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

Katowice, 24 czerwca 2025

Recenzja rozprawy doktorskiej

mgra inż. Piotra Biczyka

Tytuł rozprawy: *Wykrywanie zmian danych wejściowych będących celowymi modyfikacjami ukierunkowanymi na wpływanie na wyniki działania modeli uczenia maszynowego*
Autor: mgr inż. Piotr Biczyk
Promotor: dr hab. Marek Sikora, prof. Pol. Śl.
Jednostka: Politechnika Śląska, Wydział Automatyki, Elektroniki i Informatyki

1 Tematyka rozprawy

Rozprawa doktorska mgra inż. textbfPiotra Biczyka koncentruje się na problematyce wykrywania ukierunkowanych modyfikacji danych wejściowych (ataków adwersaryjnych) w modelach uczenia maszynowego operujących na danych tabelarycznych, co bezpośrednio odwzorowuje zaproponowany temat rozprawy. Zagadnienie to plasuje się na styku trzech dynamicznie rozwijających się obszarów: bezpieczeństwa systemów AI, eksploracji danych oraz inżynierii oprogramowania. Autor trafnie zauważa, że większość istniejących badań w tym zakresie poświęconych jest zagadnieniom ataków adwersaryjnych na modelach przetwarzających obrazy, natomiast modele przetwarzające dane tabelaryczne, dominujące w praktyce biznesowej, są relatywnie słabo chronione.

Kluczowym wyzwaniem, które podejmuje rozprawa, jest identyfikacja prób omijania klasyfikatora w scenariuszu czarnej skrzynki. Oznacza to, że algorytm detekcyjny dysponuje jedynie: oryginalnym zbiorem uczącym; wynikami bieżącej inferencji oraz historią poprawnych zachowań modelu, nie mając wglądu w jego architekturę ani parametry. Takie założenie wiernie odzwierciedla realia usług typu ML–as–a–Service, gdzie użytkownik otrzymuje wyłącznie odpowiedź modelu, bez możliwości introspekcji.

Autor poddaje analizie taksonomię ataków NIST i precyzyjnie umiejscawia własny wkład w kategorii ataków omijania (ang. evasion attacks) naruszających integralność modelu na etapie predykcji. Rozprawa syntetyzuje dorobek ostatnich siedmiu lat badań nad FGSM, PGD, ZOO, HopSkipJump, PermuteAttack i innymi algorytmami generowania perturbacji, pokazując, że dla danych tabelarycznych istnieją specyficzne ograniczenia domenowe (dyskretność, zależności międzypolowe, reguły biznesowe), które radykalnie zmieniają krajobraz zagrożeń w porównaniu z obrazami RGB. Ta diagnoza stanowi punkt wyjścia do projektu nowej metody detekcyjnej.

Główny wkład pracy stanowi dwutorowe podejście zweryfikowane na 22 otwartych zbiorach danych:

1. Ekstrakcja atrybutów diagnostycznych.
   Autor proponuje oryginalny zestaw cech pochodnych opisujących subtelne odchylenia statystyczne pomiędzy rozkładem bieżących przykładów wejściowych a rozszerzoną, wielowymiarową projekcją zbioru uczącego.

2. Algorytm klasyfikujący zdarzenia adwersaryjne.
   W drugim etapie rozważane jest meta-zadanie klasyfikacji binarnej „atak vs. brak ataku".

Pewną nowością koncepcyjną jest wprowadzenie pojęcia uw
walności ataku w przestrzeni tabelarycznej: autor formalizuje
cech binarnych a łatwością przenoszenia perturbacji pomiędzy n
nymi (LR – XGBoost – SVM). Ta obserwacja otwiera perspekty
automatycznym twardnieniem modeli na etapie treningu.

Podsumowując, tematyka rozprawy jest aktualna, inieco interdyscyplinarna i społecznie istotna. Praca łączy solidną podbudowę teoretyczną z praktycznym ukierunkowaniem

na wdrożenia i otwiera perspektywy dla:

- podniesienia transparentności oraz bezpieczeństwa modeli ML stosowanych w obszarach regulowanych,
- budowy platform monitorujących odporność klasyfikatorów na ataki w czasie rzeczywistym,
- rozszerzenia badań nad *explainable AI* w kontekście detekcji anomalii adwersaryjnych.

2 Ocena merytoryczna

Rozprawa została podzielona na pięć zasadniczych rozdziałów – od omówienia celów badawczych, przez analizę literatury i propozycję własnej metody, aż po bogaty korpus eksperymentów i podsumowanie wniosków. Taki układ jest klarowny, prowadzi czytelnika krok po kroku i spełnia standardy monografii naukowej.

W rozdziale 1 Autor przedstawia tło badań, motywację i konkretne cele. Tekst prowadzi od implementacj wybranych ataków, przez konstrukcję detektora działającego w trybie czarnej skrzynki, aż po weryfikację skuteczności na zróżnicowanych zbiorach danych. Lista celów głównych i szczegółowych ułatwia weryfikację, czy zostały później osiągnięte. Na pochwałę zasługuje umieszczenie pracy w kontekście doktoratu wdrożeniowego oraz wskazanie praktycznej współpracy z przemysłem, co czyni badania realnie użytecznymi. Brak jednak przedstawienia tezy pracy, jaką Autor planuje zweryfikować.

Rozdział 2 zawiera przegląd literatury, definicji przykładów adwersaryjnych i funkcji odległości, taksonomię ataków wg NIST, aż po szczegółowe omówienie technik FGSM, PGD, ZOO czy HopSkipJump. Autor omawia zarówno scenariusz białej, jak i czarnej skrzynki, wskazując braki w aktualnym stanie wiedzy dla danych tabelarycznych. Wywód jest logiczny, poparty odniesieniami do publikacji i stanowi mocne uzasadnienie dla podjęcia dalszych badań.

Jedną z najważniejszych części dysertacji jest rozdział 3. Autor wprowadza dwuetapową koncepcję: (1) ekstrakcję atrybutów diagnostycznych opisujących lokalne i globalne odchylenia statystyczne oraz (2) klasyfikator „atak / brak ataku". Szczegółowo opisano

fazę monitoringu, sposób budowy modelu zastępczego oraz zestaw piętnastu miar diagnostycznych, w tym innowacyjne wskaźniki zgodności predykcji i celów w sąsiedztwie.

W rozdziale 4 poświęcony jest eksperymentom i uzyskanym rezultatom. Dobór zbiorów danych użytych w badaniach zapewnia dużą zmienność liczby cech (5–1776) i rozmiaru (569–34465 przykładów), co pozwala ocenić skalowalność rozwiązania. W analizie wyników Autor omawia skuteczność detekcji (F_1 i AUC), wpływ intensywności ataku, ważność poszczególnych atrybutów diagnostycznych oraz ograniczenia w wykrywaniu typów ataku.

Ostatni rozdział jest bardzo szerokim podsumowaniem uzyskanych wyników i odniesieniem do podanyych wcześniej celów pracy. Autor wskazuje również ograniczenia oraz proponuje dwa kierunki dalszych badań.

Struktura rozprawy jest spójna, a kolejne części ściśle wynikają z postawionych celów badawczych. Język pracy jest klarowny, terminologia konsekwentna, jednak rysunki mogłyby być nieco czytelniejsze (w niektórych przypadkach), warto też rozróżnić tabele od rysunków. Z poważniejszych zastrzeżeń zwróciłbym uwagę jedynie na brak pełnego porównania z metodami semi-white-box oraz stosunkowo krótką refleksję nad implikacjami etycznymi.

3 Uwagi i pytania

1. W rozdziale 1 przedstawiono cele badawcze, ale nie zidentyfikowano wprost głównej tezy ani hipotez, które miałyby zostać zweryfikowane w ramach rozprawy.

2. We wstępie warto wyraźnie odróżnić cele ogólne od szczegółowych pytań badawczych, aby czytelnik miał jasność, jak poszczególne rozdziały odpowiadają na postawione problemy.

3. Sekcja „Pytania badawcze" mogłaby zostać rozbudowana o krótki opis, w jaki sposób każde pytanie prowadzi do weryfikacji konkretnej hipotezy.

4. Autor w kilku miejscach (np. na rys. 2.2, rys. 4.1) wstawił do dokumetu przygotowanym w LaTeX zrzut ekranu tabeli z jednego z klasycznych edytorów tekstów (choć

w sumie, jednak nie rozumiem dlaczego), w mojej ocenie nie powinien nazywać tego rysunkiem, a najlepiej skorzystać ze środowiska `table`.

5. Część ilustracji (np. rys. 2.1) jest mało czytelna. Może warto byłoby w takim przypadku przygotować wyraźny rysunek na podstawie źródeł (podając na podstawie jakich materiałów został przygotowany) zamiast dodawać bardzo nieczytelny, oryginalny rysunek.

6. Wzory matematyczne pojawiają się bez konsekwentnej numeracji. Ponumerowanie wszystkich równań ułatwiłoby odwoływanie się do nich w tekście.

7. W opisie metodyki doboru hiperparametrów podano jedynie zakresy wyszukiwania. Brakuje informacji, czy przeprowadzono strojenie (np. grid search albo random search) i jakie wartości ostatecznie przyjęto.

8. W analizie wyników zastosowano głównie miary F_1 i AUC. Czy są one najwłaściwsze dla badanych problemów (np. przy silnej niezrównoważonej klasie)? Nie podano w pracy powodu wybrania tych dwóch miar, a może warto było rozważyć również także precyzje i czułość lub wprost podać macierze błędów.

9. Co prawda w rozdziale 4 Autor wspomina o zrównoważnoej dokładności klasyfikacji (nazywając ją „balanced accuracy"), ale jednak w innym kontekście i też bez szereszego wyjaśnienia możliwości wykorzystanie tej miary zależnie od konkretnego przypadku.

10. Liczba reduktów (~ 2500 przy $\epsilon = 0{,}05$) sugeruje znaczne zapotrzebowanie na moc obliczeniową. Czy oszacowano czas wykonania i zużycie pamięci oraz ich wpływ na możliwość wdrożenia rozwiązania produkcyjnego?

11. Kod źródłowy i skrypty eksperymentalne nie zostały udostępnione. Publiczne repozytorium (np. Git) ułatwiłoby replikację wyników i ocenę poprawności zaproponowanych rozwiązań.

4 Ocena formalna

Rozprawa doktorska mgra inż. Piotra Biczyka reprezentuje wysoki poziom zarówno pod względem merytorycznym, jak i formalnym. Tekst prowadzony jest klarownym, precyzyjnym stylem naukowym. Autor wyraźnie formułuje cel główny oraz cele szczegółowe, choć nie wspomina o tezie pracy.

Wyniki eksperymentów są odtwarzalne: Autor precyzyjnie opisuje środowisko obliczeniowe i narzędzia (m.in. biblioteka BrightBox oraz Adversarial Robustness Toolkit), wskazując konkretne wersje oprogramowania i publiczne repozytoria wykorzystywanych komponentów. Takie podejście podnosi wiarygodność rezultatów oraz ułatwia potencjalną weryfikację przez innych badaczy.

Bibliografia nie jest bardzo dobra, ale wystarczająca. Liczy 84 pozycje i obejmuje zarówno klasyczne prace (np. Pawlak 1991, Breiman 2001 – choć z oczywistych względów to nie jest bezpośrednio związane z tematem rozprawy), jak i najświeższe artykuły z lat 2023–2025, tu już przede wszystkim bezpośrednio związane z pracą i dotyczące ataków adwersaryjnych na dane tabelaryczne (m.in. He et al. 2025, Yin et al. 2024). Pokazuje to dobrą orientację autora w aktualnym stanie wiedzy.

Pod względem edytorskim rozprawa jest przygotowana poprawnie, pomimu kilku mankamnetów skład pracy jest przejrzysty. Poprawy wymaga jedynie czytelność tabel (a w zasadzie przygotowanie ich zamiast zrzutów ekranu) i niektórych rysunków.

5 Konkluzja

Stwierdzam, że przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie poprawnie zdefiniowanego problemu naukowego. Recenzowana dysertacja mgra inż. Piotra Biczyka 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 prowadzenia badań naukowych.

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