RDITT – hpi. 11.06.2025 r.

Kraków, dnia 26 maja 2025 roku

Dr hab. inż. Tomasz Hachaj
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej
Katedra Informatyki Stosowanej
al. A. Mickiewicza 30, Pawilon C-2
30-059 Kraków

Recenzja Rozprawy Doktorskiej mgr inż. Piotra Biczyka pt. „Wykrywanie zmian danych wejściowych będących celowymi modyfikacjami ukierunkowanymi na wpływanie na wyniki działania modeli uczenia maszynowego”

Promotor: dr hab. Marek Sikora
Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki

Niniejsza recenzja została opracowana w oparciu o uchwałę Rady Dyscypliny Informatyka Techniczna i Telekomunikacja nr 107/2025 z dnia 25 marca 2025 podpisanej przez prof. dr hab. inż. Andrzeja Polańskiego, Przewodniczącego niniejszej Rady; która to (uchwała) wyznacza mnie na Recenzenta. Otrzymałem pocztą zawiadomienie w dniu 08 maja 2025 informujące mnie o podjęciu powyższej uchwały.

1. Ocena wyboru tematy i tez rozprawy

W pracy postawiono i starano się wykazać prawdziwość hipotezy badawczej, że możliwe jest:
(strona 8)
„(...) stworzenie, zweryfikowanie i implementacja metody wykrywania ataków adwersaryjnych na modele ML (machine learning – przypisek recenzenta) trenowane na danych tabelarycznych i pełniące rolę systemów klasyfikacyjnych. W celu zapewnienia uniwersalności, metoda ma działać w reżimie czarnej skrzynki tzn. bez dostępu do modelu poddawanego diagnostyce, w szczególności bez dostępu do jego architektury oraz parametrów. Działanie metody powinno opierać się na obserwacji decyzji podejmowanych przez diagnozowany model. Przy czym zakładamy, że metoda diagnostyczna ma dostęp do wyników działania diagnozowanego modelu na niezaburzonym, pozbawionym przykładów adwersaryjnych, zbiorze przykładów na których model ten był trenowany i walidowany”.

(strona 46)

„W zakresie niniejszej pracy leżała weryfikacja hipotezy mówiącej że analiza atrybutów diagnostycznych może wskazać istotne prawdopodobieństwo wystąpienia modyfikacji danych wejściowych noszącej znamiona modyfikacji celowej (ataku adwersaryjnego).”

Autor Rozprawy doprecyzował poszczególne elementy tezy wykorzystując popularne miary perturbacji, skuteczności ataku oraz atrybutów diagnostycznych, których ścisłą definicję zawarł w drugim i trzecim rozdziale Dysertacji. W celu wykazania prawdziwości tezy pracy, zrealizowano szereg celów szczegółowych, których opis stanowi treść kolejnych rozdziałów Pracy.

Hachaj

Praca została wykonana w ramach doktoratu wdrożeniowego realizowanego we współpracy z firmą QED Software sp. z o.o., która specjalizuje się m.in. w rozwoju technik wyjaśnialności modeli ML. Autor pracy deklaruje (strona 9), że wyniki niniejszej pracy zostaną wykorzystane w działaniach QED Software związanych z rozwojem platformy służącej do monitorowania i ciągłego audytu modeli ML.

Tematyka recenzowanej Rozprawy dobrze lokuje się w obszarze zainteresowań współczesnej informatyki, w szczególności w zakresie zaawansowanych zagadnień podatności model ML na ataki adwersaryjne oraz wykrywanie tego typu szkodliwych aktywności. **Wybór problemu, tezy rozprawy oraz jej celów oceniam pozytywnie.**

2. Ocena zawartości rozprawy

Przedstawiona do recenzji Praca składa się ze 134 stron maszynopisu. Zawiera spis treści, pięć numerowanych rozdziałów, spis pozycji literatury, który zawiera 84 pozycje. Całość rozprawy napisana jest w języku polskim.

Układ i zawartość rozdziałów merytorycznych jest zasadniczo poprawna. W pierwszym rozdziale Autor zarysowuje tematykę doktoratu oraz ogólnie przedstawia podjętą problematykę badawczą. Przedstawia tu również hipotezę badawczą rozprawy, przyczyny, dla których Doktorant zainteresował się tą tematyką, informuje o współpracy z partnerem biznesowym. Przedstawiony jest również układ pracy.

W drugim rozdziale przedstawiono teorię i aktualny stan wiedzy (przegląd literatury) dotyczące ataków adwersaryjnych na modele uczenia maszynowego w tym taksonomię tego typu ataków, metody ich wykrywania. Zwrócono szczególną uwagę na modele przetwarzające dane tabelaryczne, które są przedmiotem zainteresowania niniejszej rozprawy.

W trzecim rozdziale Doktorant opisał proponowaną metodę wykrywania ataków adwersaryjnych, którą stworzył w celu udowodnienia hipotezy badawczej. Dyskutuje tu zastosowanie modeli zastępczych, rolę atrybutów diagnostycznych, konstrukcję klasyfikatorów wykrywających potencjalne ataki oraz przebieg monitoringu modeli uczenia maszynowego. Modele zastępcze utworzone zostały za pomocą aproksymacji oryginalnych modeli przy pomocy metod z obszaru zbiorów przybliżonych. Operacja ta wymagała dyskretyzacji atrybutów, konstrukcji przybliżonych reduktorów (które aproksymowały oryginalny model), zgodnie z algorytmem zaimplementowanym w pakiecie RoughSets [65], którego koncepcja jest omówiona w sekcji 3.1.2 Dysertacji. Użyte przez Autora atrybuty diagnostyczne (patrz sekcja 3.2) to rozmiar sąsiedztwa, niepewność oraz szereg metryk spójności. Atrybuty te służą do oceny spójności predykcji modelu oraz pomagają w zrozumieniu indywidualnych błędnych klasyfikacji. Autor definiuje również klasyfikatory, których rolą jest rozpoznanie faktu ataku na model, używa do tego algorytmów losowych lasów oraz XGBoost. Przygotowanie oraz przebieg eksperymentu symulującego cykl diagnostyczny (monitorowanie modeli) opisany jest w sekcjach 3.3 oraz 3.4.

W rozdziale czwartym Doktorant przedstawia i omawia przebieg eksperymentów mających dowieść skuteczności zaproponowanego w trzecim rozdziale podejścia. Opisane tu prace badawcze obejmowały: przygotowanie modeli uczenia maszynowego działających na publicznie dostępnych zbiorach danych; implementację wybranych metod ataków; analizę z wykorzystaniem modeli zastępczych sposobu działania modeli uczenia; analizę statystyczną wartości atrybutów diagnostycznych w celu stwierdzenia czy występują istotne różnice pomiędzy stanami z atakiem oraz

bez ataku; stworzenie klasyfikatora umożliwiającego wykrywanie stanu podejrzenia o atak. Autor wyselekcjonował 22 publicznie dostępne tabelaryczne zbiory danych, przedstawiono je w Tabeli 4.1. Wybór tych konkretnych zbiorów w wyczerpujący sposób umotywował w sekcji 4.1.1 pracy. Autor szczegółowo i rzetelnie omawia poszczególne etapy badania przedstawiając wyniki w postaci tabelarycznej oraz na wykresach. Przedstawione wyniki dowodzą skuteczności zaproponowanego w pracy podejścia.

Ostatni piąty rozdział jest podsumowaniem całości Rozprawy. Autor uzasadnia tu spełnienie zakładanej hipotezy badawczej. Informuje również, że jego praca badawcza posłużyła firmie QED Software do stworzenia biblioteki BrightBox [41], która znalazła zastosowanie przy wdrażaniu monitoringu oraz audytu modeli uczenia maszynowego. Proponuje również dalsze kierunki badawcze będące rozszerzeniem tematyki tej Dystertacji.

Warto podkreślić, że praca [41] „BrightBox — A rough set based technology for diagnosing mistakes of machine learning models", której Doktorant jest współautorem ukazała się w renomowanym czasopiśmie Applied Soft Computing.

3. Oryginalne osiągnięcia autora rozprawy

W recenzowanej pracy zawarto kilka wartościowych i oryginalnych koncepcji oraz rozwiązań, dokonano ich implementacji oraz uzyskano wyniki wzbogacające naszą wiedzę w zakresie omawianych zagadnień. Do najważniejszych osiągnięć autora należy zaliczyć:

1. Opracowanie, analizę, weryfikację oraz implantację przedstawionych w trzecim rozdziale algorytmów.
2. Rozbudowaną analizę statystyczną o charakterze porównawczym zaproponowanych metod w oparciu o szeroki zestaw danych.
3. Dbałość o rzetelność i powtarzalność przeprowadzonych badań poprzez staranny dobór i wykorzystaniu powszechnie dostępnych zbiorów danych.

Wymienione osiągnięcia są oryginalne i znaczące. Na nich zatem przede wszystkim opieram ogólnie pozytywną ocenę rozprawy. Udowodniają one postawioną tezę rozprawy oraz stanowią wkład w rozwój algorytmów pozwalających na monitorowanie oraz wykrywanie zmian danych wejściowych ukierunkowanych na wpływanie na wyniki działania modeli uczenia maszynowego.

4. Uwagi dyskusyjne i krytyczne

Jak już wcześniej wspomniałem, przedłożona do recenzji rozprawa pod względem merytorycznym i redakcyjnym napisana jest w większości poprawnie. W mojej ocenie Autor nie ustrzegła się jednak pewnych braków i nieścisłości. Moje uwagi dyskusyjne oraz krytyczne będę wymieniał w kolejności chronologicznej, zgodnie porządkiem wystąpienia uciążliwości oraz nieścisłości w Rozprawie. Nie zaznaczałem licznych błędów interpunkcyjnych.

1. Pierwszą uciążliwością, która od razu rzuca się w oczy jest brak numeracji praktycznie wszystkich wzorów oraz definicji, które występują w pracy. Wzorów nie jest przesadnie dużo i nie ma żadnych merytorycznych ani redakcyjnych przesłanek, aby każdy formalizm miał swój unikalny numer, do którego można by się odwołać. Brak takich unikalnych odwołań

zmusza do stosowania formy opisowej w postaci numeru strony oraz przedmiotu definicji czy równania, do czego będę musiał się uciec pisząc tą recenzję.

2. Strona 4: "Wprowadzenie i rozpowszechnienie komputerów umożliwiło automatyzację procesów decyzyjnych samych w sobie." Trafniejszym określeniem byłoby „Wprowadzenie i rozpowszechnienie komputerów pozwoliło na wykonywanie procesów decyzyjnych w sposób automatyczny".
3. Strona 4: "Jako stan bazowy można przyjąć okres (...)" – z tekstu nie wynika czego dotyczy ten „stan bazowy".
4. Strona 6: "przenoszalności" – po polsku lepiej brzmi "przenośności".
5. Strona 11: Definicja ataku adwersaryjnego w rozdziale 2.1.1 Autor zapewne zakłada, że kodowanie etykiet klas odbywa się na przy pomocy notacji one-hot-bit (przeciwdziedzina f to R^k, gdzie k jest liczbą klas). Warto byłoby to uściślić dla jasności dalszego wywodu.
6. Strona 12: "Zakłada się przy tym, że x' pozostaje w określonym sąsiedztwie x, determinowanym przez użytą funkcję odległości L_p oraz maksymalną odległość ϵ [43]" - te informacje należałoby zawrzeć w formalnej (matematycznej) definicji (2.2) jako ograniczenia nakładane na rozwiązanie w procesie optymalizacji.
7. Strona 12: $L\infty$ to odległość Czebyszewa (pozostałe miary zostały nazwane w tekście). W dalszej części Dysertacji Autor wielokrotnie używa normy oznaczonej jako || ||, która nie jest zdefiniowana w rozdziale 2.1.4. Zakładam, że to zapewne norma euklidesowa (L2).
8. Strona 13 "W szczególności, L2 jest często używana, gdy chcemy równomiernie rozłożyć zakłócenia." - co autor ma na myśli pisząc w kontekście odległości euklidesowej o "równomiernym rozkładzie zakłóceń"?
9. Strona 14, wzór ASR (Attack Success Rate) - co we wzorze oznacza symbol "δ_i"? Zakładam, że f_{target} to model będący przedmiotem ataku.
10. Strona 14: Autor definiuje również "Minimal Perturbation Norm", warto byłoby zdefiniować ten wskaźnik formalnie przy pomocy wzoru.
11. "Szczególną cechą ataków adwersaryjnych jest ich transferowalność (przenoszalność). Przykłady adwersaryjne x', zaprojektowane tak, aby oszukać model f_1, wprowadzają w różny od f_1 i niezależnie wytrenowany model f_2" – Zaproponowana tu definicja przenośności ataków adwersaryjnych jest wieloznaczna i trudna w interpretacji. Należałoby przedstawić ją w sposób ścisły.
12. Strona 14. "Modele o podobnych architekturach lub wytrenowane na porównywalnych zestawach danych są bardziej podatne na wspólne podatności adwersaryjne" – nasuwa się pytanie, czy istnieje ścisła miara podobieństwa architektur oraz ścisła miara podobieństwa zbiorów danych, który bezpośrednio koreluje z podatności modelu na ataki adwersaryjne? Jeśli tak, to powinno się te miary przedstawić właśnie w tym miejscu.
13. Strona 14-15: "Transferowalność między modelami f_1 i f_2 można analizować na podstawie ich granic decyzyjnych B_1 i B_2 (...) im większe jest nakładanie się B_1 i B_2 w przestrzeni cech (...)" Autor zapewne ma na myśli obszary (hiperprzestrzenie) decyzyjne a nie granice a "nakładanie się" oznacza część wspólną.
14. Strona 15: W definicji współczynnika transferowalności $T(f_1, f_2)$ brakuje wytłumaczenia, co oznacza symbol P oraz |, definicja wydaje się być oparta na prawdopodobieństwie warunkowym. Jest to istotny współczynnik, którego definicja wymaga doprecyzowania.
15. Strona 16: "Model z płaskim minimum w krajobrazie funkcji straty" - zapewne chodzi o zależność pomiędzy topologią funkcji straty jako funkcji parametrów wejściowych modelu.
16. Strona 16: "Współczynnik transferowalności ataku między modelami odpornymi, a standardowymi" podobnie jak w wypadku uwagi 14 w definicji $T(f_{robust}, f_{standard})$ brakuje wytłumaczenia, co oznacza symbol P oraz |.

17. Strona 17: "Ataki na dostępność dążą do zmniejszenia specyficzności klasyfikacji czyli generowania tzw. fałszywych alarmów." Zapewne Autor ma na myśli współczynnik "specificity" definiowany jako stosunek liczności prawdziwych negatywnych przypadków klasyfikacji (TN) dzielonych przez sumę liczności TN oraz fałszywych pozytywnych przypadków klasyfikacji (FP). Warto to zdefiniować, żeby uniknąć niejednoznaczności.
18. Strona 17: "Przedstawiony powyżej model zagrożeń nie dotyka takich zagadnień jak" lepiej brzmi „(...) nie obejmuje takich zagadnień jak (...)".
19. Strona 17: "np w przypadku celu będącego erozją zaufania do działań systemu" lepiej brzmi "np. w przypadku chęci obniżenia zaufania do systemu".
20. Strona 20: W tekście Dysertacji brakuje odwołania do rysunku 2.1.
21. Strona 20: „Kluczowym aspektem taksonomii NIST jest etap cyklu życia modelu, na którym przeprowadzany jest atak" Lepiej brzmi „cyklu życia modelu, w którym przeprowadzany jest atak".
22. Strona 23: Rys. 2.2 - to nie jest rysunek, tylko tabela. Informacje o zawartości tabeli powinny znajdować się nad nią a nie pod nią.
23. Strona 24: "gdzie: ϵ jest parametrem kontrolującym siłę perturbacji (...)" nie ma potrzeby ponownego definiowania parametrów, które zostały zdefiniowane wcześniej.
24. Strona 25: "W rzeczywistych zastosowaniach aplikacjach często konieczne jest" Lepiej brzmi "w praktyce".
25. Strona 25: "x $\in$ X na wyjścia y $\in$ Y" - czym są X i Y?
26. Strona 25: czym jest p(y) w definicji f_{target}?
27. Strona 27: "(...) wybór tej, która maksymalizuje stratę lub zmienia decyzję modelu. Algorytm (...)" poniższe podpunkty nie są algorytmem, informują jedynie o sposobie inicjalizacji zmiennych i problemie optymalizacyjnym.
28. Strona 27: "(...) przybliża się gradienty poprzez losowe perturbacje w przestrzeni wejściowej" - poniższe przybliżenie nie wyklucza możliwości dzielenia przez zero.
29. Strona 27: "Tak klasa metod" zapewne powinno być "Ta klasa metod".
30. Strona 27: "W najprostszym ataku typu Boundary Attack [13] aktualizacja zaburzenia odbywa się poprzez" - w mianowniku powinna być zapewne norma euklidesowa (patrz uwaga 7). Niezdefiniowana $\| \ \|$ jest używana kilkukrotnie na tej stronie oraz na kolejnych, nie będę się już odnosił do tego faktu w dalszych częściach tej recenzji.
31. Strona 27: "W metodzie tej, kolejne przybliżenie perturbacji uzyskuje się z użyciem przekształcenia:" - co oznacza symbol v we wzorze?
32. strona 28: "W metodzie ZOO [18] wprowadza się kierunkowe zaburzenia" – przy symbolu funkcji straty L nie ma argumentów, do tej pory zawsze były - czy to niedopatrzenie, czy litera L oznacza teraz coś innego niż do tej pory?
33. Strona 29: Jak zdefiniowana jest "najmniejsza skuteczna permutacja"?
34. Strona 29: "Przykładowo, w metodzie GenAttack (...)" przedstawiony poniżej algorytm jest klasycznym szablonem algorytmu genetycznego, bez wiedzy o działaniu krzyżowania, mutacji i selekcji w praktyce nie wiadomo, na jakiej zasadzie działa to konkretne zastosowanie.
35. Sekcja 2.5 a szczególnie sekcja 2.5.7 - część informacji zawartych w tej sekcji pojawiła się już we wcześniejszych rozdziałach.
36. Strona 36 - "Metody tworzenia modeli zastępczych zastępczych" – powtórzenie wyrazu „zastępczych".
37. Strona 36: "Metoda zakłada co najmniej możliwość wytrenowania modelu zastępczego dla nowego modelu objętego monitoringiem." – co w tym kontekście oznacza "co najmniej"? Założenie o możliwości wytrenowania modelu zastępczego wydaje się wystarczające.
38. Strona 37: "teoria RST" - teoria RS lub RST.

39. Strona 37: "https://qed.pl/nasze-rozwiazania/brightbox/", "https://qed.pl/nasze-rozwiazania/infoframes/" - oba linki przekierowują do tej samej strony "https://qedsoftware.com/"
40. Strona 37: "Metoda kwantylowa została wybrana specjalnie dla tego podejścia ze względu na jej efektywność obliczeniową oraz odporność" Brakuje sprecyzowania na co dokładnie ta metod jest „odporna".
41. Strona 42: "gdzie zakłada się że ($\kappa > 0.9$ wskazuje na udane przybliżenie)" - nawias jest niepotrzebny
42. Strona 43: "zgadzają się z wartościami docelowymi lub oryginalnymi predykcjami modelu" - czym różnią się od siebie wartości docelowe od oryginalnych predykcji modelu?
43. Strona 47: "W niniejszej pracy metodę przetestowano przy użyciu trzech oraz sześciu różnych ataków." - z kontekstu trudno się domyślić, ile dokładnie było rodzajów ataków i w jakich konfiguracjach zostały przeprowadzone.
44. Strona 50: Na rysunku 3.1 przedstawiono schemat do którego nie odwołano się w tekście, zapewne to „poniższy schemat" ze strony 49.
45. Strona 53: "oraz do niezaburzonych danych treningowych" Lepiej brzmi „danych treningowych" ewentualnie "oryginalnych danych treningowych".
46. Strona 53: "Do badan" powinno być "do badań"
47. Strona 53: "Do badan wybrano ataki charakteryzujące się niskim profilem działania" nie spotkałem się do tej pory w zagadnieniach ML z określeniem "niski profil działania".
48. Strona 54: "Stworzony wcześniej klasyfikator, bazujący na najczęstszych parametrach ataków, zostanie sprawdzony w wybranych scenariuszach w których intensywność ataku jest przeskalowana" - nie jest dla mnie jasne co Autor miał na myśli formując ten cel eksperymentu.
49. Strona 54: "Eksperymenty prowadzone były na zestawie 22 zbiorów danych tabelarycznych, udostępnionych w serwisie OpenML" - autor powinien wyspecyfikować o które zbiory danych dokładnie chodzi podając linki, z których można je pobrać. Tabela 4.1 nie ma odnośnika w tekście dysertacji. W tabeli 4.1 powinny też znaleźć się informacje dotyczące typów danych poszczególnych atrybutów w każdym ze zbiorów (np. liczba atrybutów numerycznych, kategorycznych itp).
50. Strona 59: Jakie jądra zostały wybrane do klasyfikatora SVM? Informacja ta pojawia się dopiero na stronie 75. Znacznie czytelniejsze byłoby umieszczanie informacji o hiperparametrach w tej samej sekcji, w której znajduje się opis metody.
51. Strona 60: "Sztuczne sieci neuronowe" - hiperparametry sztucznej sieci neuronowej zostały bardzo dokładnie opisane, natomiast brak informacji o konfiguracji SVM i XGBoost. Jakie były parametry pozostałych modeli? Informacja ta pojawia się dopiero na stronie 76. Znacznie czytelniejsze byłoby umieszczanie informacji o hiperparametrach w tej samej sekcji, w której znajduje się opis metody.
52. Strona 63: po raz kolejny opisywana jest metody ZOO, wystarczyłoby zdefiniować ją tylko raz. Ta sama uwaga dotyczy metody HopSkipJump (strona 65), PermuteAttack (strona 67), BIM (strona 68), FGM (strona 70), PGD (strona 71).
53. Strona 64: niekonsekwencja oznaczeń - do tej pory zakłócenia (perturbacje) oznaczane były symbolem δ, na tej stroni pojawia się Δx.
54. Strona 71: warunek $\|x' - x\|_p << \varepsilon$ jest tożsamy z $\| \delta \|_p < \varepsilon$, który byłby spójny ze wcześniejszymi oznaczeniami.
55. Strona 73: rysunek 4.1 powinien być tabelą.
56. Strona 75: parametr γ nie został wcześniej zdefiniowany - do czego służy?
57. Tabela 4.2 i 4.3 - opisy powinny pojawiać się nad a nie pod tabelami.

58. Strona 79, rysunek 4.2: czy różnice w odporności na poszczególne typy ataków pomiędzy wybranymi klasyfikatorami są istotne statystycznie?
59. Strona 81: "Przypomnijmy, krótko znaczenie tych atrybutów" Nie ma potrzeby przypominania znaczenia zdefiniowanych wcześniej pojęć. Wystarczy odwołać się do ich definicji, bardzo pomocne byłoby zastosowanie numeracji równań, których niestety w pracy brakuje.
60. Strona 87: "Do konstrukcji klasyfikatora binarnego wykrywającego ataki wykorzystano zbiór zawierający 264 wektory danych wejściowych. Każdy wektor danych odpowiadał jednej z kombinacji 22 pierwotnych zbiorów danych tabelarycznych x 3 modele je klasyfikujące x 4 przypadki ataków (3 ataki i 1 zbiór niezaatakwoany), i zawierał, wyliczone dla części diagnostycznej zbioru danych: agregaty wartości atrybutów diagnostycznych (opisane w dalszej części tej sekcji), liczność zbioru diagnostycznego, liczbę klas decyzyjnych, balanced accuracy pierwotnego klasyfikatora przetwarzającego dane diagnostyczne. Zawierał też etykietę binarną - atak lub brak ataku." Z tego zdania trudno jest wywnioskować ile faktycznie atrybutów miał zbiór treningowy, jaki był typ atrybutów oraz które atrybuty były zależne.
61. Strona 97 - kod w Python (zapewne biblioteka Keras) jest powtórzeniem wcześniej zdefiniowanej sieci (strona 60). Nie ma potrzeby wielokrotnego definiowania tej samej topologii sieci, wystarczy odwołać się do wcześniejszych sekcji.
62. Strona 121 - "Random Forest dla systemów wymagających stabilności i interpretiowalności" - interpretowalność działania klasyfikatora Random Forest jest dyskusyjna - w praktyce trudno wyobrazić sobie, aby użytkownik był wstanie poprawnie zinterpretować dziesiątki lub setki drzew decyzyjnych.

Należy zaznaczyć, że wymienione powyżej uwagi nie wpływają w sposób bardzo istotny na poznawcze oraz utylitarne wartości zaproponowanych rozwiązań i metodologii ich oceny.

5. Podsumowanie

Przytoczone powyżej uwagi polemiczne nie umniejszają wartości merytorycznej pracy, która stanowi oryginalny wkład Autora w zagadnienia związane z podatnością model ML na ataki adwersaryjne oraz wykrywanie tego typu szkodliwych aktywności.

Podsumowując recenzję stwierdzam, że moja generalna opinia o Rozprawie doktorskiej „Wykrywanie zmian danych wejściowych będących celowymi modyfikacjami ukierunkowanymi na wpływanie na wyniki działania modeli uczenia maszynowego" mgr inż. Piotra Biczyka jest pozytywna. Uważam, że przedstawiona do recenzji Rozprawa zawiera samodzielne i oryginalne, zaproponowane przez Autora rozwiązanie trudnego i aktualnego problemu badawczego, o potencjalnie licznych praktycznych zastosowaniach. Przedstawiona teza rozprawy została dowiedziona a postawione cele i zadania pracy zrealizowane. Odpowiada to ustawowym wymaganiom stawianym rozprawom doktorskim. **Na tej podstawie wnioskuję o dopuszczenie Rozprawy do publicznej obrony w celu uzyskania przez jej Autora stopnia doktora w dziedzinie nauk inżynieryjno – technicznych w dyscyplinie naukowej informatyka techniczna i telekomunikacja.**

Tomasz Hachaj