prof. dr hab. inż. Andrzej Obuchowicz
Uniwersytet Zielonogórski
Instytut Sterowania i Systemów Informatycznych
a.obuchowicz@issi.uz.zgora.pl

Zielona Góra, 24 marca 2026

Recenzja rozprawy doktorskiej

Pana mgr inż. Mateusza Gniewkowskiego

zatytułowanej:

Rigorous Evaluation: Towards Trustworthy Representations in Machine Learning

opracowana na zlecenie

Rady Dyscypliny Naukowej

Informatyka Techniczna i Telekomunikacja

Politechniki Wrocławskiej

1. Problem badawczy i jego znaczenie

Jednym z najistotniejszych wyzwań przed twórcami modeli AI jest weryfikacja ich wiarogodności. Na ile nauczony na skończym zestawie danych model będzie prawidłowo reagował na dane pozyskiwane w rzeczywistych warunkach. W szczególności dotyczy to modeli typu *czarna skrzynka*, w których dane przetwarzane są numerycznie i o silnym stopniu zrównoleglenia, a w tych realiach niezwykle trudno jest uzyskać klarowne wyjaśnienie podjętej decyzji. Praktyka uczy, że często otrzymywane modele decyzyjne, czy predykcyjne, wykazują niski poziom stabilności, reagują na korelacje pozorne, lub na niewielkie zmiany zmiennych semantycznie nieistotnych itp. Próba podniesienia wiarygodności modelu, zwłaszcza stoso-

wanego w zadaniach krytycznych, dotyczących zdrowia i bezpieczeństwa człowieka, jest niezwykle istotna i godna podjęcia. Autor dysertacji proponuje niezwykle ciekawe rozwiązanie w postaci systematycznego iteracyjnego wielokryterialnego audytu jakości modelu, który umożliwia stopniowe udoskonalanie wyuczonych relacji, i stawia na stronie 4 hipotezę, że tego typu postępowanie poprawia solidność, stabilność, wiarygodność i bezpieczeństwo modeli uczenia maszynowego.

2. Wkład autora

Zasadniczym wkładem autora dysertacji jest propozycja procedury weryfikacji niezawodności modelu oraz, w konsekwencji, jego udoskonalenia w postaci iteracji audyt-pomiar-poprawa. W cyklu weryfikacyjnym po jakościowej i ilościowej analizie dowodów wprowadzane jest celowane douczenie reprezentacji bez pełnego przeuczenia całego modelu. Efektywność proponowanego podejścia doktorant wykazywał na dwóch praktycznych realizacjach na danych tekstowytch i graficznych, uzyskując bardziej odporny i łatwiej interpretowalny model. Istotnym elementem proponowanej metodyki w postaci pętli audyt-pomiar-poprawa jest wieloaspektowa analiza stanu modelu z wykorzystaniem zestawu metod audytu i pomiaru jakości reprezentacji, jej odporności i wiarygodności, zawierający opracowane i przystosowane techniki klasteryzacji, miary uogólniania i detekcji odchyleń, czy weryfikacji zgodności dowodów eksplikacji. Ponadto ewaluacja modelu jest wsparta testami antagonistycznymi i metodami ataku sterowanego wyjaśnieniami, tak, aby na etapie poprawy wzmocnić niezawodność systemu uodparniając go na tego typu ataki.

3. Poprawność

Opiniowana praca, zawarta na 134 stronach, składa się z 6 rozdziałów uzupełnionych jednym dodatkiem, zawierającym listę publikacji doktoranta, i bibliografią zawierającą wyczerpujący zbiór 142 publikacji obejmujących stan badań prowadzonych w obszarze zainteresowania pracy.

Motywacja podjęcia tematu pracy, sformułowanie problemu i, na jego podstawie, postawienie tezy pracy zostały zawarte w pierwszym wprowadzającym rozdziale. W rozdziale tym przedstawiono również listę opisującą wkład autora rozprawy

do dziedziny badań, a także opis zawartości następujących kolejnych rozdziałów. Zadaniem rozdziału drugiego jest przybliżenie czytelnikowi podstaw koncepcyjnych i metodyki związanej z uczeniem reprezentacji, zagadnieniami wyjaśnialności w sztucznej inteligencji, wykrywaniem błędów i odpornością modeli. W rozdziale tym wprowadzona została i ujednolicona notacja stosowana w rozprawie, a także zdefiniowane kryteria stosowane w dalszych rodziałach pracy. Wyniki badań doktoranta zostały zawarte w rozdziałach 3, 4 i 5. Proponowany przez autora dysertacji schemat poprawy wiarogodności modelu w postaci audyt-pomiar-poprawa został przedstawiony w rozdziale trzecim jako studium przypadku modelu przetwarzającego tekst, a ściślej na zadaniu klasyfikacji HTTP/URL. Kolejno wprowadzono on schemat bazowy, przeprowadzono systematyczny audyt i ocenę reprezentacji oraz zaproponowano ukierunkowane metody poprawy jakości i odporności reprezentacji. Niezwykle ciekawe rozwiązanie zaproponowano w rozdziale czwartym, w którym wykorzystano ataki adwersyjne jako narzędzie diagnostyczne. Reakcja modelu na powyższe ataki stanowi informację o czułości przesłanek służących do wyjaśnienia decyzji systemu i ich stabilności na z pozoru pomijalne ingerencje. Tego typu analiza zagrożeń jest podstawą do generowania ukierunkowanych interwencji mających na celu poprawę odporności. W rozdziale piątym przedstawiono adaptację wprowadzonej wcześniej metodyki audyt-pomiar-poprawa z uwzględnieniem oceny w warunkach czynników stresujących do klasyfikacji obrazów. Całość pracy została podsumowana w ostatnim szóstym rodziale rozprawy, w którym omówiono najistotniejsze wyniki, jak rónież rekomendacje dotyczące rozwoju wiarygodnych modeli uczenia maszynowego.

Obrana struktura pracy jest właściwa i poprawnie zrealizowana. Na wyróżnienie zasługują dobrze napisane wprowadzenia i podsumowania poszczególnych rozdziałów pracy, dzięki nim nie umyka czytelnikowi główna myśl pracy. Jest to istotne wobec wielowątkowości informacji zawartych w poszczególnych rozdziałach.

Mankamentem doskwierającym czytelnikowi jest brak spisu symboli i skrótów tak licznie stosowanych w tekście pracy. Niestety wiele z tych skrótów w ogóle w pracy nie zostało wyjaśnione, a część wyjaśnień pojawia się po kolejnym (nie pierwszym) pojawieniu się w tekście. Podobnym mankamentem jest wybiórcze wyjaśnianie symboli występujących we wzorach, szczególnie w rozdziale drugim. W ten sposób podany wzór definicyjny nie wnosi żadnej informacji dla czytelnika. Być może

są to standardowe oznaczenia dla głęboko zanurzonych badawczo w obszarze pracy czytelników. Nie mniej dla osób nie będących w głównym nurcie badań, tekst pracy, bogaty w liczne skróty i wzory bez wpełni wyjaśnionych symboli, jest fragmentami nieczytelny i wymaga od czytelnika poszukiwania wyjaśnień poza dysertacją.

Odmienne wrażenia sprawiają fragmenty pracy, w których doktorant prezentuje przebieg i wyniki własnych badań. Są one przejrzyste, a wyciągane wnioski są zasadnie argumentowane. Na wyróżnienie zasługuje również wysoka jakość edytorska pracy.

Jeżeli chodzi o poprawność metodyczną pracy, to należy stwierdzić, że doktorant wybrał niestandardowy model dowodzenia postawionej przez siebie hipotezy, który bazuje na studium dwóch przypadków. Tego typu model jest, w pewnym sensie, wymuszony specyfiką podjętego problemu badawczego, jakkolwiek wnioski z pracy w pełni odnoszą się do klas zadań reprezentowanych przez oba analizowane przypadki, dalsze uogólnianie nie musi być w pełni uzasadnione. Jednakże trudno sobie wyobrazić jak, w tym przypadku, w sposób formalny przedstawić założenia, przesłanki, hipotezę i sam dowód. Zwłaszcza, że rezultatem pracy jest zalgorytmizowana metodyka postępowania. Oceniający tę dysertację sugerowałby raczej zrezygnowanie ze stawiania hipotezy, a raczej skupił się na przedstawieniu celu pracy i wykazaniu, że przedstawione rozwiązanie ten cel realizuje w pełni, badź w zakresie ograniczonym do pewnej klasy modeli uczenia maszynowego.

Wobec powyższego warto postawić sobie pytania:

1. Czy zaproponowana metodyka iteracji audyt-pomiar-poprawa, zastosowana w pracy do modeli klasyfikacyjnych, w sposób bezpośredni może być przeniesiona na inne modele uczenia maszynowago jak predykcyjne, aproksymacyjne itd.?

2. Na ile zespoły metod audytu i pomiaru jakości reprezentacji, które zaproponowano w pracy do analizowanych realizacji na danych tekstowych i graficznych, są wystarczające dla modeli wspomnianych we wcześniejszym pytaniu, albo w zastosowaniu metodyki do innego typu danych, np. sygnałów strumieniowych (akustycznych, radiowych itp.)?

4. Wiedza kandydata

Realizując podjęte w rozprawie doktorskiej zadania doktorant wykazał się głęboką wiedzą z zakresu modeli uczenia maszynowego, w szczególności w obszarze technik reprezentacji danych tekstowych i obrazowych, grupowania danych i detekcja danych odstających, metod generowania wyjaśnień i technik oceny odporności uczneia maszynowego na ataki adwersyjne. Zawartość treści w rozdziale drugim, także w wielu fragmentach późniejszych, dogłębnie i wyczerpująco zapoznaje czytelnika z aktualnym stanem wiedzy dziedzinowej, a także opis przebiegu badań w rozdziałach następnych wskazuje na szeroką wiedzę autora rozprawy w powyższych obszarach i umiejętne z niej korzystanie. Ponadto doktorant wykazał się poprawnością terminologiczną i szeroką znajomością literatury przedmiotu.

5. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach, w tym uwagi krytyczne, głównie natury redakcyjnej i edytorskiej, i wymagania zdefiniowane w art. 187 ustawy *Prawo o szkolnictwie wyższym i nauce* (Dz. U. 2024 r., poz. 1571 z późn. zm.) **stwierdzam, że**: rozprawa doktorska pana mgr inż. Mateusza Gniewkowskiego zawiera oryginalne rozwiązanie problemu naukowego, kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja oraz posiada umiejętność samodzielnego prowadzenia pracy naukowej.

W związku z powyższym **wnoszę** o przyjęcie rozprawy i dopuszczenie jej do publicznej obrony.