Częstochowa, dn. 25 kwietnia 2025 r.

prof. dr hab. inż. Marcin Korytkowski
Katedra Sztucznej Inteligencji
Wydział Informatyki i Sztucznej Inteligencji
Politechnika Częstochowska
al. Armii Krajowej 36
42-200 Częstochowa

Recenzja

rozprawy doktorskiej mgra inż. Stanisława Loty Analiza wyjaśnialności algorytmów uczenia maszynowego w zadaniach detekcji ataków w programowalnych sieciach komputerowych

Niniejszą recenzję opracowano zgodnie z uchwałą Senatu Społecznej Akademii Nauk z dnia 07/04/2025. Promotorem jest prof. dr hab. inż. Leszek Rutkowski.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Autor rozprawy doktorskiej porusza niezwykle ważny temat jakim jest wykorzystanie różnych technik analizy wyjaśnialności algorytmów dla potrzeb oceny systemów bazujących na technikach ML wykrywających ataki DDoS w sieciach definiowanych programowo (SDN- Software-Defined Network).

Recenzowana praca doktorska koncentruje się na trzech aspektach tak zdefiniowanego problemu:

1) Ocenie różnych technik wyjaśnialności użytych algorytmów, w tym stworzeniu kilku autorskich ich implementacji bazujących na koncepcji A. Rosenfelda
2) Wskazaniu praktycznego rozwiązania mającego na celu ochronę zasobów sieciowych (usług) przed atakami z rodziny DDoS bazującego na technikach ML
3) Ochronie systemów bazujących na technikach AI, zabezpieczających systemy sieciowe przed atakiem typu Adversrial Machine Learning oraz Data Poisoning.

Recenzowana praca stanowi wkład w wykorzystaniu technik inteligencji obliczeniowej w szeroko rozumianym bezpieczeństwie sieci komputerowych oraz interpretowalności modeli AI.

2. Zawartość rozprawy

Recenzowana praca mgra inż. Stanisława Loty składa się z dziesięciu rozdziałów, spisu rysunków, tabel oraz bibliografii. Dokument liczy 168 stron.

Pierwszy rozdział (Wstęp) jest dokładnym opisem tematyki podjętej w pracy, omawia jej zawartość oraz przedstawia tezę, która brzmi następująco:

„Autorskie propozycje i implementacje kryteriów wyjaśnialnej sztucznej inteligencji umożliwiają skuteczne wyjaśnienie podejmowanych decyzji przez modele uczenia maszynowego w procesie detekcji rozproszonych ataków odmowy usługi w sieciach definiowanych programowo."

W tym rozdziale autor przedstawia charakterystykę sieci SDN (Software- Defined Network), problem zapewnienia bezpieczeństwa systemów komputerowych w przypadku ataków DDoS, obecny stan światowej literatury podejmującej tę tematykę w powiązaniu z technikami „Soft Computing", a także pobieżny opis narzędzi służących do analizy wyjaśnialności działania technik AI typu SHAP, LIME (ocena wpływu poszczególnych cech na jakość działania wytworzonego (wytrenowanego) modelu).

Rozdział drugi stanowi znaczące rozwinięcie problemów opisanych we wstępie oraz zawiera kilka nowych ważnych w kontekście podjętej tematyki punktów, do których zaliczyć należy podpunkt 2.5 opisujący niezwykle ważny w obszarze sztucznej inteligencji problem Adversarial Machine Learning oraz Data Poisoning, czyli problem ataków na systemy uczące się poprzez zatruwanie danych treningowych.

Rozdział trzeci (Przegląd literaturowy) autor poświęcił analizie aktualnego stanu wiedzy światowej w następujących aspektach:

1) Wykorzystaniu modeli AI w zapobieganiu atakom DDoS
2) Wykorzystaniu technik XAI w zapobieganiu atakom DDoS
3) Wykorzystaniu technik zatruwania danych w kontekście ataków na systemy bazujące na AI służące do wykrywania ataków DDoS

Rozdział czwarty prezentuje środowisko badawcze zbudowane przez autora rozprawy pozwalające na analizę zdarzeń w sieci i zbieranie logów dla potrzeb budowy zestawów treningowych. Zgodnie z założeniem przyjętym w pracy całość rozwiązania bazuje na sieci komputerowej definiowanej programowo SDN. Zaprojektowana infrastruktura nie jest mocno rozbudowana, jednakże wystarczająca dla potrzeb zbudowania danych treningowych. Niestety nie odnalazłem w nim dokładnych informacji dot. sposobu generowania ruchu sieciowego symulującego atak DDoS.

W rozdziale piątym autor przedstawił technikę budowy własnego zestawu danych treningowych o nazwie DVCDDoS2022, który zawiera informacje dotyczące ruchu sieciowego wytworzonego w warunkach laboratoryjnych, który autor nazwał „normalnym" oraz ruch szkodliwy- zapis zdarzeń podczas sztucznie generowanego ataku DDoS. Dwie wspominane klasy zawierają mniej więcej 50% rekordów każda z 107220 wszystkich. Do weryfikacji poprawności danych i ewentualnej ich korekty w zestawie użyto algorytmu XGBoost.

W rozdziale tym w punkcie 5.3 zaprezentowano autorską technikę selekcji cech o nazwie SelectDVC. Opiera się ona na dość oczywistej koncepcji wyboru do zestawów treningowych tych cech, które najczęściej były wskazywane przez inne różnorodne metody selekcji (SelectKBest w dwóch wariantach: test chi-kwadrat, test Anova, Recursive Feature Elimination, Principal Component Analysis,

FeatureWiz). Autor uznał, że prawdopodobieństwo tego, że dana cecha jest istotna rośnie wraz z częstością pojawienia się jej w przypadku zastosowania innych metod. W zaproponowanym opisie procedury SelectDVC nie uwzględniono wag określających istotność danego sygnału wyznaczonego w poszczególnych algorytmach.

Rozdział szósty autor poświęcił porównaniu skuteczności algorytmu NGBoost z bardzo popularnymi i często wykorzystywanymi w praktyce Decision Tree, XGBoost, LightGBM oraz Random Forest w zadaniu wykrywania ataków DDoS. W założeniu porównaniu podlegać miało wydajność, złożoność obliczeniowa poszczególnych metod oraz ich wyjaśnialność. Autor zdefiniował kilka ważnych pytań, na które uzyskał odpowiedzi na podstawie przeprowadzonych symulacji i analizy otrzymanych wyników, tj. autor wykazał m.in., że model oparty na algorytmie NGBoost może z powodzeniem być stosowany w systemach detekcji ataków DDoS w sieciach SDN. Wykazał, także że zastosowanie technik selekcji cech przyczynia się do osiągnięcia najlepszych wyników w detekcji ataków DDoS w sieciach SDN (przeprowadzono redukcję z 77 do 19 sygnałów wejściowych) co znacząco wpłynęło na zwiększenie wydajności wykorzystanych algorytmów. Jednocześnie zaproponował on autorski sposób określenia metryki D zaproponowanej przez Rosenfelda.

W rozdziale siódmym autor pracy podjął temat związany z określeniem wyjaśnialności klasyfikacji ataków DDoS z wykorzystaniem techniki SHAP. W tym punkcie pracy doktorskiej zaproponowano algorytm, którego zadaniem jest umożliwienie usuwania poszczególnych cech z zestawu treningowego przy zachowaniu jak najwyższego poziomu jakości działania modeli użytych do klasyfikacji. Na podstawie przeprowadzonych symulacji autor wykazał, że możliwe jest znalezienie optymalnego kompromisu między liczbą cech a dokładnością użytego modelu XGBoost przy wykorzystaniu do oceny techniki SHAP. Autor zaproponował autorską metodę wyliczania metryki F, dzięki czemu możliwe stało się ograniczenie liczby cech używanych do wyjaśnienia predykcji modelu. W efekcie tych działań możliwe stało się zdefiniowanie optymalnej liczby cech, która zapewni równowagę między wysoką dokładnością a wyjaśnialnością modelu.

W rozdziale ósmym autor podejmuje tematykę odporności systemów machine learning na niewielkie zmiany w danych treningowych i testujących. Dla potrzeb określenia miary odporności zaprezentowano autorskie implementacje metryk monotoniczności, wierności, niekompletności oraz niewierności, oparte na mechanizmie wprowadzania celowych, kontrolowanych perturbacji do danych wejściowych. W kolejnych punktach tego rozdziału autor przedstawił wyniki przeprowadzonych eksperymentów wraz z ich analizą. Ostatecznie autor wykazał, że wszystkie analizowane algorytmy wykazują zdolność do zapewnienia stabilności każdego z użytych modeli ponadto zaproponowane metryki dostarczają dodatkowego narzędzia do oceny nie tylko skuteczności, ale też transparentności i interpretowalności modeli wykrywających ataki.

Rozdział dziewiąty został poświęcony niezwykle ważnej tematyce w światowej literaturze związanej z treningiem algorytmów z rodziny machine learning, a mianowicie „zatruwaniem" danych uczących (ang. data poisoning). Problem został poddany analizie na podstawie danych dot. ataków DDoS. Przeprowadzono szczegółową badania pozwalające ocenić wydajność modeli poddanych takim atakom

oraz oceniono skuteczność procesu oczyszczania danych. Badania oparto na metodzie SHAP do analizy wpływu zatruwania na wyjaśnialność modeli oraz zaproponowano nową autorską metrykę S, mierzącą stabilność wyjaśnień modeli pod wpływem tych zaburzeń. Ostatecznie, uzyskane wyniki wskazują na konieczność stosowania zaawansowanych metod oczyszczania danych i stałego monitorowania integralności zbiorów danych treningowych w zastosowaniach rzeczywistych.

Rozdział dziesiąty stanowi podsumowanie osiągniętych wyników w całej pracy oraz propozycję dalszych badań w kierunku rozwoju technik związanych z wyjaśnialnością modeli ML.

3. Ocena rozprawy

W ramach rozprawy doktorskiej Pan mgr inż. Stanisław Lota zaproponował zbiór oryginalnych technik związanych z zapewnieniem bezpieczeństwa infrastruktury IT w oparciu o algorytmy ML przed atakami DDoS z uwzględnieniem aspektu wyjaśnialności zastosowanych modeli. Tematyka pracy jest aktualna i a otrzymane rezultaty mogą być przydatne w zastosowaniach praktycznych w obrębie cyberbezpieczeństwa.

Oryginalny dorobek autora przedstawiony w pracy polega na:

1. opracowaniu autorskiej metody selekcji cech SelectDVC i zaprezentowanie jej skuteczności na przykładach związanych z wykrywaniem ataków DDoS,
2. opracowaniu środowiska badawczego opartego na sieciach programowalnych i stworzenie na jego podstawie autorskich zbiorów danych treningowych,
3. przeprowadzeniu analizy porównawczej algorytmów klasyfikacji zaadaptowanych do zadania detekcji ataków DDoS w sieciach definiowanych programowo,
4. przeprowadzeniu analizy porównawczej algorytmów klasyfikacji zaadaptowanych do zadania detekcji ataków DDoS w sieciach definiowanych programowo,
5. opracowaniu autorskich implementacji metryk D (Ocena wyjaśnialności modeli poprzez wydajność i złożoność), F (Ocena wyjaśnialności poprzez liczbę cech potrzebnych do wyjaśnienia decyzji), S (Ocena stabilności wyjaśnień modeli) zdefiniowanych przez A. Rosenfelda oraz autorskich propozycji ich modyfikacji,
6. opracowaniu autorskich implementacji metryk monotoniczności, wierności, niewierności oraz niekompletności,
7. opracowaniu autorskich wersji algorytmów ataków Flip-Label oraz oczyszczania zainfekowanych etykiet label-sanitization,
8. przeprowadzeniu kompleksowych eksperymentów w zakresie testowania odporności modeli ML (NGBoost, XGBoost, LightGBM, Decision Tree, Random Forest) na ataki zatrucia danych i perturbacje,

Rozprawa doktorska uwidacznia wysoką ogólną wiedzę teoretyczną i praktyczną oraz umiejętność prowadzenia pracy naukowej mgra inż. Stanisław Lota,

Niezależnie od mojej pozytywnej oceny pracy, nasunęły mi się następujące uwagi i pytania:

1) Najnowsze rozwiązania w zakresie bezpieczeństwa sieci w przypadku ochrony przed atakami DDoS wykorzystują jako jedną z podstawowych informacji adresy IP źródłowe i docelowe oraz porty TCP do których kierowany jest ruch sieciowy. W szczególności niezwykle istotny, do wykrycia ataku tego typu, jest nagły przyrost w przedziale czasowym ruchu z różnych adresów IP. Autor w rozdziale 5.1 podkreśla, że celowo usuwa tę informację z danych treningowych.
2) W pracy nie potrafiłem odnaleźć informacji dotyczącej jakiego typu usługi serwowane przez hosty podlegały atakom DDoS? Czy otrzymane wyniki badań są zależne od usług, na które przeprowadzony jest atak DDoS?
3) Czy zdaniem autora wynik działania oryginalnego algorytmu „SelectDVC" nie powinien być uzależniony od liczby użytych technik selekcji oraz od oceny ich jakości działania?

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że rozprawa doktorska „Analiza wyjaśnialności algorytmów uczenia maszynowego w zadaniach detekcji ataków w programowalnych sieciach komputerowych" prezentuje oryginalne rezultaty stanowiące rozwiązanie problemu naukowego oraz wkład w rozwój dyscypliny informatyka techniczna i telekomunikacja. Pan mgr inż. Stanisław Lota wykazał się umiejętnością samodzielnej pracy badawczej, znajomością literatury światowej i wiedzą w zakresie wykorzystania aktualnych technik uczenia maszynowego, najnowszych rozwiązań w zakresie projektowania i budowy sieci komputerowych typu SDN. Rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie informatyka techniczna i telekomunikacja. Recenzowana praca spełnia wymagania Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2022 r. poz. 574 z późn. zm.). w dyscyplinie naukowej informatyka techniczna i telekomunikacja. Wnoszę o jej przyjęcie i dopuszczenie do dalszych etapów postępowania doktorskiego.