Prof. dr hab. inż. **Jerzy Stefanowski**
Członek korespondent PAN
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2
60-965 Poznań

Poznań, 2024

Recenzja rozprawy doktorskiej mgr Katarzyny Woźnicy pt. „Towards Trustworthy Automated Data Science"

1. Uwagi ogólne

Niniejsza recenzja jest sporządzona w odpowiedzi na pismo z Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiego (uchwała nr 722/2024 w sprawie wyznaczenia recenzentów rozprawy doktorskiej w postępowaniu w sprawie nadania stopnia doktora mgr Katarzyny Woźnicy), podpisanego przez profesorów J. Arabasa i A. Janickiego.

Przedmiotem recenzji jest oprawione wydanie manuskryptu zatytułowanego „**Towards Trustworthy Automated Data Science**", który jest rozprawą doktorską p. mgr Katarzyny Woźnicy.

Konstrukcyjnie ta rozprawa obejmuje zarówno wstęp, podstawowe informacje o rozważanych metodach (napisane specjalne do tego wydania) jak i odpowiedniki pięciu artykułów (opublikowanych przez doktorantkę w różnych układach współautorskich). Można więc uznać, że podstawą rozprawy jest tzw. kolekcja artykułów.

Odnosząc się do wymagań ustawowych oraz tzw. zwyczajowych (regulowanych przez niektóre Rady Dyscyplin lub Naukowe PAN) rozprawa doktorska nie musi być typowym maszynopisem, lecz może być przedstawiona w formie powiązanego tematycznie zbioru artykułów (opublikowanych lub przyjętych do druku w czasopismach naukowych, określonych przez Ministra właściwego do spraw nauki) – art. 187 ustawy. Ponadto inny artykuł ustawy stwierdza, że doktorant/ doktorantka powinien mieć choć jeden artykuł opublikowany w czasopiśmie naukowym lub w recenzowanych materiałach z konferencji międzynarodowej, które w roku opublikowania artykułu w ostatecznej formie były ujęte w wykazie sporządzonym przez Ministra – co jest w przypadku recenzowanej rozprawy spełnione. Także w mojej opinii spełniono wymóg powiązania tematycznego.

Recenzowana rozprawa doktorska dotyczy problematyki zaawansowanych systemów uczących się, tzw. predykcyjnych, choć w rozprawie są one rozważane dla zadania klasyfikacji przykładów. Doktorantka odnosi się szerzej do nauki o danych, lub w moim odczuciu **eksploracji danych** (poprawniejszy termin ang. **Data Science** /skrót DS/– choć dla mnie wiele z omawianych zadań odpowiada etapom procesu Odkrywania Wiedzy z Danych – ang. Knowledge Discovery from Data), która obejmuje nie tylko uczenie samego modelu predykcyjnego, lecz etapy przetwarzania wstępnego danych (w tym niezwykle istotny etap tzw. inżynierii doboru cech) oraz oceny i eksploatacji modelu. Takie zadania mają obecnie wiele zastosowań praktycznych.

W ogólności omawiana eksploracja danych jest to dość złożony proces, związany charakterem analizowanych danych i oczekiwań w danych zastosowaniu, często o charakterze wysoce eksploracyjnym i iteracyjnym. Ze względu na wykorzystywanie coraz bardziej złożonych systemów uczących się nie jest łatwe nawet dla specjalistów dostrajanie parametrów. Kontekst zastosowania (i dostępnej wiedzy dziedzinowej) może odgrywać kluczową rolę w tych eksploracyjnych krokach, nawet w stosunkowo dobrze zdefiniowanych procesach, takich jak modelowanie predykcyjne z wykorzystaniem systemów uczących się (ang. machine learning – w skrócie ML).

We wstępie rozprawy doktorantka odnosi się do podstawowych etapów tego procesu – inżynierii danych, tworzenia modeli predykcyjnych (wybór modelu oraz dostrajanie ich parametrów) oraz ich eksploatacji i przedstawia (w zasadzie zgodnie z literaturą) znane argumenty za możliwością automatyzacji tych etapów (przede wszystkim dwóch pierwszych). Od lat możemy obserwować takie wysiłki przede wszystkim w zakresie tzw. **automatyzowanego uczenia maszynowego** (tzw. **AutoML**), lecz także choć częściowej automatyzacji pewnych etapów w przetwarzaniu wstępnym danych. Systemy AutoML, takie jak np. biblioteki Pythona Auto-sklearn, wykorzystują heurystyczne strategie albo optymalizacje Bayesowską do automatycznego wybierania algorytmów i dostrajania

parametrów. Badania i eksperymenty wykazały, że różne systemy AutoML często mogą optymalizować przetwarzanie danych i dostarczać szybciej trafniejsze modele predykcyjne niż ręcznie dobierane - ma to duże znaczenie dla mniej doświadczanych badaczy oraz tzw. inżynierów ML w praktyce przemysłowej (co przyczynia się do tzw. demokratyzacji stosowania uczenia maszynowego w zastosowaniach). Szerszy obraz obecnych trendów w automatyzacji eksploracji danych można przeanalizować w wielu artykułach, np. Automating Data Science: Prospects and Challenges. Tijl De Bie et al Communications of ACM (2022) [cytowany także przez doktorantkę jak poz. 42].

W mojej opinii, choć w dużym stopniu można osadzić przedstawione rezultaty jednak w automatyzacji uczenia maszynowego (pomimo szerszego tytułu rozprawy) to jednak część z rozdziałów pracy mieści bardziej w rozszerzeniach obecnego wątku systemów klasy AutoML, lecz tylko niektóre wychodzą poza to, zwłaszcza z uwagi na interesujące wykorzystywanie metod **wyjaśnialności sztucznej inteligencji** (ang. XAI) – co jest także specjalnością zespołu MI2 prof. Przemysława Biecka, z którego pochodzi doktorantka. Wydaje mi się, że rozważane tworzenie modeli (np. automatyczny wybór, lecz przede wszystkich konfiguracja i dostrajanie w oparciu o meta-cechy oraz ich transfer do innych danych / zadań) oraz eksploatacja (tutaj raczej różne spojrzenie na ocenę modeli) ma oryginalny charakter.

Ponadto uważam takie systemy (jak rozważa doktorantka) powinny być zintegrowane z doświadczeniem ludzkim i wiedzą specjalistyczną w danej dziedzinie (jeśli jest dostępna tzw. wiedza dziedzinowa, np. ontologie), z naciskiem na uzupełnianie i ulepszanie pracy ekspertów ludzkich, ale w mojej opinie jednak nie na pełną mechanizację – co także dostrzegam w części przedstawionego materiału.

Przechodząc do hipotez i celów rozprawy – doktorantka wychodzi od postulatu potrzeby silniejszej współpracy z ekspertami / użytkownikami systemów oraz postulatu większego **budowania ich zaufania do samego procesu automatyzacji**. Mogę tutaj podzielać przedstawione motywacje, że użytkownik (nawet doświadczony) nie będzie ufał automatyzacji w wersji „black box", lecz będzie oczekiwał pewnej dodatkowej informacji, przejrzystości i wspierania zrozumienia wykonywanych operacji.

Pani mgr Katarzyna Woźnica zamierza to osiągnąć w dwóch różnych perspektywach działania: (1) wspierania metod oceny zarówno przetwarzanie wstępnego jak i doboru meta-cech i porównywania modeli predykcyjnych; (2) wprowadzenia wiedzy dziedzinowej do procesu meta-uczenia.

Zostało to przedstawione w rozdziałach 1.2.1 oraz przede wszystkim w rozdziale 1.3 maszynopisu jako pięć hipotez (wymieniam je w punkcie 2 mojej recenzji) – które mają być zweryfikowane przez wyniki (propozycje metod oraz eksperymentalnie) w opublikowanych 5 artykułach.

Same artykuły są trzykrotnie pracami konferencyjno-warsztatowymi (dobra lokalizacja przy konferencji ECMLPKDD) oraz pracami w dwóch bardzo dobrych czasopismach Machine Learning oraz Knowlede-based Systems (co bardzo dobrze oceniam). Wszystkie artykuły są wieloautorskie, lecz dwie pracy to wyłącznie działa Pani Katarzyny Woźnicy i jej promotora.

W mojej opinii tematyka, motywacje rozprawy są dobrze uzasadnione, interesujące oraz dotyczą otwartych i nowych problemów badawczych. Wątek wykorzystania metod wyjaśnialności i dodatkowego wykorzystywania wiedzy (zarówno dziedzinowej jak i przenoszenia doświadczeń z wcześniej badanych zbiorów do meta-uczelnia) są dość oryginalne. Reasumując przedstawione cele są niewątpliwie warte dysertacji doktorskiej.

2. Ocena struktury i zawartości rozprawy

Recenzowana rozprawa jest napisana w języku angielskim. Dostarczony maszynopis zawiera 8 rozdziałów oraz spis literatury = łącznie liczy 190 stron, choć jest to chyba związane z tym, że rozdziały 3-7 są odpowiednikami opublikowanych już wcześniej artykułów.

Na początku rozprawy dołączono polskojęzyczne streszczenie, które jest dość zgrabnie napisane, choć zawiera zdecydowanie za dużo „slangu ML" i „kalek" terminów z języka angielskiego – co jest zaskakujące dla czytelnika.

Organizacji maszynopisu jest dobra. Niewątpliwie autorskie są dwa pierwsze rozdziały, gdzie w rozdziale pierwszych Autorka najpierw bardzo krótko wprowadza motywacje do potrzeby

automatyzacji procesu uczenia systemów ML oraz wsparcia dostrajania ich parametrów. Jest to dość standardowa argumentacja zgodna z literaturą przedmiotu oraz potrzebami praktycznymi, także z odniesieniem do bibliotek dostępnych dla badaczy oraz inżynierów ML pracujących w przemyśle. Właściwe jest przejście do automatyzacji uczenia maszynowego szerzej do automatyzacji eksploracji danych, wraz z wcześniejszymi etapami przygotowania danych (co nie jest tak rozwinięte w bibliotekach AutoML). Podział całości automatyzacji na inżynierię danych, eksploracje modeli oraz eksploatacji zapożyczony z literatury jest bardzo dobrze przedstawiony i opisany. W drugiej części wstępu w miarę klasycznie zarysowano tzw. braki w obecnym stanie wiedzy i sformułowano pięć hipotez badawczych. Zastanawiając się na ich sformułowaniem – brzmią bardziej jak tezy niż pytania w hipotezach – ale uznaję, że są dobrze zarysowane.

Drugi rozdział jest bardzo krótkim wprowadzeniem do wybranych zagadnień i operacji w ramach automatyzacji eksploracji danych (ang. AutoDS). Autorka starała się sformalizować zapis matematycznie – choć z uwagi na to, że późniejsze rozdziały związane z artykułami i tak tego nie wykorzystują – odbieram to jako rodzaj „sztuki notacji dla samej sztuki". Dalsza część rozdziału jest dużo bardziej interesująco, gdyż zawiera one b. zwięzłe przedstawienie metod będących odniesieniem albo wykorzystywanych w dalszych rozdziałach. Z uwagi na ten związek podrozdział hyperparameter portfolio mógł by być bardziej dogłębnie omówiony z szerszym obrazem obecnego stanu badań. Za to techniki wyjaśnialności systemów predykcyjnych ML są zwięźle, ale dobrze przedstawione. Zamieszczone ilustracje są czytelne.

Kolejne rozdziały 3 – 7 odpowiadają postawianym hipotezom badawczym, tj. rozdział 3 H1: Validation of preprocessing techniques can simplify the exploration operations within Data Exploration – w praktyce dotyczy porównania technik zastępowanie nieznanych wartości atrybutów i ich wpływu na trafność modeli predykcyjnych. Kolejny rozdział 4 realizuje hipotezę H2 (Levaraging exploratory machine learning models extends the validation process for the quality of meta-models) i w mojej opinii zawiera bardzo oryginalne podejście do użycia metod wyjaśnialności do oceny wpływu tworzonych meta-cech dla jakości meta-uczenia. Rozdział 5ty to odpowiedź na hipotezę H3 (Integrating the meta-measure with probabilistic interpretation of model performance enhances benchmarks validation and provides easily comprehensible insights into model quality, facilitating model monitoring by non-expert in the field) – wprowadzający nowy sposób porównywania predykcji modeli poprzez EPP score oraz obszerne studia eksperymentalne jego wykorzystania.

Dalsze dwa rozdziały dotyczą bardziej wykorzystywania wiedzy w poprawie meta-uczenia. I tak rozdział 6 jest poświęcony hipotezie H4 (considering the specificity of the new prediction may lead to a greater transfer of hyperparamters compared to transferring from genetic datasets …) – w zasadzie nie ma typowej wiedzy dziedzinowej, lecz właśnie przenoszenie informacji z wcześniej rozwiązywanych zadań, danych do nowych poprzez hiperparametry w ramach nowej propozycji tzw. skonsolidowanego uczenia się. No i ostatni rozdział 7 to hipoteza H5 (Incorporating ontologie sinto Data Exploration enables the injection of domain-specific information and modeling knowledge from diverse prediction problems enriching the understanding of data) i zawiera oryginalną propozycję uwzględniania ontologii medycznych w strukturze SeFNet.

Na ile sprawdziłem oryginalne prace – to rozdziały w zasadzie są ich kopią – stąd patrzę na całość maszynopisu jako tzw. tematycznie spójna kolekcja artykułów.

Pani mgr Katarzyna Woźnica słusznie poprzedza każdy z rozdziałów odniesieniem się do konkretnego artykułu oraz przedstawienia odniesienia / związku tego tekstu z problematyką automatyzacji eksploracji danych oraz postawioną konkretną hipotezą. Przy czym z uwagi na dość dużą różnorodność tekstów, ich konstrukcji oraz pewnych ograniczeń (np. prace konferencyjne albo tzw. workshopowe mają najczęściej tzw. limit maksymalnej liczby stron) te rozdziały są jakby „nierówne" dla czytelnika. W niektórych (np. 6) mamy w ramach tzw. related works powtórzenie pojęć już wcześniej zdefiniowanych, a w innych przypadkach można by oczekiwać podania więcej szczegółów, zwłaszcza związanych z eksperymentami – niestety często mi brakuje wszystkich szczegółów, wyborem metod lub interpretacją ich działania. Lecz przy przyjętej koncepcji przedstawienia rozprawy jako kolekcji artykułów jest to niestety typowe (i trochę utrudnia czytanie). Być może dodanie specjalnych załączników przy niektórych z artykułów, by to ulepszyło. Trochę można ponarzekać, że może pisanie klasycznego maszynopisu by było lepsze. Lecz podsumowując, jak na moje doświadczenie z recenzowania doktoratów jako kolekcje artykułów – rozprawa p.

Katarzyny Woźnicy jest całkiem dobrze przemyślana, po to, aby pokazać spójność tematyczną oraz sprawnie skonstruowana jako „mikro książka".

Zauważmy sprawdzając profil Google scholar p. Katarzyny Woźnicy, że jest współautorką aż 26 prac (wyrazy uznania z mojej strony), więc i tak wspólnie z promotorem całkiem nieźle je wybrali.

Całość maszynopisu kończy krótkie podsumowanie (rozdział 8), gdzie zgodnie z nazwą Autorka syntetycznie opisała dokonania i w dość ograniczony sposób wspomniała o ograniczeniach i potrzebie dalszych badań. Brakuje mi jednak wyraźniejszego przedstawienie oryginalności swoich propozycji na tle aktualnego stanu badań (tzw. own contribution in the context of earlier works, z j. ang.) oraz dyskusji przydatności tych propozycji dla budowy zaufania do systemów AutoML i AutoDS (patrz moje uwagi w punkcie 4 recenzji). Pomimo tej uwagi rozdział ten jest skonstruowany w miarę standardowo w odniesieniu do wymagań metodologii nauki oraz zasad pisania prac naukowych.

3. Ocena wkładu oryginalnego

W poprzednim punkcie w opisie poszczególnych rozdziałów wielokrotnie wskazywałem wiele oryginalnych propozycji doktorantki w zakresie postawionych pięciu hipotez badawczych. W stosunku do każdej z nich możemy w rozdziałach 3-7 odnaleźć autorskie i ciekawe dokonania. Jednak w mojej opinii za najważniejsze wskazuję trzy następujące (zwłaszcza w kontekście tematyki rozprawy doktorskiej związanej z automatyzacją uczenia maszynowego i eksploracji danych):

- Zaproponowanie użycia technik wyjaśnialnej sztucznej inteligencji (w zakresie oceny znaczenia cech) przy ocenie meta-cech budowanych dla zadania meta-uczenie – tj. propozycje przedstawione w rozdziale 4tym zatytułowanym „Towards explainable meta-learning"). Większość zadań w ramach automatycznego meta-uczenia się, gdzie przenosi się doświadczenie z analizy jednych zbiorów danych na inny związane jest z wyborem właściwych meta-cech. Doktorantka pokazała, że popularne techniki XAI związane z analizą znaczenia cech, takie jak permutacyjne ocena ważności cech, profile Ceterius Paribus (dla mnie odmiana znanych (j. ang) partial dependency plots stosowanych często w statystycznej eksploracji danych oraz w ramach XAI) oraz bardziej statystyczne narzędzia wykorzystujące H-Statystykę Friedmana i wykresy triplot mogą być skutecznie wykorzystane do oceny ważności rozważanych (nawet b. złożonych) meta-cech i badania ich interakcji oraz skorelowania, a także wpływu na strojenie tzw. zastępczego modelu (poprzez tzw. Individual Conditional Expectation profiles – patrz rozdział 4.8). Użyteczność tych propozycji została sprawnie przedstawiona na podstawie realizacji procesy meta-uczenia się na tzw. danych z OpenML100 data sets, gdzie rozważano wiele meta-cech pochodzących zarówno ze typowych statystyk danych, lecz także uruchamiania klasycznych algorytmów i ich parametrów. Zwróćmy uwagę, że obecne podejścia skupiają się na poszukiwaniu najlepszego meta-modelu, ale na ogół nie wyjaśniają, w jaki sposób te różne cechy przyczyniają się do jego skuteczności. Pomimo, że przedstawiona propozycja twórczo wykorzystuje znane już metody to w mojej opinii jest bardzo ciekawym rozwiązaniem dla głębszego zrozumienia znaczenia i wpływu meta-cech na strojenie tego modelu oraz w pełni realizuje przedstawioną motywację (i chyba najpełniej tytuł) rozprawy. Zgodnie z moja wiedzą nie spotkałem się z takim rozwiązanie w literaturze, dlatego podkreślałbym jej oryginalność w zakresie włączenia użytkownika w ocenę i współpracę z metodami metą-uczenia.
- Zaproponowanie nowego sposobu porównywania predykcji modeli uczenia maszynowego wykorzystujące podejście z meta-wskaźnikiem EPP (z j. ang. Elo-based Predictive Power). Doktorantka inspirowała się podejściem stosowanych w analizie rozgrywek sportowych do oceny względnego rankingu graczy na podstawie ich bezpośrednich pojedynków (tzw. Elo-system rankingowy) do przeniesienia na uporządkowanie modeli predykcyjnych testowanych porównawczo na wielu zbiorach danych. W stosunku do rozwiązań dominujących w literaturze (głównie wykorzystujących nieparametryczne testy statystyczne z poprawkami) ta propozycja jest dość nietypowa i oryginalna. Ponadto pozwala m.in. na przyrostową aktualizację tego współczynnika oraz oceny szansy wygrania (lepszego działania modelu) na nierozważanych nowych danych oraz przede wszystkim pokazuje ciekawą probabilistyczną interpretację wygranej oraz różnic w ocenach dwóch modeli. Propozycja jest wsparta analizą teoretyczną (pokazujące dość solidne odniesienia do podstaw statystycznych – co jest dużą częścią tego

rozdziału) i obszernymi eksperymentami obliczeniowymi popularnych modeli na benchmarkach VTAB oraz przede wszystkim OpenML – które są ciekawie intepretowane dla odbiorcy. Tą oryginalną propozycję można odebrać jako kolejne spojrzenie na ocenę i porównywania modeli predykcyjnych uczonych z wielu zbiorów danych – co jest problemem samym w sobie dla środowiska uczenia maszynowego. Jej związek z oceną modeli dla automatycznego uczenia maszynowego istnieje potencjalnie, choć w treści tej pracy nie został silnie wyeksploatowany, dlatego żałuję, że nie rozwinięto tego we wstępie do tego rozdziału albo rozdziale końcowym manuskryptu rozprawy.

- Propozycja tzw. consolidated learning przedstawiona w rozdziale 6. Pomimo umieszczenia go w części związanej z uwzględnianiem tzw. wiedzy dziedzinowej w AutoDS, dla mnie esencja tej propozycji to specyficzne wykorzystanie doświadczeń o poprzednio rozwiązywanych problemach danych (poprzez podobne zmienne) dla lepszego skonstruowania zbioru do meta-uczenia oraz lepszego przekazania informacji do strojenia hiperparametrów (w rozważanych podejścia do meta-uczenia z optymalizacją tzw. hyperparameter porfolio) co może doprowadzić lepszej realizacji samego meta-uczenia. Jest to zaprezentowane na przykładzie odpowiednio skonstruowanej zestawu danych metaMIMIC pochodzącego z repozytoriów danych medycznych – zostało to zaproponowane jako rodzaj specjalizowanego tzw. benchmarku dla tego typu skonsolidowanego meta-uczenia (zgodnie z moim zrozumieniem opisu jest to także oryginalne dokonanie w zakresie propozycji specjalizowanych danych dla tej odmiany meta-uczenia). Następnym osiągnięciem jest pokazanie lepszego działania zaproponowanego podejść do prostszej optymalizacji w przeniesieniu informacji o hiperparametrów dla meta-klasyfikatora wzmacnianych drzew XGBoost. Ponadto takie podejście może potencjalnie zwiększać transparentność samego procesu meta-uczenia, co jest zgodne z celami rozprawy.

Ponadto dostrzegam dwa kolejne osiągniecia:

- Po pierwsze, przedstawienie w rozdziale 7mym bezpośredniego wykorzystanie wiedzy dziedzinowej o rozważanych problemach. Doktoranta zaproponowała specjalną strukturą meta-danych o cechach o nazwie ang. Semantic Feature Net, która dokonuje powiązania wybranych cech pomiędzy parą dwóch zbiorów danych. Przydatność tej propozycji pokazano z wykorzystaniem danych medycznych z kolekcji wybranych repozytoriów (stworzenie takiego specjalizowanego repozytorium można uznać także za oryginalny, acz bardziej praktyczny wkład dla środowiska badawczy). Ciekawsze metodologicznie jest dyskutowanie podobieństw pomiędzy parą tak powiązanych dwóch zbiorów danych, które wychodzi od znanych miar podobieństw termów (tutaj cech) i zostanie uogólnione dla szerszego podobieństwa (propozycja Dataset Ontology Semantic Similarity) – bardziej przydatnego dla zadań meta-uczenia. Autorka przedstawia tutaj przykład użycia (w rozdziale 7.5) z dokładną analizą podobnych cech pomiędzy zbiorami danych medycznych oraz wykorzystania ich w ramach strojenia parametrów meta-uczenia.
- Po drugie, w rozdziale 3 doktorantka przedstawiła studium eksperymentalne porównujące wiele popularnych metod podstawiania nieznanych wartości atrybutów w zadaniach klasyfikacyjnych oraz statystycznie przeanalizowała otrzymane wyniki – w zasadzie dochodząc do znanej wcześniej konkluzji mówiącej, że nie ma uniwersalnie najlepszej metody. Nie do końca zgadzam się z Autorką twierdzącą, że nie było wcześniej już studiów eksperymentalnych na tej temat – może tutaj zaprezentowane jest obszerniejsze (ale akurat tutaj odniesienie się do literatury nie jest właściwe – patrz uwagi krytyczne). Ponadto akurat takie studium porównawcze ma znaczenie dla ogólnej wiedzy o przetwarzaniu wstępnym danych a wydaje mi się, że nie jest zbyt powiązane z automatycznym meta-uczeniem, co jest przedmiotem rozprawy. Podejrzewam, że doktorantka chciała wykorzystać tą pracę jako ilustracje możliwości oceny dla tzw. inżynierii oryginalnych cech w AutoDS – lecz dla mnie jest to dyskusyjne i nie wyeksploatowane – sam artykuł jest zbyt oderwany tematycznie od głównego wątku tematyki rozprawy.

Krótko podsumowując wymienione osiągnięcia uznaję, że doktorantce w pełni udało się odpowiedzieć na postawionych pięć hipotez roboczych, które przerodziły się w cele rozprawy. Wprawdzie część z pojęć albo metod była już wcześniej znana (lub dyskutowana w literaturze, lecz w

innym kontekście), ale tutaj zostały w przemyślany i uporządkowany sposób połączone z własnymi propozycjami doktorantki i jej współautorów w nowe podejścia rozszerzając obecny stan wiedzy o meta-uczeniu.

4. Uwagi dyskusyjne i polemiczne

Nie kwestionując wartości wyników zawartych w rozprawie zgłaszam poniżej kilka uwag krytycznych i dyskusyjnych. Część dotyczy wskazania tych aspektów, które można było potraktować szerzej w rozprawie lub mogą być przedmiotem dalszych badań i mam nadzieję, że otworzą dyskusję naukową podczas samej obrony:

- Motywacja rozprawy doktorskiej jest bardzo ambitna – budowa zaufania odbiorców (cyt. Ze wstępu „to improve trust") do systemów AutoML i AutoDS. Natomiast same hipotezy są na ogół bardzo konkretne (co dobrze) i związane z konkretnymi zadaniami w procesie eksploracji danych i samego meta-uczenia. Część z osiągnięć jest nastawiana na predykcje oraz poprawę optymalizacji parametrów. Tylko niektóre bezpośrednio odnoszą się do interpretacji meta-cech, lepszego zrozumienia pomiędzy cechami i tylko jedna z propozycji bezpośrednio wykorzystuje metody wyjaśnialności cech. Pomimo, że to są ciekawe propozycje, to jednak niekoniecznie odnoszą się wprost do sztucznej inteligencji godnej zaufania (którą się definiuje z innymi charakterystykami niż tylko przejrzystość procesu albo objaśnienia). W artykułach z rozprawy brakuję tych dodatkowych elementów, nie ma też studiów z ludźmi, odbiorcami na ile oni oceniają przedstawione propozycje. Samo eksperymentalne badania zaufania do systemów sztucznej inteligencji jest ciągle bardzo trudne, ale warto by ten wątek dyskutować w tej rozprawie jeśli jest to jej główna motywacja. Nie ma także zbyt wielu studiów potencjalnej interpretacji wyników przez ew. odbiorców (takich jak w rozdziale 4tym), wiele wyników eksperymentów odnosi jednak do predykcji albo właściwości strojenia parametrów. W mojej opini przy tak sformułowanym tytule rozprawy chociaż rozdział podsumowujący mógł być rozbudowany o dyskusje na ile zaprezentowane metody przyczyniają się do poprawy zaufania do systemów AutoML i AutoDS. Prosiłbym, aby doktorantka omówiła to podczas obrony.
- Praca jest jednak kolekcją trochę odrębnych artykułów, z których każdy prezentuje różnorodne zagadnienia. Doktorantka sprawnie umieściła ja w ogólnym schemacie AutoDS np. wg. propozycji Tijl De Bie et al. Niektóre z artykułów dobrze się w to wpisują (patrz poprzedni punkt recenzji) ale niektóre (zwłaszcza rozdział 3) wydają się wstawione trochę „na siłę" a sama hipotez H1 jest dla mnie sformułowana zbyt szeroko w stosunku do dokonania (studium porównawcze metod zastępowania nieznanych wartości atrybutów). Prosiłbym doktorantkę o rozszerzenie informacji o bezpośrednim wykorzystaniu tego w AutoDS oraz przekonania mnie, dlaczego nie rozważano innych metod przetwarzania wstępnego (np. selekcji cech).
- Ponadto specyfika artykułów, zwłaszcza konferencyjnych, ogranicza część zawartych w nich informacji. Co z punktu widzenia czytelnika utrudnia analizę zwłaszcza eksperymentów. Przykładowo w większości używa się zespołu gradientowo wzmacnianych drzew jako meta-modelu, a jako tzw. landmarkes prostych klasyfikatorów jak k-NN, drzew losowych, uogólnionych modeli regresji = nie przedstawiono do końca uzasadnień takich wyborów oraz nie podano wszystkich parametrów, itd. Utrudnia to tzw. odtwarzalność wyników (ang. reproducibility). Domyślam się, że wynikało to z inspiracji literaturowych (np. pozycja [96]) ale może we wstępach do rozdziałów można było dać rozszerzone informacje. Tak samo w rozdziale 5.4.1 nie mam przekonania do masowego uczenia algorytmów z olbrzymią ilością losowo ustawianych konfiguracji hiperparametrów (patrz dół strony 94) – czy mógłbym prosić o uzasadnienie podczas obrony.
- W powyższym kontekście zastanawia mnie, na ile zaprezentowane wyniki będą się uogólniały na inne meta-modele, np. sieci neuronowe?
- Nie do końca zgadzam się z często pisanym zdaniami, że nie było wcześniejszych prac a własne propozycje są unikalne. Doradzałbym więcej ostrożności. Przykładowo techniki zastępowania nieznanych wartości atrybutów są dość klasyczne. Rozważmy artykuł przeglądowy A survey on missing data in machine learning. Tlamelo Emmanuel et al z Journal of Big Data (2021), który

podaje wiele prac eksperymentalnych. W przypadku prostszych klasyfikatorów wiele z tych technik było pod koniec poprzedniego wieku badane eksperymentalnie przez prof. Jerzego Grzymałę Busse i współpracowników (w tym piszącego niniejszą recenzję) – to jest jeden z przykładów, że dyskusja tzw. related works w rozdziale 2 powinna być obszerniejsza. Ponadto problematyka meta-uczenia była też intensywniej wcześniej badana w Polsce, patrz głównie prace prof. Norberta Jankowskiego, i częściowo prof. W.Ducha z UMK w Toruniu (chociażby książka z 2011 Meta-learning in computational intelligence wydana przez Springer - niestety nie cytowana przez Doktorantkę)

- Podobna uwaga dotyczy opisu technik optymalizacji (głownie portfolio hiperparametrów, która jest później często wykorzystywana w wielu z artykułów – rozdziałów. W mojej opinii powinna być bardziej szczegółowo omówiona w rozdziale 2 a także zarysowania przez doktorantke swoich propozycji na tle stanu literatury i innych propozycji rozszerzeń tej metody.
- W przypadku propozycji wskaźnika EPP w rozdziale 5, zastanawiam się na ile można oceniać czy drobna różnica wartości mierników (np. AUC) jest statystycznie znacząca – np. w definicjach na str 84 -85.
- Wiedza dziedzinowa w rozdziałach 6 i 7 – zwłaszcza ontologie są budowane wyłącznie dla danych medycznych (co przy okazji jest ciekawe, i stanowi rodzaj osiągniecia dla tej poddziedziny). Lecz równocześnie otwiera to pytanie o ogólność i doświadczenie z modelowaniem innych problemów niemedycznych.
- Jeśli dobrze zrozumiałem sieć Semantic Feature Net była budowana „ręcznie" ekspercko. Pytanie do doktorantki czy widzi możliwości półautomatycznego wspierania ekspertów.

Sama praca jest na ogół dobrze zredagowana i pisana poprawnym językiem angielskim. Zauważyłem trochę drobnych błędów redakcyjnych lub edytorskich, albo niejasności terminologicznych. Przykładowo:

- Na początku rozdziału 4.7 i 4.8 zaskakująco zdania rozpoczyna się od małych liter; patrz np. str 67 „the black box"
- Str 70 opis jest „in Figure 4.4 on the right panel" – chyba na tym rysunku panele są położone jeden pod drugim
- Str 83 jest źle formatowane dolne indeksy np. Player_i
- Str 100 jest transofiming -> transforming ; itd
- Na str. 23, jest zmieniana kolejność w opisie danych D, raz D(X,Y) a później dla danych testowych D^{new}(Y,X)
- Czasami piszę się features, a czasami też variables
- W polskojęzycznym streszczeniu jest za dużo angielskiego „slangu" – np. tworzenia nowych algorytmów i frameworksów; technika imputacji albo kalek z języka angielskiego (np. kluczowa konieczność)

Powyższe uwagi nie podważają mojej dobrej oceny rozprawy i osiągnięć pani mgr Katarzyny Woźnicy.

5. Konkluzja końcowa

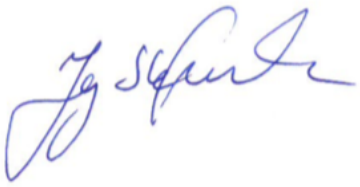
Recenzowana rozprawa charakteryzuje się bardzo dobrym poziomem merytorycznym i zawiera niezwykle interesujące rezultaty, które mogą być przydatne dla badań nad problematyką automatyzacji systemów uczących się oraz przetwarzania danych.

Wszystkie postawione hipotezy rozprawy zostały pozytywnie zweryfikowane.

W mojej opinii doktorantka wykazała się umiejętnościami prowadzenia badań naukowych – zwłaszcza w zakresie wprowadzanie nowych metod oraz intensywnej jej oceny eksperymentalnej, pomysłowością oraz pracowitością. Należy podkreślić bardzo dobre umiejętności matematyczne oraz ciekawe odniesienia do podejść statystycznych, co jest rzadkie w polskich rozprawach poświęconych tej tematyce.

Ponadto chciałbym stwierdzić, że w moim przekonaniu zawartość rozprawy, przedstawione wyniki oraz udokumentowane w spisie literatury publikacje pani mgr Katarzyny Woźnuct (w tym zwłaszcza artykuły przyjęte do wygłoszenia na najważniejszych międzynarodowych konferencjach ECML/PKDD oraz dwa w prestiżowych czasopismach) znacznie przewyższają poziom typowych wymagań stawianych badaniom doktorskim. Zwróćmy także uwagę, że Doktorantka rozwiązała wiele problemów naukowych, a nie tylko jeden.

Podsumowując, uważam, że opiniowana rozprawa Pana mgr Katarzyny Woźnicy **spełnia z nadmiarem warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesieniu do rozpraw doktorskich i może być dopuszczona do publicznej obrony**.



Prof. dr hab. inż. Jerzy Stefanowski