dr hab. Jan Kozak, prof. UE
Katedra Uczenia Maszynowego
Wydział Informatyki i Komunikacji
Uniwersytet Ekonomiczny w Katowicach

Katowice, 30 kwietnia 2026

Recenzja rozprawy doktorskiej

mgra inż. Mateusza Gniewkowskiego

Tytuł rozprawy: *Rigorous Evaluation: Towards Trustworthy Representations in Machine Learning*
Autor: mgr inż. Mateusz Gniewkowski
Promotor: dr hab. inż. Henryk Maciejewski
Promotor pomocniczy: dr inż. Tomasz Surmacz
Jednostka: Politechnika Wrocławska, Wydział Elektroniki

1 Tematyka rozprawy

Rozprawa doktorska mgra inż. Mateusza Gniewkowskiego podejmuje zagadnienie rzetelnej ewaluacji modeli uczenia maszynowego z perspektywy ich wiarygodności i odporności w realnych warunkach wdrożenia. Punktem wyjścia jest obserwacja, że dominujący w literaturze paradygmat oceny modeli oparty na agregowanych metrykach klasyfikacyjnych – takich jak dokładność czy F1 – jest niewystarczający do gwarantowania niezawodności systemów działających poza środowiskiem testowym. Praca wpisuje się tym samym w jeden z najważniejszych nurtów współczesnego uczenia maszynowego, dotyczący bezpieczeństwa, interpretowalności i odporności modeli w obliczu przesunięcia

rozkładu danych, ataków adwersaryjnych oraz korelacji pozornych.

Doktorant sformułował centralną hipotezę badawczą zakładającą, że wiarygodność, odporność i bezpieczeństwo modeli uczenia maszynowego można istotnie poprawić poprzez systematyczny, wielokryterialny audyt jakości modelu, umożliwiający przyrostowe doskonalenie wyuczonych reprezentacji. W tym celu zaproponował spójną dyscyplinę badawczą ujętą jako pętla *audyt* $\rightarrow$ *pomiar* $\rightarrow$ *poprawa* (ang. *audit – measure – improve*), która integruje metody wyjaśnialności (XAI), metryki jakości przestrzeni reprezentacji oraz testy poza-dystrybucyjne (OOD) w powtarzalny cykl diagnostyczno-naprawczy.

W ramach realizacji celu badawczego Doktorant podjął następujące działania:

- opracowanie systematycznej procedury audytu i poprawy modeli opartej na analizie jakości reprezentacji i zachowania na danych spoza rozkładu uczącego;
- zaproponowanie portfela metod do oceny jakości i odporności reprezentacji, obejmującego klasteryzację jako wskaźnik korelacji pozornych, detekcję OOD, lokalne i globalne audyty XAI oraz geometryczną analizę przestrzeni osadzeń;
- opracowanie metodyki oceny adwersaryjnej z wykorzystaniem metod wyjaśnialności, określonej jako *weaponized XAI*;
- zaproponowanie ukierunkowanych metod poprawy jakości reprezentacji zarówno dla danych tekstowych (strojenie kontrastowe i destylacja z protokołem *human-in-the-loop*), jak i wizualnych (funkcja straty ukierunkowana na zgodność lokalizacyjną atrybucji z adnotowanymi obiektami);
- empiryczną walidację proponowanego podejścia w studiach przypadków, w tym w zastosowaniu do klasyfikacji ruchu HTTP/URL z użyciem modelu RoBERTa.

Tematyka pracy jest wysoce aktualna i osadzona w kluczowych dyskusjach toczonych w środowisku badań nad AI – zarówno w wymiarze technicznym (bezpieczeństwo i odporność modeli głębokiego uczenia), jak i regulacyjnym (EU AI Act, NIST AI Risk Management Framework). Proponowane podejście wychodzi naprzeciw rosnącemu zapotrzebowaniu na narzędzia systematycznej oceny i dokumentowania zachowania modeli w warunkach dystrybucyjnego przesunięcia.

2 Ocena merytoryczna

Rozprawa podzielona jest na sześć rozdziałów (wliczając wprowadzenie i podsumowanie), uzupełnionych załącznikiem z wykazem osiągnięć naukowych Doktoranta. Struktura jest przejrzysta i logicznie spójna – każdy rozdział kończy się podsumowaniem, które syntetyzuje uzyskane wyniki i wskazuje na ich miejsce w szerszym kontekście pracy.

W rozdziale pierwszym Doktorant precyzyjnie motywuje wybór tematu, odwołując się do praktycznych zagrożeń wynikających ze stosowania modeli niezdolnych do sygnalizowania własnej niepewności lub podatnych na korelacje pozorne. Wart podkreślenia jest staranny aparat definicyjny: Doktorant explicite rozróżnia IID, przesunięcie rozkładu (covariate shift, label shift, concept shift), generalizację OOD, detekcję OOD oraz odporność adwersaryjną, co znacząco ułatwia śledzenie wywodu w kolejnych rozdziałach.

Rozdział drugi zawiera obszerny przegląd literaturowy obejmujący zarówno klasyczne metody (TF-IDF, Word2Vec, SVM, CNN), jak i najnowsze modele transformatorowe (BERT, RoBERTa, generatywne LLM), a także techniki klasteryzacji, detekcji OOD i wyjaśnialności modeli. Aktualność przeglądu – uwzględniającego prace z lat 2022–2024 – stanowi mocny punkt dysertacji i odróżnia ją korzystnie od prac, w których przegląd literatury kończy się na roku 2020. Pewną uwagę można sformułować co do proporcji: obszerność tego rozdziału (ponad 40 stron) zbliża go miejscami do podręcznikowego opisu standardowych technik, zamiast skupiać się wyłącznie na aspektach bezpośrednio istotnych dla późniejszego wywodu.

W rozdziale trzecim Doktorant prezentuje pierwsze studium przypadku – klasyfikację ruchu HTTP/URL opartą na modelu RoBERTa – i ilustruje pełną pętlę audyt–pomiar–poprawa w praktyce. Autor nie ogranicza się do oceny standardowych metryk klasyfikacyjnych, lecz systematycznie analizuje geometrię przestrzeni osadzeń za pomocą klasteryzacji i detekcji OOD, odsłaniając problemy niewidoczne z perspektywy tradycyjnej ewaluacji. Wartościowym wkładem jest tu metoda Sec2Vec, która wykazuje istotną poprawę interpretowalności i odporności modeli wykrywania złośliwego ruchu sieciowego.

Rozdział czwarty rozszerza portfel ewaluacyjny o ocenę adwersaryjną. Doktorant wprowadza koncepcję *weaponized XAI* – celowego wykorzystania metod atrybucji jako narzędzi konstrukcji ataków adwersaryjnych – i weryfikuje ją na klasycznych klasyfikato-

rach tekstowych oraz na zdalnie dostępnych modelach generatywnych (ChatGPT, LLaMA, OpenChat). Pomysł jest oryginalny i interesujący naukowo; Doktorant przeprowadza ponadto ocenę percepcyjnej jakości wygenerowanych próbek z udziałem ludzi, co zwiększa wiarygodność wyników. Warto jednak zaznaczyć, że analiza odporności samych metod XAI na manipulacje (stabilność atrybucji w sensie *explanation robustness*) mogłaby zostać pogłębiona.

W rozdziale piątym Doktorant proponuje lekką, architektonicznie agnostyczną funkcję straty opartą na mierze zgodności lokalizacyjnej (Class Robustness Score, CRS), której celem jest ograniczenie polegania modelu na korelacjach tła i zwiększenie odporności na ataki czarnoskrzynkowe bez degradacji dokładności na danych czystych. Wyniki walidacji na zbiorze ImageNet są przekonujące. Pewnym ograniczeniem jest brak porównania z metodami regularyzacji ukierunkowanej na uwagę modelu (np. podejściami z rodziny *Right for the Right Reasons*), co utrudnia pełną ocenę przewagi zaproponowanego rozwiązania.

Dysertację zamyka rozdział szósty, w którym Doktorant nie tylko syntetyzuje kluczowe wyniki, ale też rzetelnie omawia ograniczenia metodologiczne, kwestie etyczne i kierunki dalszych badań. Dołączony praktyczny przewodnik dla wdrożeń w formie zwięzłej listy kontrolnej (*lifecycle checklist*) stanowi wartościowy dodatek skierowany bezpośrednio do praktyków.

Do mocnych stron dysertacji należy zaliczyć: wyraźną i dobrze uzasadnioną hipotezę badawczą, spójność metodologiczną pętli audyt–pomiar–poprawa realizowanej konsekwentnie przez całą pracę, wysoką aktualność przeglądu literaturowego oraz oryginalność koncepcji *weaponized XAI*. Solidna liczba publikacji w recenzowanych materiałach konferencji ECAI 2023, ICCS 2025 i innych potwierdza dojrzałość naukową Doktoranta.

3 Uwagi i pytania

1. Doktorant proponuje systematyczną pętlę audyt–pomiar–poprawa, jednak kryteria stopowalności tego cyklu nie zostały explicite zdefiniowane. W jaki sposób praktyk powinien stwierdzić, że kolejna iteracja przyniesie już jedynie marginalny zysk? Czy Doktorant widzi możliwość sformalizowania warunków zakończenia procesu?

2. Koncepcja *weaponized XAI* rodzi pytanie o naturę przewagi operacyjnej: czy skuteczność ataku nie wynika po prostu z faktu, że metody atrybucji wskazują te same cechy, które klasyfikator wykorzystuje do podjęcia decyzji – co jest trywialne z perspektywy teorii? Jaką konkretną przewagę daje ukierunkowanie przez XAI w porównaniu z metodami bez wiedzy o modelu, takimi jak TextFooler czy BERT-Attack?

3. Miara CRS definiowana jako odsetek masy atrybucji pokrywającej adnotowany obiekt zależy bezpośrednio od jakości stosowanej metody XAI. Jak Doktorant ocenia wrażliwość proponowanej funkcji straty na wybór metody atrybucji użytej jako sygnał uczący? Czy przeprowadzono eksperymenty ablacyjne porównujące różne metody atrybucji jako podstawę tej funkcji?

4. Protokół *human-in-the-loop* zastosowany w rozdziale 3 wymaga udziału ekspertów domenowych. Jak Doktorant szacuje koszt adnotacji niezbędny do uzyskania mierzalnej poprawy jakości reprezentacji? Czy istnieje możliwość wspomagania eksperta przez metody semi-supervised lub aktywnego uczenia?

5. Eksperymenty w rozdziale 4 obejmują ataki na modele generatywne (ChatGPT, LLaMA, OpenChat) konfigurowane do klasyfikacji zero-shot. Modele te mogą istotnie zmieniać swoje zachowanie w wyniku wewnętrznych aktualizacji ze strony dostawców. Jak Doktorant ocenia wpływ tej niestacjonarności na reprodukowalność prezentowanych wyników?

6. W rozdziale 5 zaproponowana funkcja straty jest testowana przede wszystkim na architekturach splotowych (ResNet). Czy Doktorant przeprowadził lub planuje eksperymenty na modelach Vision Transformer (ViT), w których interpretacja zgodności lokalizacyjnej dla mechanizmów uwagi jest mniej oczywista niż dla map gradientowych?

7. Rozprawa analizuje odporność modeli m.in. na ataki czarnoskrzynkowe. Czy Doktorant zbadał przenaszalność (ang. *transferability*) tych ataków między różnymi architekturami? Informacja ta byłaby istotna dla oceny ogólności zaproponowanych metod poprawy odporności.

8. Proponowany *lifecycle checklist* jest wartościowym uzupełnieniem praktycznym.

Czy Doktorant rozważał jego udostępnienie w formie otwartego narzędzia programistycznego lub biblioteki (podobnie jak publiczne repozytoria kodu towarzyszące publikacjom)? Jakie są plany dotyczące upowszechnienia opracowanych metodyk?

9. Zaproponowana pętla audyt–pomiar–poprawa dotyczy modalności tekstowej i wizualnej. Czy w ocenie Doktoranta daje się ją w naturalny sposób rozszerzyć na dane tabelaryczne lub multimodalne? Jakie modyfikacje portfela metod byłyby w takim przypadku konieczne?

10. Kwestia skalowalności obliczeniowej proponowanych metod audytu na bardzo dużych modelach fundacyjnych (rzędu GPT-4 czy Llama-3 70B) pozostaje w dysertacji otwarta. Czy Doktorant przewiduje istotne bariery w zastosowaniu zaproponowanego podejścia do modeli wielkoskalowych, gdzie pełne obliczenie atrybucji może być prohibitywnie kosztowne?

11. W kontekście regulacyjnym Doktorant odwołuje się do EU AI Act i NIST AI Risk Management Framework. Czy w ocenie Doktoranta zaproponowany *lifecycle checklist* przekłada się bezpośrednio na wymagania formalne stawiane systemom AI wysokiego ryzyka w rozumieniu tych regulacji?

4 Ocena formalna

Rozprawa mgra inż. Mateusza Gniewkowskiego napisana jest w języku angielskim na wysokim poziomie zarówno pod względem merytorycznym, jak i formalnym. Język pracy jest precyzyjny i konsekwentny, terminologia techniczna stosowana poprawnie i spójnie. Godnym naśladowania zwyczajem jest wprowadzenie przez Doktoranta własnych definicji roboczych już w rozdziale pierwszym – pozwala to uniknąć niejednoznaczności przy takich pojęciach jak OOD generalization, OOD detection czy adversarial robustness, które w literaturze bywają stosowane wymiennie.

Wykaz literatury jest szeroki i aktualny – obejmuje zarówno prace fundamentalne, jak i publikacje z lat 2022–2024 dotyczące modeli LLM, wyjaśnialności i regulacji AI.

Znaczna część wyników zaprezentowanych w dysertacji opiera się na pracach opublikowanych wcześniej przez Doktoranta w recenzowanych wydawnictwach (m.in. ECAI 2023, ICCS 2025), co potwierdza oryginalność naukową pracy i jej zakorzenienie w bieżącym dyskursie.

Pewną uwagę można sformułować co do proporcji rozdziału drugiego: jego obszerność sprawia, że miejscami zbliża się do podręcznikowego opisu standardowych technik. Selektywne skrócenie tego rozdziału – z zachowaniem jedynie aspektów bezpośrednio istotnych dla późniejszego wywodu – mogłoby poprawić proporcje dysertacji bez uszczerbku dla jej wartości merytorycznej. Ilustracje użyte w pracy są czytelne i dobrze opisane.

5 Konkluzja

Stwierdzam, że przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie poprawnie zdefiniowanego problemu naukowego. Recenzowana dysertacja mgra inż. Mateusza Gniewkowskiego spełnia wymagania stawiane rozprawom doktorskim przez Ustawę Prawo o szkolnictwie wyższym i nauce, Dz. U. 2023, poz. 742 z późn. zm. Praca dowodzi ogólnej wiedzy teoretycznej i praktycznej Doktoranta w dyscyplinie Informatyka Techniczna i Telekomunikacja oraz potwierdza jego umiejętność samodzielnego prowadzenia badań naukowych.

W związku z powyższym wnioskuję o przyjęcie rozprawy doktorskiej oraz o dopuszczenie mgra inż. Mateusza Gniewkowskiego do publicznej obrony.