tabeli 3., zmiany te wynoszą 3,59 w przypadku warstw wytworzonych w atmosferze 1 : 1 oraz 4,21 w atmosferze 9 : 1 $Ar : O_2$. Różnice te najpewniej związane są z niedokładnością wykonania pomiaru.

$$\frac{1}{d^2} \quad (2)$$

Gdzie:
d – odległość umieszczenia podłoży od targetów w trakcie trwania procesu rozpylania magnetronowego



Rysunek 5. Obrazy stopnia między warstwą tlenków wolframu a podłożem z krzemionki amorficznej: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]



Rysunek 6. Profile przekrojów powierzchni wytworzonych cienkowarstwowych struktur z zaznaczoną grubością: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]

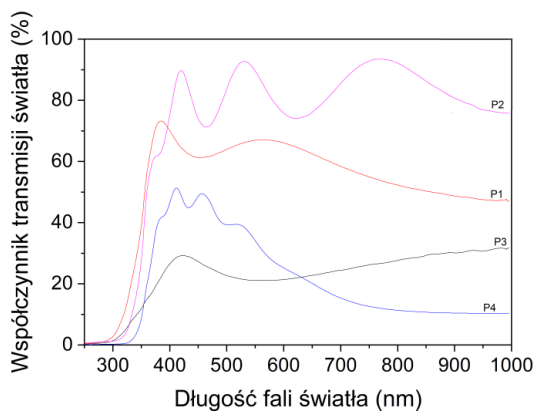
Tabela 3. Zestawienie wartości zmierzonych grubości oraz obliczonej szybkości nanoszenia dla poszczególnych warstw

Nazwa	Grubość [nm]	Szybkość nanoszenia [Å/s]
P1	157	0,44
P2	564	1,57
P3	165	0,46
P4	695	1,93

Źródło: Opracowanie własne

3.2. Właściwości optyczne

Analizę właściwości optycznych wykonano na podstawie charakterystyk transmisji światła. Na rysunku 8. porównano przebiegi charakterystyk współczynnika transmisji światła dla wszystkich wykonanych próbek. Przy warstwach wytworzonych dla mniejszych odległości target-podłoże w trakcie procesu nanoszenia tlenków zaobserwowano więcej interferencji, co przekłada się na ich większą grubość i szybkość nanoszenia w stosunku do tych wytwarzanych w większych odległościach. Średnie wartości współczynnika transmisji światła z zakresu widzialnego są większe dla warstw wykonanych w atmosferze $Ar : O_2$ wynoszącej 1 : 1. Na podstawie charakterystyk współczynnika transmisji wyliczono wartość fundamentalnej krawędzi absorpcji $\lambda_{cut\text{-}off}$. Mniejsze wartości uzyskano dla warstw wykonanych w większych odległościach podłoże-target. Wartości zestawiono w tabeli 4. Najlepszą przezroczystością charakteryzowała się warstwa P2, najmniejszą przezroczystość miała próbka P3.



Rysunek 7. Współczynnik transmisji światła przez cienkowarstwowe powłoki tlenków wolframu – zestawienie wszystkich wytworzonych próbek [opracowanie własne]

Tabela 4. Wartości fundamentalnej krawędzi absorpcji $\lambda_{cut\text{-}off}$ wykonanych cienkowarstwowych tlenków wolframu

Nazwa	$\lambda_{cut\text{-}off}$ [nm]
P1	317
P2	326
P3	304
P4	341

Źródło: Opracowanie własne

3.3. Właściwości elektryczne

Właściwości elektryczne określono przez pomiar rezystancji powierzchniowej, sondą czteroostrzową, oraz obliczenie rezystywności tlenków. Wartości pokazano w tabeli 5. Proporcja gazów w komorze w trakcie procesu rozpylania wynosząca 1 : 1 stosunku Ar : O_2 pozwoliła na uzyskanie warstw o charakterze dielektrycznym. Ich rezystancja powierzchniowa wynosiła powyżej $2 \cdot 10^{10}$ Ω/□ (zakres przyrządu pomiarowego). Rezystancja warstw wykonanych w atmosferze 9 : 1 była mniejsza i charakterystyczna dla materiałów półprzewodnikowych. Także mniejsze odległości targetu od podłoża pozwalają na uzyskanie mniejszych wartości rezystancji warstw.

Tabela 5. Wartości parametrów elektrycznych wszystkich wykonanych cienkowarstwowych tlenków wolframu

Nazwa	Rezystancja powierzchniowa [Ω/□]	Rezystywność [Ωcm]
P1	powyżej $2 \cdot 10^{10}$	powyżej 10^5
P2	powyżej $2 \cdot 10^{10}$	powyżej 10^6
P3	$1{,}3 \cdot 10^5$	2,15
P4	$4{,}1 \cdot 10^4$	2,85

Źródło: Opracowanie własne

3.4. Zwilżalność powierzchni

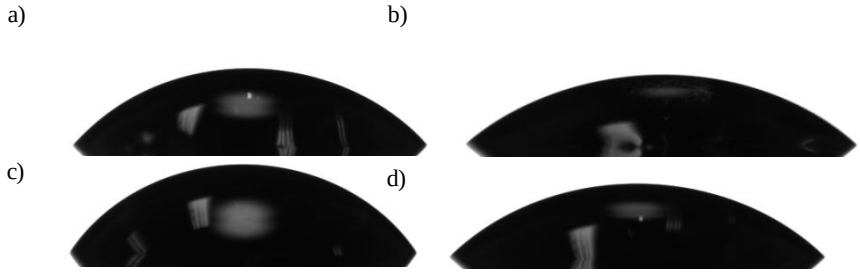
Analizę zwilżalności powierzchni wykonano przez pomiar kąta zwilżania z użyciem trzech cieczy o znanych napięciach powierzchniowych:

- wody dejonizowanej – 72,8 $\frac{mN}{m}$;
- glikolu etylenowego – 48,0 $\frac{mN}{m}$;
- alkoholu etylowego – 22,4 $\frac{mN}{m}$.

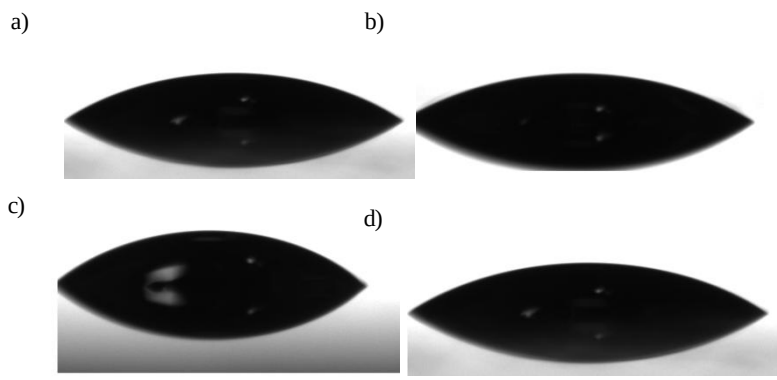
Krople cieczy umieszczano na powierzchniach każdej z warstw, używając w tym celu precyzyjnego dozownika. Na podstawie pomiaru kąta zwilżania oraz zależności Zismana [12], kolejnym krokiem było wyliczenie krytycznego napięcia powierzchniowego oraz swobodnej energii powierzchniowej. Wszystkie zmierzone oraz wyliczone wartości zebrano w tabeli 6-7. Fotografie kropli poszczególnych cieczy w zależności od warstwy zamieszczono na rysunkach 8-10. Krytyczne napięcie powierzchniowe wskazuje wartość napięcia powierzchniowego cieczy, dla którego badana warstwa zostanie całkowicie zwilżona. Im mniejsze napięcie powierzchniowe posiada dana ciecz, tym lepszą adhezję ta ciecz wykazuje w stosunku do powierzchni. Im większe napięcie powierzchniowe charakteryzuje ciecz, tym większy będzie kąt zwilżania obserwowany między nią a badaną powierzchnią. Analizując wygląd poszczególnych kropel na powierzchni wykonanych warstw, można zaobserwować, że woda dejonizowana posiadała najmniejszą adhezję do podłoża, natomiast najlepiej powierzchnie tlenków zwilżał alkohol etylowy, charakteryzujący się najmniejszą wartością napięcia powierzchniowego. Zależność kosinusa kąta zwilżania w funkcji napięcia powierzchniowego, to zależność Zismana [12]. Zależności te, dla wszystkich badanych próbek, pokazano na rysunku 11. Na ich podstawie wyliczono wartości swobodnej energii powierzchniowej, czyli wartość energii wynikającej z oddziaływań między cieczą a badaną powierzchnią. Posłużono się również wzorem [12]:

$$\gamma_s = \frac{(b \cdot \gamma_c + 1)^2}{4b} \quad (1)$$

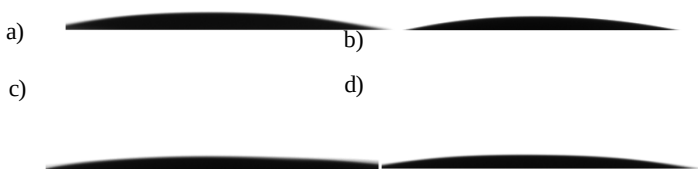
gdzie: γ_s – swobodna energia powierzchniowa, γ_c – krytyczne napięcie powierzchniowe, b – wartość bezwzględna współczynnika kierunkowego prostej aproksymującej wykres Zismana.



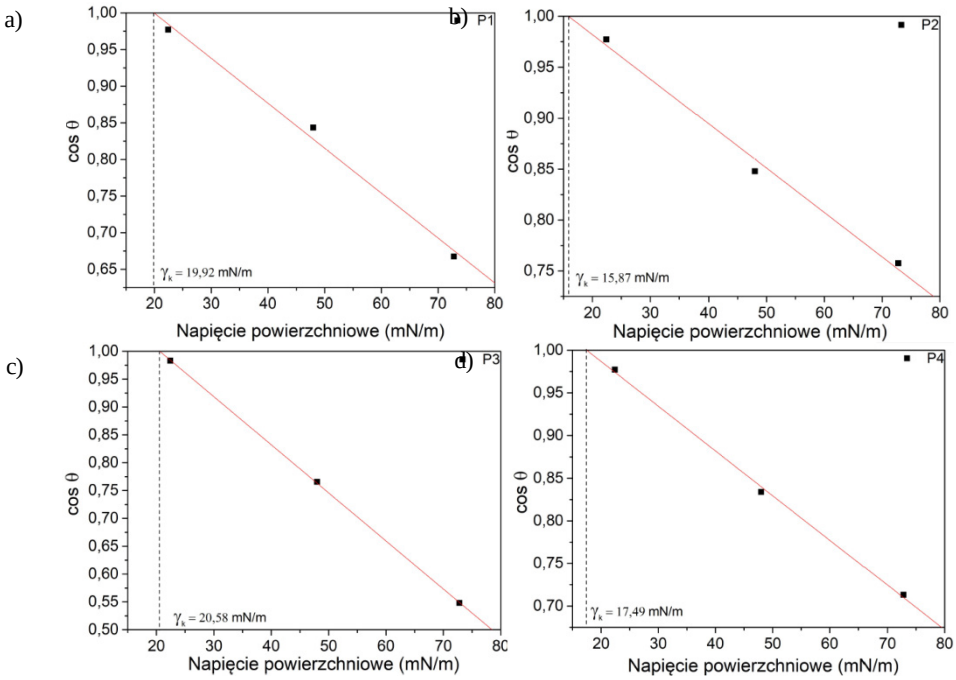
Rysunek 8. Wyniki pomiarów kąta zwilżania wody dejonizowanej na powierzchni cienkich warstw tlenków wolframu: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]



Rysunek 9. Wyniki pomiarów kąta zwilżania glikolu etylenowego na powierzchni cienkich warstw tlenków wolframu: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]



Rysunek 10. Wyniki pomiarów kąta zwilżania alkoholu etylowego na powierzchni cienkich warstw tlenków wolframu: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]



Rysunek 11. Wykresy Zismana – zależność zmian kosinusa kąta zwilżania θ w funkcji napięcia powierzchniowego cieczy dla poszczególnych powłok z tlenku wolframu: a) P1, b) P2, c) P3, d) P4 [opracowanie własne]

Tabela 6. Wartości kąta zwilżania dla wszystkich badanych cieczy wykonanych cienkowarstwowych tlenków wolframu

Nazwa	Kąt zwilżania θ [°]		
	Woda dejonizowana	Glikol etylenowy	Alkohol etylowy
P1	48,1 ±1,0	32,5 ± 1,3	12,3 ± 0,4
P2	40,7 ± 0,7	32,0 ± 2,1	12,2 ± 0,6
P3	56,8 ± 0,5	40,1 ± 0,9	10,5 ± 1,9
P4	44,5 ± 1,5	33,5 ± 2,4	12,2 ± 0,6

Źródło: Opracowanie własne

Tabela 7. Wartości krytycznego napięcia powierzchniowego oraz swobodnej energii powierzchniowej wszystkich wykonanych cienkowarstwowych tlenków wolframu

Nazwa	Krytyczne napięcie powierzchniowe $[\frac{mN}{m}]$	Swobodna energia powierzchniowa $[\frac{mN}{m}]$
P1	19,92	51,29
P2	15,87	65,55
P3	20,58	40,17
P4	17,49	56,86

Źródło: Opracowanie własne

Zaobserwowano, że dla mniejszych wartości krytycznego napięcia powierzchniowego warstwy posiadają większą swobodną energię powierzchniową. Podyktowane jest to faktem, że w warstwach posiadających mniejsze krytyczne napięcie powierzchniowe występuje więcej oddziaływań między cieczą a zwilżaną warstwą. Mniejsze

wartości krytycznego napięcia powierzchniowego posiadają te z próbek, który zostały wykonane w zarówno mniejszych odległościach, jak i mniejszej zawartości argonu podczas procesu rozpylania magnetronowego.

Na podstawie wyglądu kropel cieczy oraz wartości kąta zwilżania przez nich powierzchni można zauważyć, że kąt zwilżania alkoholu etylowego znajduje się na stałym poziomie dla każdego wykonanego tlenku wolframu. Kąt zwilżania glikolu etylenowego oraz wody dejonizowanej jest największy dla powłoki wytworzonej w atmosferze Ar : O_2 o proporcji 9 : 1 oraz odległości targetu od podłoża wynoszącej 16 cm. Zwiększenie zawartości argonu w mieszaninie gazów w atmosferze podczas rozpylania oraz umieszczenie podłoża dalej od targetu skutkuje pogorszeniem zwilżalności wykonanej warstwy.

Chropowatość powierzchni wykonanych tlenków wskazuje na brak znaczących różnic odchyleń przy zadanych parametrach chropowatości, natomiast wyniki pomiarów zwilżalności powierzchni warstw sugerują inne właściwości chemiczne każdej z nich. Warstwy posiadają różne właściwości adsorpcyjne zarówno grup hydroksylowych, jak i wodorotlenowych. W przyszłości przedstawione badania można zatem rozszerzyć i rozważyć zastosowania tak wykonanych tlenków wolframu jako warstwy gazoczułej do detekcji gazów w czujnikach chemicznych.

4. Podsumowanie

Przedmiotem niniejszej pracy była analiza właściwości optycznych, elektrycznych i właściwości powierzchni cienkowarstwowych powłok tlenków wolframu wykonanych za pomocą rozpylania magnetronowego. Na potrzeby badań naniesiono cztery warstwy, w oddzielnych procesach, różniących się między sobą atmosferą gazową w komorze (Ar : O_2 1 : 1 oraz 9 : 1) oraz odległością podłoża od targetu (8 cm i 16 cm). Mniejsze odległości targetu od podłoża oraz większe zawartości argonu w atmosferze podczas procesu pozwoliły na uzyskanie większych grubości oraz szybkości nanoszenia. Pomiar transmisji światła warstw wykazał więcej interferencji dla tych, wytworzonych w odległości 8 cm target-podłoże, co ma bezpośrednie przełożenie na ich większą grubość. Większą średnią wartość współczynnika transmisji uzyskano dla warstw wykonanych w proporcji gazów 1 : 1 podczas procesu. Tlenki wolframu wytworzone w tej atmosferze miały dielektryczny charakter, natomiast rezystancja tych wykonanych przy wyższej zawartości argonu klasyfikowała je jako półprzewodniki. Morfologia powierzchni wykazała jednolitość i ciągłość wszystkich wykonanych powłok cienkowarstwowych, a pomiar chropowatości nie wykazał różnic w ich chropowatościach. Ocena zwilżalności powierzchni wykazała jednak różnorakie właściwości adsorpcyjne każdej z warstw. Najmniejszy kąt zwilżania zaobserwowano dla alkoholu etylowego. Mniejsze odległości targetu od podłoża oraz stosunek gazów procesowych Ar : O_2 wynoszący 1 : 1 poprawiały zwilżalność warstw przez wodę dejonizowaną i glikol etylenowy.

Uzyskane wyniki są obiecujące. Pomiar zwilżalności sklasyfikował wykonane warstwy jako kandydatów do zastosowania w chemicznych czujnikach gazów, ze względu na właściwości adsorpcyjne grup hydroksylowych i wodorotlenowych. Badania te można zatem w przyszłości rozszerzyć.

Literatura

1. Babu M.B., Madhuri K.V., *Synthesis and electrochromic properties of nanocrystalline WO_3 thin films.* Physica B: Condensed Matter 584, 2020, 412068.
2. Filipescu M., Ossi P.M., Dinescu M., *WO_x cluster formation in radio frequency assisted pulsed laser deposition,* Applied Surface Science 254, 2007, s. 1347-1351.
3. Polášek J., Johánek V., Ostroverkh A., Mašek K., *Methanol oxidation on pure and platinum-doped tungsten oxide supported by activated carbon,* Materials Chemistry and Physics 228, 2019, s. 147-159.
4. Torabi Goodarzi M., Ranjbar M., *Atmospheric flame vapor deposition of WO_3 thin films for hydrogen detection with enhanced sensing characteristics,* Ceramics International 46, 2020, s. 21248-21255.
5. Mohamed S.H., Mohamed H.A., Abd El Ghani H.A., *Development of structural and optical properties of WO_x films upon increasing oxygen partial pressure during reactive sputtering,* Physica B: Condensed Matter 406, 2011, s. 831-835.
6. Yamamoto A., Abe Y., Kawamura M., Sasaki K., *Effects of oxygen gettering and target mode change in the formation process of reactively RF sputtered WO_x thin films,* Vacuum 66, 2002, s. 269-273.
7. Haviar S., Čapek J., Batková Š., Kumar N., Dvořák F., Duchoň T., Zeman P., *Hydrogen gas sensing properties of WO_3 sputter-deposited thin films enhanced by on-top deposited CuO nanoclusters,* International Journal of Hydrogen Energy, 43, 2018, s. 22756-22764.
8. Pan L., Han Q., Dong Z., Wan M., Zhu H., Li Y., Mai Y., *Reactively sputtered WO_3 thin films for the application in all thin film electrochromic devices,* Electrochimica Acta 328, 2019.
9. Xu X., Yazdi M.A.P., Rauch J.Y., Salut R., Billard, A., Potin V., Martin N., *Tungsten Oxide Thin Films Sputter Deposited by the Reactive Gas Pulsing Process for the Dodecane Detection,* Materials Today: Proceedings 2, 2015, s. 4656-4663.
10. Arfaoui A., Ouni B., Touihri S., Mannoubi T., *Investigation into the optoelectrical properties of tungsten oxide thin films annealed in an oxygen air,* Materials Research Bulletin 60, 2014, s. 719-729.
11. https://www.artisantg.com/info/Mikropack_DH2000_Datasheet.pdf.
12. Selvakumar N., Barshilia H.C., Rajam K.S., *Effect of substrate roughness on the apparent surface free energy of sputter deposited superhydrophobic polytetrafluoroethylene coatings: A comparison of experimental data with different theoretical models,* Journal of Applied Physics 108, 2010.

Badanie wpływu wybranych parametrów procesu rozpylania magnetronowego na morfologię powierzchni oraz właściwości optyczne i elektryczne cienkich warstw tlenków wolframu

Steszczenie

Cel pracy zdefiniowano jako analizę właściwości optycznych, elektrycznych i powierzchni cienkowarstwowych powłok na bazie tlenków metali wytworzonych w różnych warunkach procesowych, metodą rozpylania magnetronowego. Wykonano cztery warstwy przy jednakowym ciśnieniu i czasie trwania procesu rozpylania magnetronowego, ale przy zmiennej odległości targetu od podłoża (8 i 16 cm) oraz proporcji gazów Ar:O_2 wynoszących 1:1 oraz 9:1. Na potrzeby tej analizy wykonano szereg badań, w tym pomiary morfologii powierzchni, grubości, chropowatości oraz pomiar współczynnika transmisji światła i rezystancji powierzchniowej. Wyniki badań uzupełniono o pomiar zwilżalności powierzchni tlenków, wykorzystując w tym celu wodę dejonizowaną, glikol etylenowy oraz alkohol etylowy. Pokazano, że wszystkie naniesione warstwy były ciągłe. Zwiększenie zawartości argonu w mieszaninie gazów podczas procesu

skutkowało uzyskaniem większych grubości warstw, zmianę charakteru z dielektrycznego na półprzewodnikowy oraz pogorszeniem zwilżalności. Mniejsza odległość targetu od podłoża pozwoliła na zaobserwowanie większej ilości interferencji w charakterystykach współczynnika transmisji światła oraz polepszenie zwilżalności cienkich warstw. Tak wytworzone tlenki wolframu w przyszłości mogłyby znaleźć zastosowanie jako warstwy sensorowe.
Słowa kluczowe: tlenek wolframu, właściwości optyczne, właściwości elektryczne, morfologia powierzchni, właściwości adsorpcyjne

Influence of selected magnetron sputtering process parameters on the morphology, optical and electrical properties of tungsten oxide thin films

Abstract

The purpose of this study was the analysis of the optical, electrical, surface properties of thin films based on tungsten oxide prepared by magnetron sputtering, using various sputtering conditions. Four sets of films were prepared by magnetron sputtering using the same process pressure and deposition time but different target-substrate distance equal to 8 cm and 16 cm, and gas ratios 1:1 and 9:1. For the purpose of this paper, a numerous investigations were performed: surface morphology, measurements of surface thickness and roughness, examination of transmission coefficient, and sheet resistance. Wettability of the thin films was also studied with the use of deionized water, ethylene glycol, and ethanol. It was shown, that all prepared thin films were continuous. Increasing the fraction of argon during sputtering resulted in higher thickness, change of electrical character from dielectric to semiconducting, and deterioration of wettability. Shorter distance between substrate and target caused more interferences in transmittance spectra and better wettability. Such prepared thin films will probably be applicable as sensor films.
Keywords: tungsten oxide, optical properties, electrical properties, surface morphology, adsorption properties

Indeks Autorów

Andruloniw P. 132
Bedla M. 175
Borkowski M. 175
Ciepielewski R. 86
Deniziak S. 175
Kaczmarski D. 175
Kępys A. 162
Kiełczawa Sz. 191
Krechowicz A. 175
Król P. 57
Lubańska A. 191
Łukawski G. 175
Marszałek K. 7
Mazur M. 191
Miedzińska D. 86
Mrzygłód B. 105
Opaliński A. 105
Partyka B. 29
Pawełkiewicz M. 175
Pięta P. 175
Piwowarczyk M. 43
Płonka J. 142
Regulski K. 105
Rutkiewicz K. 142
Słowiński G. 117
Syska E. 162
Tutaj A. 57
Ungeheuer K. 7
Zwierzykowski P. 29, 132