Dr hab. inż. Roman Stanisław Deniziak, prof. PŚk
Katedra Systemów Informatycznych
Politechnika Świętokrzyska
ul. 1000-lecia Państwa Polskiego 7
25-314 Kielce

Kielce, 23.06.2025

# Recenzja rozprawy doktorskiej dla Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej

**Tytuł rozprawy:** Wykrywanie zmian danych wejściowych będących celowymi modyfikacjami ukierunkowanymi na wyniki działania modeli uczenia maszynowego

**Autor rozprawy:** mgr inż. Piotr Biczyk

**Promotor rozprawy:** dr hab. inż. Marek Sikora, prof.

## 1. Cel, zakres i charakter rozprawy

Tematyka pracy doktorskiej dotyczy zagadnień z zakresu uczenia maszynowego w kontekście bezpieczeństwa systemów informatycznych. Autor skupił się na problemach ataków adwersaryjnych polegających na zakłócaniu danych wejściowych modeli uczenia maszynowego w taki sposób, aby zakłócić działanie modelu. Przy czym badania zostały ograniczone do systemów klasyfikacji danych tabelarycznych.

Cel pracy został przez Autora określony jako: „stworzenie, zweryfikowanie i implementacja metody wykrywania ataków adwersaryjnych na modele ML trenowane na danych tabelarycznych i pełniących rolę systemów klasyfikacyjnych". Cel miał być osiągnięty przy założeniu, że metoda ma zakładać brak wiedzy o architekturze i parametrach modelu (czarna skrzynka). Uzyskanie efektu końcowego Autor zaplanował poprzez realizację następujących celów szczegółowych:

1. Opracowanie metod ataku modeli ML klasyfikujących dane tabelaryczne.
2. Opracowanie metody aproksymującej działanie modelu ML.
3. Opracowanie i weryfikacja metody wykrywania ataków.
4. Weryfikacja działania opracowanej metody na przykładach ataków adwersaryjnych.

Realizacja tych prac stanowiła jednocześnie kolejne etapy metodyki badawczej Doktoranta.

W pracy nie zdefiniowano tezy pracy, zatem praca ma głównie charakter projektowy, co zostało określone w opisie celu pracy. Chociaż Autor jako element pracy w p. 3.2.1 określił hipotezę badawczą mówiącą, że „analiza atrybutów diagnostycznych może wskazać istotne prawdopodobieństwo wystąpienia modyfikacji danych wejściowych noszącej znamiona

modyfikacji celowej". Hipoteza ta została potwierdzona eksperymentalnie, zatem można stwierdzić, że praca obejmowała też zagadnienia naukowe.

## 2. Zawartość i strona redakcyjna rozprawy

Rozprawa doktorska składa się z 5 rozdziałów i wykazu bibliografii. Pierwszy rozdział stanowi wstęp do tematyki pracy i zawiera motywację oraz cele pracy. W rozdziale 2 Autor przedstawił problematykę ataków adwersaryjnych, opisał typy oraz metody takich ataków a także metody wykrywania takich ataków. Bardziej szczegółowo przedstawił zagadnienia ataków adwersaryjnych na modele przetwarzające dane tabelaryczne. Rozdział 3 zawiera opis proponowanej metody wykrywania ataków. Autor opisał zasady tworzenia modeli zastępczych, przedstawił proponowane atrybuty diagnostyczne i metodę tworzenia klasyfikatorów rozpoznających ataki. Opis wykonanych eksperymentów przedstawiono w rozdziale 4, stanowiącym najobszerniejszą część pracy. Oprócz wykazania skuteczności opracowanej metody Autor wykonał eksperymenty statystyczne pokazujące różne własności proponowanych rozwiązań. Na zakończenie przedstawiono analizę i dyskusję dotyczącą uzyskanych wyników. W rozdziale 5 zawarto podsumowanie pracy i kierunki dalszego rozwoju badań.

Zastosowany układ pracy jest logiczny i kolejność przedstawianych zagadnień jest poprawna. Jednak ze względu na dużą liczbę wzorów i symboli przydałby się wykaz ważniejszych oznaczeń. Przydatne byłyby także spisy tabel i rysunków. Do wielu rysunków Autor nie odwołuje się w tekście. Brakuje również numeracji równań, być może wtedy Autor uniknąłby powtórzeń tych samych równań (str. 40-41 i str. 50). Również Autor niepotrzebnie powtarza pewne opisy (np. na stronie 81). Czytelność pracy byłaby większa, gdyby pewne opisy takie jak wyjaśnienia kluczowych pojęć i algorytmy były przedstawione w sposób bardziej formalny np. w postaci numerowanych definicji i pseudokodów.

## 3. Analiza źródeł

Bibliografia przedstawiona w pracy zawiera 84 pozycji, obejmujących głównie artykuły w czasopismach zagranicznych i referaty z konferencji międzynarodowych. Dużą część bibliografii stanowią tzw. preprinty. W znacznej liczbie pozycji brakuje pełnych informacji o publikacji, zwykle brakuje tytułu wydawnictwa, co utrudnia identyfikację wskazanej publikacji. Bibliografia jest stosunkowo nowa ponad 25% pozycji jest z ostatnich 5 lat a ponad 80% pozycji jest z ostatnich 10 lat.

Przegląd literatury dotyczącej tematyki pracy jest przedstawiony w rozdziale 2. Autor przedstawia przegląd literatury z zakresu ataków na modele ML jak i na metody wykrywania tych ataków. Pominięte zostały zagadnienia związane z obroną modeli ML przed atakami adwersaryjnymi, wydaje się, że tematyka ta jest bardzo bliska tematyce pracy i powinna być również omówiona w rozprawie. W p. 2.5 gdzie Autor skupia się na klasyfikatorach przetwarzających dane tabelaryczne, przedstawił ataki na te modele, ale nie przedstawił żadnej metody wykrywania takich ataków. Ponieważ jest to główna tematyka pracy warto było by wskazać jak sobie do tej pory radzono z takimi atakami.

## 4. Metodyka badań

Punktem wyjścia do badań przeprowadzanych w ramach pracy była analiza istniejących rozwiązań w zakresie wykrywania ataków adwersaryjnych na modele ML. Autor zauważył, że istniejące rozwiązania koncentrują się głównie na modelach przetwarzających obrazy a także że brakuje wystarczająco skutecznych metod traktujących model ML jako "czarną skrzynkę". W zasadzie jest to wystarczająca motywacja do podjęcia tego kierunku badań, jednak warto byłoby jeszcze przeanalizować jaka jest skala tego problemu i czy skoro niektóre metody ataku można dostosować do danych tabelarycznych to też czy metody wykrywania nie można by dostosować do tego typu ataków.

W kolejnym kroku Autor zaproponował rozwiązanie polegające na utworzeniu modelu zastępczego dla modelu ML, zaproponowaniu zestawu atrybutów diagnostycznych oraz opracowaniu klasyfikatora wykrywającego atak. Proponowana procedura diagnostyczna dla tak opracowanego rozwiązania składa się z 3 kroków: utworzenie modelu zastępczego na podstawie obserwacji działania rzeczywistego modelu dla poprawnych danych (faza wstępna), równoległa obserwacja modelu ML i modelu zastępczego dla badanych danych wejściowych, oraz klasyfikacja wartości atrybutów diagnostycznych pod kątem wykrywania celowych zmian danych (faza monitorowania), aktualizacja klasyfikatora na podstawie wiedzy o nowych atakach i błędnych klasyfikacjach (faza aktualizacji klasyfikatora). Mimo, że proponowane rozwiązanie wydaje się sensowne to brakuje uzasadnienia, dlaczego akurat taki sposób postępowania został wybrany przez Autora.

W ostatnim kroku Autor wykonał szereg eksperymentów, które po pierwsze miały potwierdzić skuteczność zaproponowanej metody, a po drugie dostarczyć szczegółowych informacji o specyfice działania poszczególnych elementów rozwiązania. Informacje te mogą służyć do lepszego poznania zasady działania metody a jednocześnie mogą służyć dalszemu rozwojowi tego kierunku badań.

W ramach eksperymentów, najpierw Autor przygotował 3 przykładowe metody uczenia maszynowego. Następnie wykonał serię ataków na te modele wykorzystując standardowe zbiory danych i zaimplementowane przez siebie 3 metody ataków. Dla wszystkich testów zostały wyznaczone wartości atrybutów diagnostycznych. W kolejnym kroku Autor wyuczył klasyfikator do wykrywania ataków i wykonał eksperymenty wykrywania ataków. Na zakończenie zostały wykonane również eksperymenty dodatkowe, oceniające skuteczność dla dodatkowych typów ataków i modelu uczenia maszynowego bazującego na sieci neuronowej.

Na zakończenie Autor wykonał analizę uzyskanych wyników, mającą na celu uzyskanie wniosków dotyczących skuteczności oraz wad i zalet opracowanej metody.

## 5. Oryginalność uzyskanych wyników

Praca ma głównie charakter projektowy, zatem głównym wynikiem pracy jest opracowana metoda wykrywania ataków na modele ML. Metoda ta jest oryginalnym rozwiązaniem Autora i spełnia założenia określone w celu pracy. Uzyskane wyniki potwierdziły jakość opracowanej metody i wykazały słuszność zaproponowanego podejścia.

Rozprawa doktorska zawiera też inne oryginalne wyniki, które mogą być wykorzystywane w pracach badawczych z zakresu tematyki pracy. Takimi wynikami są:

1. Model zastępczy mechanizmu uczenia maszynowego zbudowany ze zbioru przybliżonych reduktów.
2. Zbiór atrybutów diagnostycznych umożliwiających ocenę działania modelu uczenia maszynowego.
3. Liczne eksperymenty pokazujące różne własności zaproponowanych rozwiązań.

Wyżej wymienione wyniki prac Autora mogą być wykorzystywane do dalszych badań nad atakami adwersaryjnymi i wykrywaniem tych ataków.

W odniesieniu do aktualnego stanu wiedzy wyniki badań uzyskane przez Autora są innowacyjne. Wyżej wymienione osiągnięcia zostały wdrożone w firmie QED Software.

**6. Uwagi krytyczne, wady i słabe strony rozprawy**

Oprócz wcześniej przedstawionych uwag dotyczących układu treści pracy, bibliografii i metodyki badań, wydaje się, że niektóre zagadnienia powinny być dokładniej przedyskutowane, wymagają wyjaśnienia lub uzupełnienia. Dotyczy to następujących problemów:

1. Praca zawiera obszerną część eksperymentalną, w której Autor wykazuje efektywność opracowanej metody. Niestety nie przedstawia żadnego porównania do istniejących metod wykrywania ataków adwersaryjnych. Takie porównanie byłoby przydatne chociażby z punktu widzenia oceny innowacyjności proponowanego podejścia.
2. Autor w swoich badaniach skupił się na modelach uczenia maszynowego w zastosowaniach klasyfikacji danych tabelarycznych. W pracy brakuje informacji jak szeroką klasą ataków są ataki na modele tego typu. Powstają pytania: jak często pojawiają się takie ataki i jak sobie radzą z tymi atakami istniejące metody? Odpowiedzi na te pytania są istotne z punktu widzenia znaczenia opracowanych metod.
3. Oprócz metod wykrywania ataków na modele ML opracowywane są również metody uodpornienia (obrony) modeli na te ataki. Autor nie omawia szerzej tego zagadnienia, ale powstaje pytanie czy uodpornianie metod ML na ataki adwersaryjne utrudnia wykrywanie takich ataków? Czy można łączyć ze sobą te dwa podejścia czy są to podejścia alternatywne?
4. Czy skuteczność wykrywania ataku zależy od jakości atakowanego klasyfikatora?
5. W celu pracy nie określono wymagań jakościowych dotyczących projektowanej metody. Zatem powstaje pytanie jak się ma uzyskany wynik oceny skuteczności wykrywania ataków do oczekiwań? Jaka jest skuteczność innych istniejących metod?
6. W p. 4.2.2 porównano skuteczność ataków na 3 modele uczenia maszynowego, Autor wskazał tu, że najbardziej odpornym modelem jest XBoost. Na jakiej podstawie pojawił się taki wniosek, bo na przedstawionym rysunku (Rys. 4.2) skutki ataków dla wszystkich metod są podobne.

Ponadto rozprawa zawiera wiele drobnych błędów redakcyjnych i językowych. Szczegółowe uwagi:

1. Str. 8 – „modelei" => „modeli".
2. Str. 13 – w równaniu nie określono znaczenia wartości *d*.
3. Str. 15 – „adwersaryjnee" => „adwersaryjne".
4. Str. 16 – „prawdopodibneństwo" => „prawdopodobieństwo"
5. Str. 18 – „zależeń" => „zależeć".
6. Str. 19 – „modułu" => „moduły".
7. Str. 24 – „grupa metoda" => „grupa metod".
8. Str. 25 – „zastosowaniach aplikacjach" => ?
9. Str. 29 – „stosowany" => „stosowane",
10. Str. 31 – „stabilności" => „stabilność",
11. Str. 34 – „heurestycznych" => „heurystycznych",
12. Str. 35 – „atakowania" => „atakowaniu",
13. Str. 36 – „zastępczych zastępczych" => „zastępczych",
14. Str. 37 – „niewpewnymi" => „niepewnymi",
15. Str. 37-38 – „zgodność i zdolności" => ?
16. Str. 38 – „to z jej" => „to jej",
17. Str. 40 – „liczbą" => „liczba"
18. Str. 46 – „normalnych działających" => „normalnie działających"?
19. Str. 46 – „Wejściem do tych klasyfikatorów są wartościach atrybutów diagnostycznych oraz etykietach decyzyjnych …",
20. Str. 48 – „zastosowania w których" => „zastosowania, w których",
21. Str. 49 – „przez model wejściowy przez model zastępczy"?
22. Str. 50 – wzory są powtórzone ze stron 40-41,
23. Str. 71 – „FGW" => „FGM",
24. Str. 74 – „Pytnom" => „Phyton",
25. Str. 75 – „stosowaną praktyką w literaturze", raczej „stosowaną praktyką przedstawioną w literaturze",
26. Str. 78 – „przesunięcie" => „przesunięcia",
27. Str. 80 – „.." => „."
28. Str.80 – „obliczono i na tej podstawie obliczono"?
29. Str. 82 – „juz" => „już",
30. Str. 104 – „problematyczna" => „problematyczne,
31. Str. 113 – „przekrywania" => „przykrywania",
32. Str. 117 – „o ani dostępu" => „ani bez dostępu".

Wyżej wymienione uwagi nie podważają wartości uzyskanych wyników badań i poprawności opracowanych rozwiązań. Ale omówienie tych zagadnień pozwoliłoby na lepszą ocenę uzyskanych wyników badań, lepsze zrozumienie znaczenia uzyskanych wyników i podkreśliłoby innowacyjność uzyskanych wyników. Natomiast usunięcie błędów redakcyjnych i językowych zwiększyłoby czytelność i jakość pracy.

## 7. Znaczenie uzyskanych wyników

Głównym wynikiem pracy jest opracowanie nowej metody wykrywania ataków adwersaryjnych na modele ML przetwarzające dane tabelaryczne. Wykonane eksperymenty potwierdziły wysoką skuteczność metody w wykrywaniu takich ataków, co jednocześnie potwierdza praktyczną przydatność opracowanej metody w systemach bezpieczeństwa dedykowanych do rozwiązań wykorzystujących uczenie maszynowe. Znaczenie uzyskanych wyników jest nieco zawężone przez ograniczenia opracowanej metody, jednak Autor wskazał kierunki dalszych prac, które mogą poszerzyć zakres zastosowań tej metody.

Obecnie następuje szybki rozwój metod sztucznej inteligencji bazujących na różnych modelach uczenia maszynowego. Jednocześnie rozwija się cyberprzestępczość również w zakresie ataków na systemy oparte na uczeniu maszynowym. Konieczne jest zatem tworzenie coraz bardziej skutecznych i wydajnych mechanizmów ochrony. Jednym z istotnych elementów takich zabezpieczeń są narzędzia do wykrywania ataków. W tym kontekście uzyskane wyniki mają ogromne znaczenie praktyczne, szczególnie że opracowane rozwiązane jest gotowe do wdrożenia w rzeczywistych systemach.

## 8. Ocena końcowa rozprawy

Podsumowując, uważam, że rozprawa doktorska mgra inż. Piotra Biczyka przedstawia oryginalne rozwiązanie zaprezentowanego w niej zagadnienia naukowego i konstrukcyjnego. Autor podjął w niej problem, który ma istotne znaczenie w dziedzinie nauk inżynieryjno-technicznych w zakresie dyscypliny naukowej Informatyka Techniczna i Telekomunikacja. Trafnie określił założenia dotyczące jego analizy i uzyskane wyniki potwierdził wykonanymi eksperymentami. Wykazał się dobrą znajomością ogólnej wiedzy teoretycznej i praktycznej z zakresu tematyki pracy, a także umiejętnością samodzielnego prowadzenia pracy naukowej. Stwierdzam, że recenzowana praca spełnia wymogi stawiane rozprawom doktorskim w Ustawie z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce (Dz.U. poz. 1668, art. 187) i niniejszym wnoszę o dopuszczenie jej do publicznej obrony.