RDIT wpł. 03.04.2024
M. Jkes

Dr hab. inż. Joanna Domańska
Instytut Informatyki
Teoretycznej i Stosowanej PAN
ul. Bałtycka 5, 44-100 Gliwice

Gliwice, 29 marca 2024 r.

Recenzja w postępowaniu habilitacyjnym
dr inż. Małgorzaty Bach

Podstawa formalna recenzji

Recenzja została wykonana na podstawie uchwały nr 140/2023 Rady Dyscypliny "Informatyka Techniczna i Telekomunikacja" Politechniki Śląskiej z dnia 19 grudnia 2023 w sprawie powołania komisji habilitacyjnej w postępowaniu o nadanie stopnia doktora habilitowanego dr inż. Małgorzacie Bach w dziedzinie nauk inżynieryjno-technicznych, w dyscyplinie informatyka techniczna i telekomunikacja, na mocy której zostałam powołana na recenzenta w składzie wyżej wymienionej komisji, jako wynik wskazania mojej osoby przez Radę Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej.

Podstawą prawną oceny osiągnięć naukowych Kandydatki ubiegającej się o stopień doktora habilitowanego jest art. 221 ust. 8 Ustawy z dnia 20 lipca 2018 r. - Prawo o szkolnictwie wyższym i nauce (t.j.: Dz.U. z 2023 poz. 742, z późn.zm.), a w zakresie kryteriów branych pod uwagę przy tej ocenie - art. 219 ust.1 pkt 2 wspomnianej ustawy. Dokumentację i materiały dotyczące przedmiotowego postępowania habilitacyjnego otrzymałam 5 lutego 2024 r.

Źródłem danych do wykonania recencji była dokumentacja sporządzona przez dr inż. Małgorzatę Bach, przekazana w formie elektronicznej.

Podstawowe dane o kandydatce

Pani dr inż. Małgorzata Bach w 1988 roku ukończyła studia magisterskie na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej w Gliwicach i obroniła pracę magisterską dotyczącą "Opracowania pakietu operacji graficznych dla mikroprocesora ComPAN". Promotorem pracy był prof. dr hab. inż. Stanisław Kozielski.

Pani dr inż. Małgorzata Bach uzyskała stopień doktora nauk technicznych uchwałą Rady Wydziału Automatyki, Elektroniki i Informatyki Politechniki Śląskiej w Gliwicach z dnia 1 czerwca 2004 r. Praca doktorska pod tytułem "Metody konstruowania zadań wyszukiwania w bazach danych w procesie translacji zapytań formułowanych w języku naturalnym" została napisana pod kierunkiem prof. dr hab. inż. Stanisława Kozielskiego.

W ramach kariery zawodowej Habilitantka była zatrudniona w Katedrze Informatyki Stosowanej na Wydziale Automatyki, Elektroniki i Informatyki Politechniki

Domańska

Śląskiej na początku jako asystent stażysta w latach 1988-1989, a później kolejno jako asystent (w latach 1989-1999) oraz wykładowca (w latach 1999-2004). Od roku 2004 jest zatrudniona na stanowisku adiunkta w wyżej wymienionym miejscu.

Informacje o ocenianych osiągnięciach naukowych osoby ubiegającej się o stopień doktora habilitowanego

Przedmiotem oceny jest osiągnięcie naukowe stanowiące podstawę ubiegania się o nadanie stopnia doktora habilitowanego pt.: "Metody wstępnego przetwarzania danych w kontekście odkrywania wiedzy klasyfikacyjnej, ze szczególnym uwzględnieniem problemu niezrównoważenia klas i wielowymiarowości." Na osiągnięcie składa się cykl ośmiu powiązanych tematycznie artykułów naukowych opublikowanych w czasopismach naukowych oraz w recenzowanych materiałach konferencyjnych. Są to prace:

[MB1] Bach M., Werner A., Żywiec J., Pluskiewicz W.: *The study of under- and over-sampling methods utility in analysis of highly imbalanced data on osteoporosis*, Information Sciences, Elsevier Inc., vol. 384, 2017, s. 174-190, DOI:10.1016/j.ins.2016.09.038, IF(4.305) 45 pkt. MNiSW

[MB2] Bach M., Werner A., Palt M.: The proposal of undersampling method for learning from imbalanced datasets. Procedia Computer Science, Elsevier BV, vol. 159, 2019, s. 125-134, DOI:10.1016/j.procs.2019.09.167, CORE Conference Ranking (B), 70 pkt. MNiSW

[MB3] Bach M., Werner A.: Improvement of random undersampling to avoid excessive removal of points from a given area of the majority class, W: Computational Science – ICCS 2021: 21st International conference, Krakow, Poland, June 16-18, 2021, Proceedings: Paszynski M.[i in.](red.), Lecture Notes In Computer Science, 2021, Springer, ISBN 978-3-030-77966-5, s. 172-186, DOI:10.1007/978-3-030-77967-2-15, CORE Conference Ranking (A), 140 pkt. MNiSW

[MB4] Bach M.: New undersampling method based on the KNN approach, Procedia Computer Science, Elsevier BV, vol. 207, 2022, s. 3403-3412, DOI: 10.1016/ j.procs.2022.09.399, CORE Conference Ranking (B), 70 pkt. MNiSW

[MB5] Werner A., Bach M., Pluskiewicz W.: The study of preprocessing methods' utility in analysis of multidimensional and highly imbalanced medical data, W: Proceedings of the 11th Scientific Conference Internet in the Information Society 2016, ISBN 978-83-65621-00-9, s. 71-87, 15 pkt. MNiSW

[MB6] Bach M., Werner A.: Cost-sensitive feature selection for class imbalance problem, W: Information systems architecture and technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology - ISAT 2017. Proceedings paper / Świątek J., Wilimowska Z. (red.), Advances in Intelligent Systems and Computing, vol. 655, 2018, Springer, ISBN 978-3-319-67220-5, s. 182-194, DOI:10.1007/978-3-319-67220-5-17, 20 pkt. MNiSW

[MB7] Adamczyk P., Werner A., Bach M., Żywiec J., Czekajło A., Grzeszczak W.2, Drozdowska B., Pluskiewicz W.: Risk factors for fractures identified in the algorithm

developed in 5-year follow-up of postmenopausal women from RAC-OST-POL study, Journal of Clinical Densitometry, vol. 21, nr 2, 2018, s. 213-219, DOI: 10.1016/j.jocd.2017.07.005, IF(2,184) 20 pkt. MNiSW

[MB8] Pluskiewicz W., Adamczyk P., Werner A., Bach M., Drozdzowska B.: POLRISK: an algorithm for 10-year fracture risk prediction in the postmenopausal women from the RAC-OST-POL Study, Polskie Archiwum Medycyny Wewnętrznej, MEDYCYNA PRAKTYCZNA SP K SP ZOO, vol. 133, nr 3, 2023, s. 1-9, DOI:10.20452/pamw.16395, IF(4,8) 200 pkt. MNiSW

Osiągnięcie naukowe, które jest przedmiotem oceny, obejmuje 8 artykułów opublikowanych w 3 czasopismach naukowych i 5 recenzowanych materiałach z konferencji.

Artykuł [MB1] w roku publikacji (2017) miał tylko 45 pkt wg wykazu ministerialnego, natomiast warto podkreślić, że wg aktualnego wykazu czasopismo Information Sciences zostało ocenione na 200 pkt; a jego impact factor wzrósł z 4,305 (2017) do 8,1 (2023). Artykuł [MB8] został opublikowany w roku 2023 w czasopismie, którego wartość w wykazie ministerialnym została oceniana na 200 pkt, natomiast artykuł [MB7] - oceniony w roku 2018 na 20 pkt - w najnowszym wykazie ma 100 pkt, a jego impact factor wzrósł z 2,184 (2018) do 2,5 (2023). Podsumowując - czasopisma, w ramach których kandydatka publikowała swoje prace naukowe oceniam dobrze.

W kwestii recenzowanych materiałów konferencyjnych, w ramach których publikowane były artykuły będące częścią ocenianego osiągnięcia naukowego, najwyżej oceniam artykuł [MB3], przedstawiony na konferencji "International Conference on Computational Science (ICCS)". Konferencja ta jest oceniania w wykazie ministerialnym na 140 pkt i wprawdzie w 2023 spadła do rangi "Multiconference" na australijskim wykazie Core (wg którego ustalana była pierwotna punktacja ministerialna), ale w roku publikacji, czyli 2021, miała rangę "Core: A". Konferencja "International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES)", w ramach której kandydatka przedstawiała swoje prace [MB2] i [MB4], jest oceniana w wykazie ministerialnym na 70 pkt, a na liście Core od 2008 roku niezmiennie zachowuje rangę "Core: B". Konferencja "International Conference on Information Systems Architecture and Technology", w ramach której został wygłoszony i opublikowany artykuł [MB6], od roku 2021 na liście Core jest rozpoznawana jako "National Poland", natomiast konferencja "Scientific Conference Internet in the Information Society 2016", w ramach której zaprezentowano artykuł [MB5], nie widnieje na liście Core. Podsumowując - recenzowane materiały konferencyjne, w ramach których kandydatka publikowała swoje prace naukowe oceniam jako akceptowalne w niniejszym postępowaniu o nadanie stopnia doktora habilitowanego.

Wszystkie prace wchodzące w skład ocenianego osiągnięcia naukowego są ściśle związane z tematyką osiągnięcia, pomimo tego, że prace [MB7] i [MB8] zostały opublikowane w czasopismach przypisanych do innych dyscyplin aniżeli dyscyplina w której toczy się postępowanie o nadanie stopnia doktora habilitowanego. Artykuł [MB7] jest przypisany do następujących dyscyplin: inżynieria biomedyczna; biologia medyczna; nauki medyczne; biotechnologia, natomiast artykuł [MB8] jest przypisany do następujących dyscyplin: inżynieria biomedyczna; biologia medyczna; nauki farmaceutyczne; nauki medyczne; nauki o zdrowiu; nauki o rodzinie; biotechnologia. Wkład habilitantki w powstanie prac [MB7] i [MB8] wg mojej oceny można przypisać do dyscypliny informatyka techniczna i telekomunikacja. We wszystkich współautorskich pracach składających się na oceniane osiągnięcie naukowe wg mojej opinii kandydatka odgrywała wiodącą rolę w ramach ich powstawania.

Dane naukometryczne Habilitantki:

Na dzień wszczęcia podstępowania hanilitacyjnego kandydatka do stopnia doktora habilitowanego legitymuje się 45 opublikowanymi pracami, dla których sumaryczny współczynnik Impact Factor wynosi 32,196 Sumaryczna punktacja ministerialna wynosi 1282 pkt. Indeks Hirscha wynosi 5 wg baz Web of Science (WoS) i Scopus oraz 8 wg Google Scholar. Wszystkie publikacje indeksowane w bazie Scopus i Web of Science powstały po uzyskaniu stopnia doktora. Sumaryczna liczba cytowań wynosi 144 wg bazy Web of Science (WoS), 182 wg bazy Scopus oraz 346 wg Google Scholar.

Według mojej oceny dane naukometryczne habilitantki są akceptowalne. Godny podkreślenia jest niski (poniżej 10%) poziom autocytowań.

Informacja o spełnieniu przez kandydatkę kryterium dotyczącego wykazywania się istotną aktywnością naukową

Aktywność naukową kandydatki oceniam jako dobrą.

Jedną z form tej aktywności był udział w projektach. W jednym z nich (H2020-Working in a Collaborative Factory of the Flight Simulators Branch of RISE) habilitantka była członkiem komitetu naukowego i liderem jednego z 7 kluczowych zadań, a udział w tym projekcie wiązał się sumarycznie z 11-miesięcznym stażem w firmie LG Nexera we Wiedniu.

Na uwagę zasługuje również współpraca z pracownikami Śląskiego Uniwersytetu Medycznego (SUM) w Katowicach, której efektem - oprócz motywacji do zajęcia się tematyką opisaną w ramach ocenianego osiągnięcia i współautorstwem sześciu publikacji dotyczących zaburzeń metabolicznych szkieletu - było opracowanie kalkulatora umożliwiającego ocenę ryzyka złamań osteoporotycznych. Udział w opracowaniu tego kalkulatora stał się przyczynkiem do nominacji Kapituły Redakcji „Dziennika Zachodniego" do tytułu Osobowość Roku 2019 w kategorii Nauka.

W ramach współpracy z uniwersytetem w Winnipeg w Kanadzie (The University of Winnipeg) habilitantka została zaproszona do wygłoszenia wykładu dla studentów Wydziału Informatyki tamtejszej uczelni.

Habilitantka jest członkiem komitetu programowego „International Conference on Knowledge-Based and Intelligent Information and Engineering Systems" (KES2023) oraz współorganizatorką sesji zaproszonej w ramach tej konferencji.

Kandydatka była również recenzentką artykułów publikowanych w ramach krajowych jak i międzynarodowych konferencji naukowych, takich jak: Beyond Databases, Architectures and Structures (BDAS), International Conference on Man-Machine Interactions (ICMMI), International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), Internet in the information society. Przygotowywała również recenzje dla czasopism z listy JCR: International Journal of Information Technology and Decision Making, Knowledge-Based Systems, Information Sciences.

Informacja o osiągnięciach dydaktycznych, organizacyjnych i popularyzujących naukę kandydatki do stopnia doktora habilitowanego

Osiągnięcia kandydatki: dydaktyczne, organizacyjne i popularyzujące naukę oceniam wysoko. Prowadziła zajęcia (wykłady, ćwiczenia, laboratoria i projekty), na studiach stacjonarnych oraz niestacjonarnych, pierwszego i drugiego stopnia, w pięciu różnych uczelniach wyższych:

- Politechnika Śląska (na ośmiu kierunkach - w tym 3 podyplomowych),
- Wyższa Szkoła Biznesu w Dąbrowie Górniczej,
- Gliwicka Wyższa Szkoła Przedsiębiorczości,

- Akademia Humanitas w Sosnowcu,
- Akademia Śląska w Katowicach.

Niektóre zajęcia prowadzone były w ramach projektów dydaktycznych (kandydatka wymienia osiem takich projektów).

Do godnych uwagi zadań dydaktycznych, organizacyjnych i popularyzujących naukę zaliczam również udział w Rybnickich Targach Edukacji – prowadzenie warsztatów dla młodzieży z zakresu projektowania systemów informatycznych i metodyki SCRUM, prowadzenie warsztatów dotyczących języka SQL, dla koła informatycznego CMI Akademickiego Liceum Ogólnokształcącego ALO w Gliwicach oraz aktywny udział w cyklicznych seminariach ADAA (Advanced Data Analysis and its Applications).

Ocena osiągnięcia naukowego osoby ubiegającej się o stopień doktora habilitowanego

Tematyka ocenianego osiągnięcia dotyczy metod odpowiedniego przygotowania danych przed ekstrakcją wiedzy klasyfikacyjnej. Są do zagadnienia w dalszym ciągu bardzo istotne w świetle wzrastającej popularności wykorzystywania metod sztucznej inteligencji w wielu dziedzinach życia, w tym również w zagadnieniach związanych z szeroko rozumianą medycyną. Wprawdzie rozpoczął się już proces wykorzystywania uczenia głębokiego w rozwiązywaniu szeregu istotnych zagadnień, które mogą zostać zrewolucjonizowane dzięki tej technologii, w tym do diagnozowania chorób, ale ze względu na wprowadzane w życie europejskie regulacje dotyczące wykorzystywania sztucznej inteligencji (tzw. AI Act), interpretowalność modeli wykorzystywanych w takich zastosowaniach jak medycyna staje się bardzo ważna. Dlatego też ważne są prace mające na celu ulepszanie tzw. modeli uczenia płytkiego - co uzasadnia ważność tematyki ocenianego osiągnięcia naukowego.

Jakość danych oraz ich specyficzne charakterystyki mogą znacząco wpływać na skuteczność i efektywność procesu analizy. Dane mogą być niekompletne, zaszumione, a ich struktura może być skomplikowana, co dodatkowo komplikuje ich przetwarzanie i analizę. W związku z tym, opracowanie skutecznych metod czyszczenia, normalizacji i transformacji danych jest kluczowe dla osiągnięcia wiarygodnych wyników analizy danych i wydobycia z nich cennej wiedzy. Klasyczne modele uczenia maszynowego (czyli tzw. modele uczenia płytkiego) są wyjątkowo użyteczne w procesie tzw. wydobywania wiedzy z danych, jednak ich skuteczność zależy od wielu czynników, przede wszystkim od jakości i natury danych, na których te modele są trenowane. Dobrze przygotowane dane mogą znacznie poprawić uzyskane wyniki modelowania, podczas gdy dane o niskiej jakości mogą prowadzić do błędnych wniosków i tym samym do nieefektywnych modeli. Uczenie maszynowe wymaga danych, które są reprezentatywne, kompletne, i właściwie przetworzone. Problemy takie jak brakujące wartości, obecność wartości odstających, czy nierównomierny rozkład klas mogą znacznie utrudnić proces uczenia. Ponadto, wybór odpowiednich cech (feature selection) i ich inżynieria (feature engineering) są kluczowymi elementami wstępnego przetwarzania danych, które mają bezpośredni wpływ na zdolność modelu do nauki i generalizacji. W związku z powyższym, praca z danymi w kontekście uczenia maszynowego wymaga nie tylko umiejętności stosowania odpowiednich algorytmów uczenia maszynowego, ale również głębokiego zrozumienia danych i metod ich przetwarzania. To wszystko potwierdza,

jak ważne jest połączenie wiedzy eksperckiej z zaawansowanymi technikami analizy danych w procesie tworzenia efektywnych modeli uczenia maszynowego.

W opisie osiągnięcia naukowego stanowiącego podstawę wniosku o przeprowadzenie postępowania habilitacyjnego został zastosowany podział na kilka obszarów. Najważniejsze wydają się być dwa główne obszary dotyczące analizy danych w kontekście: niezrównoważenia/niezbalansowania klas oraz wielowymiarowości.

Problem niezrównoważenia klas

Problem nierównomiernego rozkładu klas, znany także jako problem niezrównoważonych danych jest niezwykle ważnym wyzwaniem w kontekście uczenia maszynowego, w szczególności w przypadku zadań klasyfikacyjnych. Jeżeli liczba przykładów w jednej klasie znacznie przewyższa liczbę przykładów w jednej lub więcej innych klasach - może to prowadzić do pogorszenia skuteczność modeli uczenia maszynowego, skutkując tym, że modele będą stronnicze w stronę bardziej licznych klas kosztem tych mniej licznych. Nierównomierny rozkład klas jest powszechny w wielu rzeczywistych zbiorach danych, zwłaszcza w tych związanych z zastosowaniem w medycynie (np. wykrywanie rzadkich chorób), w detekcji oszustw finansowych (gdzie większość transakcji przebiegała prawidłowo), czy też w systemach wykrywania ataków (które występują znacznie rzadziej aniżeli normalna aktywność). Przyczyna tego problemu często wynika więc z samej natury zbieranych danych - zdarzenia rzadsze są trudniejsze do zaobserwowania i tym samym są słabiej reprezentowane w danych. Habilitantka motywuje zajęcie się tą tematyką rozpoczęciem współpracy z pracownikami Śląskiego Uniwersytetu Medycznego w Katowicach - co wydaje się jak najbardziej zasadne, gdyż jak wspomniano wyżej dane dotyczące zdrowych pacjentów często stanowią większość próbek poddawanych analizie, a modele uczenia maszynowego trenowane na niezrównoważonych danych mogą lepiej radzić sobie z przewidywaniem obserwacji z klasy dominującej kosztem słabszej skuteczności w przypadku klas mniejszościowych, co w ekstremalnych przypadkach może nawet powodować ignorowanie klasy mniejszościowej, co może być wysoce problematyczne w zadaniach, gdzie wykrycie rzadkich zdarzeń jest kluczowe (np. wczesne wykrywanie choroby).

Typowe rozwiązania tego problemu bazują na modyfikacji zbioru danych poprzez nadpróbkowanie klasy mniejszościowej (Oversampling) lub podpróbkowanie klasy większościowej (Undersampling). Stosuje się również modyfikację algorytmu uczenia poprzez stosowanie funkcji kosztu w taki sposób, aby penalizować błędne klasyfikacje w klasach mniejszościowych bardziej niż w klasie większościowej albo zastosowanie technik uczenia zespołowego (ensemble learning). Ważnym aspektem jest również fakt, aby w procesie ewaluacji modelu używać miar ewaluacji wrażliwych na nierównomierny rozkład klas, takich jak miara F1, precyzja, czułość, czy obszar pod krzywą ROC (AUC-ROC), zamiast tradycyjnej dokładności.

Habilitantka w swoich pracach wpisuje się w powyżej opisane metody radzenia sobie z niezrównoważeniem klas proponując m.in. autorskie algorytmy równoważenia danych. W pracy [MB1] dokonuje weryfikacji użyteczności stosowania różnych sposobów równoważenia klas w odniesieniu do danych dotyczących złamań osteoporotycznych, wykorzystując 3 metody balansowania (SMOTE, RU, ENN), 17 różnych klasyfikatorów oraz różne wskaźniki oceny (czułość, swoistość, BAcc, G-Mean, MCC, AUC). Autorski wkład habilitantki polega na opracowaniu metody oceny analizowanych sposobów równoważenia danych w kontekście testowanych algorytmów klasyfikacji oraz opracowanie koncepcji tzw. macierzy kontrolnych pozwalających na znalezienie odpowiedniego poziomu balansowania. Autorskie algorytmy równoważenia danych opisane w publikacjach [MB2-MB4] są rozwiązaniami działającymi na poziomie danych. Algo-

rytmy podpróbkowania bazują na idei analizowania najbliższego sąsiedztwa obiektów klasy większościowej i różnią się między sobą sposobem wyboru finalnego zestawu obiektów do usunięcia. W pracy [MB2] zaproponowani algorytm KNN Order, którego działanie polega na znajdowaniu i przerzedzaniu skupisk obiektów klasy większościowej. Wykorzystano 18 zbiorów danych o różnym stopniu niezrównoważenia klas, użyto 6 klasyfikatorów, ewaluacji dokonano kilkoma metrykami (Sensitivity, Specificity, BAcc, G-Mean i Kappa). Algorytm KNN RU zaproponowany w pracy [MB3] stanowi połączenie losowego podpróbkowania i idei analizowania najbliższego sąsiedztw. Wykorzystano analogiczne zbiory danych, klasyfikatory i miary oceny - jak w pracy [MB2]. Praca [MB4] proponuje algorytm KNN Near, który polega na usuwaniu najbliższych sąsiadów dla każdego z punktów klasy większościowej. Algorytm ten łączy niejako założenia algorytmów KNN RU i KNN Order. Przetestowany został na danych syntetycznych oraz na zestawie danych wykorzystywanych do testowania poprzednich rozwiązań. Wkład autorski habilitantki polegał również na adaptacji i dostrojeniu parametrów zaproponowanych algorytmów do badanych zbiorów danych.

W ramach wymienionych publikacji habilitantka wykazuje, że zastosowanie proponowanych algorytmów w wielu przypadkach skutkuje zauważalną poprawą jakości klasyfikacji, co potwierdza wkład autorski w prace dotyczące problemu równoważenia klas. Zauważalna jest jednak pewna niekonsekwencja w raportowaniu wkładu autorskiego. Jako oryginalne osiągnięcia w pracach dotyczących niezrównoważenia klas habilitantka wskazuje m.in. wszystkie trzy wspomniane powyżej autorskie algorytmy. Natomiast „Wykazie osiągnięć ..." na str. 5 przy artykule [MB2] habilitantka zaznacza jedynie udział w opracowaniu koncepcji algorytmu KNN Order, a nie całkowite autorstwo tego algorytmu.

Redukcja wymiarowości

Redukcja wymiarowości może mieć znaczący wpływ na zdolności uogólniające modeli uczenia maszynowego - poprzez zmniejszenie liczby cech (wymiarów) danych, proces ten wpływa na:

- zmniejszenie ryzyka przeuczenia, czyli sytuacji, w której model zbyt dokładnie dopasowuje się do danych treningowych, włącznie z ich szumem, co obniża jego zdolność do generalizacji na nowych danych. Redukcja wymiarowości poprzez eliminację nieistotnych lub redundantnych cech może pomóc w zmniejszeniu tego ryzyka.

- usprawnienie obliczeń: przetwarzanie i analiza danych o mniejszej liczbie wymiarów wymaga mniej zasobów obliczeniowych. Modele trenowane na uproszczonych zbiorach danych mogą być szybsze i bardziej efektywne.

- lepszą interpretowalność: dane o mniejszej liczbie wymiarów są łatwiejsze do zrozumienia i zinterpretowania, a to z kolei pozwala na lepsze zrozumienie działania modelu i łatwiejszą identyfikację cech, które mają największy wpływ na proces decyzyjny modelu.

Redukcja wymiarowości pomaga również na oddzieleniu "szumu" w danych, poprzez usunięcie cech, które nie wnoszą wartościowej informacji. Można wtedy skoncentrować się na zmiennych, które faktycznie przyczyniają się do predykcyjności modelu, co poprawia jego uogólnienie.

Selekcja i ekstrakcja cech to dwa podstawowe podejścia stosowane w obróbce danych, które mają na celu poprawę wydajności modeli uczenia maszynowego poprzez

redukcję wymiarowości danych. Oba podejścia zmierzają do podobnego celu, ale wykorzystują do tego celu różne strategie. Selekcja cech (Feature Selection) polega na wyborze podzbioru istotnych cech z oryginalnego zestawu danych. Celem jest usunięcie nieistotnych lub redundantnych cech, które nie przyczyniają się do mocy predykcyjnej modelu lub nawet ją obniżają. Selekcja cech pozwala na zmniejszenie złożoności modelu, skrócenie czasu trenowania i często poprawę wydajności modelu. Podstawowe metody selekcji cech to: metody filtrujące (filter methods), które oceniają cechy na podstawie statystycznych miar i wybierają te, które najlepiej korelują ze zmienną docelową, niezależnie od modelu; metody opakowujące (wrapper methods), które wykorzystują predykcyjne modele do oceny kombinacji cech i wybierają te, które najlepiej sprawdzają się w danym modelu (np. backward elimination lub forward selection) oraz metody wbudowane (embedded methods), które integrują selekcję cech jako część procesu trenowania modelu (np. LASSO (Least Absolute Shrinkage and Selection Operator), który podczas trenowania modelu regresji jednocześnie dokonuje selekcji cech). Ekstrakcja cech (feature extraction) polega na przekształceniu oryginalnych danych wejściowych w nowy zestaw cech, który jest mniejszy pod względem wymiarowości, ale nadal zachowuje kluczowe informacje. Często proces ten polega na transformacji danych do nowej przestrzeni cech, w której cechy są mniej skorelowane. Podstawowe metody ekstrakcji cech to analiza głównych składowych (PCA), która zmniejsza wymiarowość danych przez projekcję ich na przestrzeń, gdzie zmienne są nieskorelowane; Linear Discriminant Analysis (LDA), która maksymalizuje separowalność między klasami przy jednoczesnej minimalizacji wariancji wewnątrz klasy oraz dwie nieliniowe techniki ekstrakcji cech stosowane głównie do wizualizacji danych wysokowymiarowych: t-Distributed Stochastic Neighbor Embedding (t-SNE) i Uniform Manifold Approximation and Projection (UMAP).

W przypadku danych medycznych, dla których bardzo ważna jest interpretowalność modelu, naturalnym wyborem w procesie redukcji wymiarowości są metody selekcji cech. Problemem metod ekstrakcji cech jest to, że cechy wytworzone w ten sposób mogą być niejasne dla ludzkiego rozumienia. Prace [MB5] i [MB6] dotyczą analizy skuteczności różnych algorytmów selekcji cech. Praca [MB5] dokonuje weryfikacji skuteczności różnych metod selekcji cech. Przedstawia zalety i wady poszczególnych rozwiązań oraz wydajność każdego z analizowanych klasyfikatorów po zastosowaniu danej metody selekcji cech. Wniosek - zmniejszenie liczby cech ułatwia interpretację wyników oraz może mieć pozytywny wpływ na dokładność klasyfikacji. Jako ważny wkład autorski zostało wskazane opracowanie wskazówek dotyczących doboru metod selekcji. Praca [MB6] zawiera ocenę czy dostosowanie wag na etapie selekcji cech może pomóc w lepszym rozróżnianiu klas, szczególnie poprzez priorytetyzowanie cech ważnych dla klasy mniejszościowej. Zostało to wskazane przez habilitantkę jako jeden z ważniejszych wkładów jej pracy w tematyce redukcji wymiarowości.

W ramach tych prac wykazano, że redukcja cech nie tylko ułatwia interpretację wyników, ale może mieć również pozytywny wpływ na dokładność klasyfikacji. Jako ważny wkład autorski zostało wskazane nie tylko opracowanie wskazówek dotyczących doboru metod selekcji oraz adaptacja i dostrojenie wybranych metod selekcji cech do uwzględnienia macierzy kosztów.

Habilitantka wskazuje dodatkowo jako oryginalny element osiągnięcia zastosowanie wykresów profili szans w procesie dyskretyzacji zmiennych ciągłych oraz ocenę korelacji, kolinearności i multikolinearności w procesie selekcji cech i wykrywania zmiennych zakłócających. Z tymi zagadnieniami związane są prace zawarte w [MB7] i [MB8].

Według mojej oceny opisane powyżej prace habilitantki, związane z problemem

redukcji wymiarowości, stanowią wkład autorski w omawianą tematykę.

Wniosek końcowy

Podsumowując całość niniejszej recenzji, uważam, zeprezentowany cykl 8 publikacji zawiera znaczący wkład habilitantki w rozwój dyscypliny naukowej Informatyka techniczna i telekomunikacja. Według mojej oceny niniejszy cykl publikacji stanowi oryginalne osiągnięcie naukowe dr inż. Małgorzaty Bach co stanowi spełnienie wymagań formalnych w odniesieniu do postępowania habilitacyjnego. Moja recenzja jest więc jednoznacznie pozytywna w kwestii tego, że oceniane osiągnięcie naukowe powinno stanowić podstawę do nadania dr inż. Małgorzacie Bach stopnia doktora habilitowanego.