dr hab. inż. **Marcin Niemiec**, prof. AGH
Instytut Telekomunikacji
Wydział Informatyki, Elektroniki i Telekomunikacji
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
al. Mickiewicza 30
30-059 Kraków
tel. 12 617 48 03
niemiec@agh.edu.pl

Kraków, 07.02.2025

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY NAUKOWEJ DYSCYPLINY INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

Tytuł rozprawy: Data Sample Optimisation for Network Traffic Classification Using Machine Learning Methods

Autor rozprawy: mgr inż. Jacek Krupski

1. Jakie zagadnienie naukowe jest rozpatrzone w pracy /teza rozprawy/ i czy zostało dostatecznie jasno sformułowane przez Autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?

Rozprawa doktorska dotyczy problematyki doboru odpowiednich próbek ruchu sieciowego w celu trenowania modeli klasyfikacji ruchu i wykrywania ataków sieciowych w oparciu o wybrane algorytmy uczenia maszynowego. Optymalizacja próbek pod kątem ich wielkości może znacząco zwiększyć wydajność/szybkość procesu uczenia, jednak kluczowym aspektem jest jednoczesny brak degradacji jakości uzyskanego klasyfikatora. W przypadku zbiorów składających się z ruchu surowego (tzn. nieprzetworzonego, bez wyliczonych/wyodrębnionych cech ruchu), optymalizację taką można przeprowadzić np. poprzez skracanie próbek danej sesji. Z kolei jeśli model uczenia maszynowego jest trenowany poprzez analizę cech ruchu sieciowego, w procesie uczenia warto brać pod uwagę ograniczoną liczbę cech, skupiając się na takich które mają widoczny pozytywny wpływ na jakość uzyskanego klasyfikatora. Poza tym optymalna pod względem wielkości próbka powinna także stanowić jak najbardziej reprezentatywną formę danego rodzaju ruchu sieciowego, tak aby wytrenowany model był jak najbardziej uniwersalny, działając poprawnie w różnych środowiskach. Stąd istnieje potrzeba badań wpływu modyfikacji próbek na jakość otrzymywanych klasyfikatorów.

Autor rozprawy postawił tezę, iż możliwa jest znacząca redukcja wielkości próby ruchu sieciowego służącej do opracowania klasyfikatora ruchu, bez zmniejszenia miar jakości tego klasyfikatora. Dzięki takiej redukcji proces uczenia będzie bardziej wydajny – będzie wymagał mniej zasobów i proces trenowania modelu będzie szybszy. Tak przedstawione zagadnienie naukowe, wraz z jednoznacznie określoną tezą rozprawy, jest jasno sformułowane przez Autora oraz przedstawia ważne i interesujące zagadnienie badawcze. Ponieważ dotyczy ono realnego problemu we współczesnej teleinformatyce, a weryfikacja tezy bazuje na wynikach testów opracowanych przez Autora modeli klasyfikacji ruchu, rozprawa ma charakter doświadczalny.

2. **Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł / w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle/ świadczący o dostatecznej wiedzy Autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?**

Początkowe rozdziały pracy mają głównie charakter przeglądowy i wprowadzający do tematyki bezpieczeństwa w środowisku sieciowym oraz podstaw działania sieci teleinformatycznych i uczenia maszynowego. Potwierdzają one dostateczną wiedzę Autora z dziedziny dotyczącej rozprawy. Przedstawione zostały podstawowe protokoły, idea enkapsulacji, sposoby i narzędzia przechwytywania oraz zapisywania ruchu sieciowego, a także cechy ruchu sieciowego. W tej części (rozdział 3.1) Autor popełnił pewne ewidentne pomyłki: pole adresu w protokole IPv6 ma długość 128 bitów, a nie 18; warstwa dostępu do sieci w modelu TCP/IP odpowiada warstwom L1 i L2 w modelu OSI, a nie warstwom L6 i L7. Jednak z kontekstu wynika, że są to raczej niezamierzone literówki, a nie błędy wynikające z braku wiedzy co do podstaw działania sieci teleinformatycznych. Autor pokrótce przedstawiał także istotę uczenia maszynowego, w tym algorytmy bazujące na drzewach decyzyjnych i klasyfikatory linowe. Przedstawiono także metryki oceny jakości modeli, takie jak precyzja, czułość i dokładność. Osobny rozdział poświęcono opisowi ośmiu baz próbek ruchu sieciowego, których użyto do testów. W tej części przedstawiono także pojęcia związane z rodzajami ruchu sieciowego, stosownymi architekturami czy oprogramowaniem złośliwym. W większości opis tych pojęć jest poprawny choć zdarzają się pewne uogólnienia/uproszczenia, np.: „*Backdoors utilise publicly unknown issues, whereas Exploits operate on the well known problems*") lub stwierdzenie, że VPN zapewnia prywatność, co nie zawsze jest prawdą (przykładowo, IPsec VPN w trybie transportowym wprawdzie zapewnia poufność danych, ale nie prywatność użytkowników).

Rozdział 2 rozprawy jest poświęcony analizie literaturowej, bezpośrednio związanej z tematyką rozprawy. Przeanalizowano prace związane z klasyfikacją ruchu i wykrywaniem ataków sieciowych przy użyciu sztucznej inteligencji, a także prace związane z problematyką optymalizacji próbek ruchu służących do trenowania modeli – zarówno ruchu surowego jak i zbioru cech ruchu sieciowego. Wprawdzie tematyka ta jest obecnie niezwykle aktualna i można byłoby odwołać się do większej liczby prac, ale przedstawiona analiza źródeł jest wystarczająca. Przeprowadzono ją w sposób właściwy, a wnioski sformułowane na jej podstawie są jasne i przekonywujące. Umiejscawia to rozprawę na tle innych badań w dziedzinie oraz właściwie identyfikuje problem badawczy. Autor w bibliografii zamieścił również 4 własne prace związane z tematem swojej rozprawy doktorskiej, opublikowane w latach 2021-2025.

3. **Czy Autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są właściwe?**

Postawione zagadnienie badawcze zostało rozwiązane przez Autora rozprawy. Aby wykazać, że możliwa jest redukcja zbioru cech ruchu sieciowego bez degradacji jakości klasyfikatora, Autor przeanalizował cechy dostępne w kilku znanych bazach. Usuwał zbędne cechy – np. te, które są zależne od infrastruktury. Kilka nadmiarowych usunięto na podstawie analizy korelacji cech. Jednak najciekawsze wyniki osiągnięto poprzez zbadanie wpływu poszczególnych cech na wartości metryk oceny klasyfikatorów. Osiągnięto to poprzez uszeregowanie cech – od tych, które mają największy wpływ, po te które mają najmniejszy wpływ na wartości metryk. Ponieważ badania wykonano na różnych bazach próbek ruchu oraz dla różnych algorytmów uczenia ze zróżnicowanymi parametrami, Autor zaproponował własny sposób wyliczenia końcowego rankingu cech. Metoda ta została formalnie opisana, a osiągnięte wyniki potwierdziły jej poprawność. W ten sposób Autor uzyskał wiedzę, które z cech są najmniej ważne i można je usunąć z procesu uczenia. Okazało się, że większość z cech można usunąć bez wyraźnej degradacji jakości metryk (choć stwierdzenie zawarte w podsumowaniu rozprawy, że do poprawnej klasyfikacji wystarczy tylko 8 cech wydaje się zbyt radykalne).

W przypadku próbek surowego ruchu, najpierw Autor wykazał potrzebę anonimizacji analizując działanie drzew decyzyjnych i określając, które cechy decydują o klasyfikacji próbek. Następnie poprzez wykonanie szeregu testów potwierdził, że jakość modeli trenowanych na danych zanonimizowanych jest lepsza oraz został określony wpływ długości próbek na dokładność klasyfikatora. Dla różnych poziomów anonimizacji określono długość próbek ruchu, przy którym wzrost dokładności osiąga wartości bliskie maksymalnych. Wyniki badań pozwoliły także porównać dwa sposoby anonimizacji i potwierdzić, że wyższy poziom anonimizacji pozwala z reguły na osiągnięcie maksymalnej dokładności szybciej (czyli przy mniejszej próbce). W tej części pracy wyniki wielokrotnie powtarzano, a w tabelkach zamieszczone zostały wartości średnie (szkoda jednak, że nie określono błędów pomiarowych). W ramach pracy przeprowadzono szereg testów, ale gwoli ścisłości należy wspomnieć, że niektóre wnioski są raczej oczywiste, np. że zbiory powinny być odpowiednio zbalansowane czy że klasyfikatory wytrenowane na podstawie jednej bazy próbek działają gorzej, gdy klasyfikują ruch sieciowy z innej bazy – szczególnie jeśli bazy te zawierają różne rodzaje ruchu czy ataki różnego typu.

Pomimo drobnych niedoskonałości, przy użyciu poprawnie wybranych metod badawczych Autor potwierdził postawioną tezę, że możliwa jest znacząca redukcja wielkości próby ruchu sieciowego, która służy do opracowania klasyfikatora ruchu, bez zmniejszenia miar jakości tego klasyfikatora. Zostało to udowodnione przy użyciu wybranych baz zawierających próbki ruchu sieciowego. Są to bazy zróżnicowane i powszechnie używane do testów przez naukowców zajmujących się klasyfikacją ruchu przy użyciu algorytmów uczenia maszynowego.

4. Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek Autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?

Oryginalny dorobek opisany w rozprawie doktorskiej dotyczy przede wszystkim sposobów redukcji próbek ruchu sieciowego w celu wytrenowania modeli klasyfikacji ruchu przy użyciu algorytmów uczenia maszynowego (przy jednoczesnym zachowaniu wysokiej jakości modelu). Praca koncentruje się na dwóch rodzajach próbek – jednym bazującym na ruchu surowym, a drugim opartym na analizie cech ruchu sieciowego. Dodatkowo zaproponowany został sposób na uporządkowanie cech ruchu w celu ich późniejszej redukcji oraz podział na poziomy anonimizacji próbek surowego ruchu sieciowego. Samodzielny wkład Autora rozprawy jest istotny oraz oryginalny w stosunku do stanu wiedzy i poziomu techniki. Recenzent wyodrębnił najważniejsze oryginalne elementów rozprawy, które zostały krótko podsumowane poniżej.

- Autor rozprawy zaproponował trzy poziomy anonimizacji danych służących do trenowania modeli uczenia maszynowego (AL0, AL1 i AL2) na podstawie próbek surowego ruchu sieciowego. W zależności od poziomu anonimizacji, z nagłówków pakietów usuwana jest oryginalna informacja o adresach MAC i IP oraz portach lub usuwane są wszystkie informacje zawarte w nagłówkach pakietów (do warstwy L4 włącznie). Potwierdzono również, że taka anonimizacja pozytywnie wpływa na proces trenowania i późniejsze działanie klasyfikatora ruchu.

- W wyniku przeprowadzonych badań wykazano, że próbki surowego ruchu sieciowego stanowiącego poszczególne sesje, można skrócić w sposób znaczący, bez zauważalnej degradacji metryk oceny jakości danego klasyfikatora. Autor wykazał, że zamiast próbek zawierających 1000 bajtów, można zastosować znacznie mniejsze próbki. Pozytywnie wpłynie to na wydajność procesu uczenia. Co więcej, na wyższym poziomie anonimizacji (AL2), zwykle redukcja długości próbek może być większa niż przy poziomie AL1 – zamiast próbek o długości 1000 bajtów, próbki mogą zawierać jedynie kilkadziesiąt bajtów, a jakość klasyfikatora nie ulegnie widocznej degradacji.

- Wyniki badań przedstawionych w rozprawie potwierdzają, że jedynie ograniczony zbiór cech ruchu sieciowego ma istotny wpływ na jakość wytrenowanego modelu. Wiele pozostałych cech nie wpływa na poprawę wartości miar klasyfikatora w sposób zauważalny. Autor na podstawie szeregu testów (różne bazy próbek ruchu, algorytmy uczenia maszynowego i parametry) i zaproponowanej metodzie (zgodnie ze wzorem nr 10) opracował ranking cech, które mają największy wpływ na jakość klasyfikatora. Dzięki temu można zdecydować, które z cech ruchu warto usunąć z próbek treningowych, tak aby zwiększyć wydajność procesu uczenia, bez degradacji jakości klasyfikatora.

- Zaproponowana została taksonomia grupująca poszczególne cechy ruchu sieciowego. Autor zaproponował klasyfikowanie poszczególnych cech w siedmiu grupach. Trzy pierwsze grupy stanowią cechy bezpośrednio odczytywane z próbek przechwyconego ruchu, a cztery kolejne to cechy wyliczone na podstawie ruchu sieciowego.

5. Czy Autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników /zwięzłość, jasność, poprawność redakcyjna rozprawy/?

Przede wszystkim należy podkreślić, że rozprawa doktorska jest bardzo staranna pod względem edycyjnym oraz została napisana jasnym, w pełni zrozumiałym językiem. Tekst jest czytelny, a przedstawione rozwiązania, szczegóły dotyczące trenowania i oceny klasyfikatorów wystarczająco jasno opisane. Autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników. Dzięki temu recenzent nie miał trudności ze zrozumieniem treści pracy. Poniżej opisano drobne uwagi językowe czy edycyjne, jednak obniżają one jakość pracy jedyne w minimalnym stopniu i nie zmieniają ogólnej pozytywnej oceny rozprawy.

Przy uważnym czytaniu można znaleźć pojedyncze błędy językowe i literówki, np.: atak „*Dos*", „*algorithmss*" czy "*section 3.3*". Występuje niekonsekwentny zapis pojęcia cyberbezpieczeństwo: „*cybersecurity*" i „*cyber security*". Pojawił się błędny zapis zmiennych w tekście objaśniającym oznaczenia we wzorze nr 10 – we wzorze występują duże litery „R", a w opisie małe. Czasem w tekście występują kolokwializmy (np.: „*Famous Wireshark is a Swiss army knife for network traffic manipulations*" czy "*information on who is talking to whom*") oraz ogólniki (np.: "*This tool [SplitCap] works quite quickly*"). Dodatkowo, w komentarzu do wyników zaprezentowanych na rysunkach 32 i 33 Autor stwierdził, że: „*graphs show significantly higher decision-tree scores in cases where we have 13 or more best features*" – jednak efekt ten występuje dopiero od 14 cechy. Wygląda to jednak bardziej na pomyłkę językową niż błąd merytoryczny spowodowany złą interpretacją wyników.

Struktura dokumentu jest prawidłowa, choć niektóre fragmenty można udoskonalić, np. warto unikać jednozdaniowych akapitów. Często znak minus jest stosowany jako myślnik (znak interpunkcyjny). Incydentalnie zdarza się przecinek zamiast kropki w zapisie ułamków dziesiętnych (np. „*3,71 GB*" czy „*12,7 GB*"). Jednak jak już wspomniano – strona językowa i edytorska pracy jest na wysokim poziomie, a przedstawione rysunki i tabele są czytelne.

6. Jakie są słabe strony rozprawy i jej główne wady?

O pewnych słabych stronach rozprawy, np. związanych z prezentacją wyników, recenzent wspomniał już w poprzednich punktach. Poniżej przedstawione zostały cztery uwagi merytoryczne – według recenzenta kluczowe dla ocenianej rozprawy doktorskiej.

- W niektórych fragmentach rozprawy brakuje wyczerpującej dyskusji uzyskanych wyników – szczególnie tych, które w jakiś sposób są intrygujące. Przykładem mogą być

wyniki zaprezentowane w tabeli 14 (klasyfikacja ruchu związanego z atakiem DoS): dlaczego uzyskano tak słabe wyniki metryk oceny jakości w przypadku modelu zarówno trenowanego jak i testowanego na zbiorze NB15-v2? Inny przykład dotyczy wyników zaprezentowanych na rysunkach 40 i 41 oraz tabeli 36: czy zasadne jest aby cechy o numerach 5 i 6 były uszeregowane na tak wysokich miejscach skoro mają niewielki wpływ na wzrost jakości modelu (a nawet można zauważyć mały spadek wartości *RF acc*) w porównaniu do cech znajdujących się na kolejnych miejscach?

- Autor rozprawy zaproponował ograniczenie zbioru cech ruchu sieciowego, które mają służyć do opracowania klasyfikatorów ruchu. W rozdziale 6.2 przedstawiono, które z cech bazy NF-UQ-NIDS-v2 warto usunąć ze zbioru próbek. Jednak czasem wygląda to bardziej na działanie intuicyjne niż na decyzję podjętą w wyniku formalnej oceny. Przykładowo, wśród odrzuconych cech znalazła się *DNS_QUERY_ID*, gdyż Autor stwierdził: „*The DNS_QUERY_ID is a string of digits, informing only about the query number and contributes nothing to the examination*". Niestety działanie to nie zostało umotywowane w sposób przekonujący. Weźmy choćby pod uwagę przewidywalne numery zapytań DNS – stanowią one znaną podatność, która może być wykorzystana przez atakującego. Wydaje się więc, że cecha ta może być jednak istotna dla klasyfikatora, np. czy nie doszło do ataku typu DNS spoofing.

- W rozprawie brakuje także umotywowania decyzji Autora co do wyboru algorytmów czy poszczególnych parametrów podczas badań. Przede wszystkim zostało wybranych 6 algorytmów uczenia maszynowego („*tests were performed using selected artificial intelligence algorithmss*"), ale w rozprawie zabrakło dyskusji dlaczego akurat ta grupa algorytmów jest właściwa czy reprezentatywna. Podobnie jest w przypadku parametrów przedstawionych w tabeli 19 – nie wiadomo dlaczego wybrano właśnie takie parametry a nie inne oraz dlaczego zaproponowano zmianę poszczególnych parametrów z takim skokiem a nie innym.

- Pewien niedosyt sprawia brak połączenia wyników badań przedstawionych w rozdziale 8 z zaproponowanym podziałem cech ruchu na grupy (autorska taksonomia opisana w rozdziale 3.6). W ramach badań Autorowi udało się uzyskać interesujące wyniki związane z uporządkowaniem cech mających największy wpływ na klasyfikatory ruchu. Jednak wyniki tych badań nie nawiązują do zaproponowanej taksonomii cech, tzn. nie wiadomo w jakim stopniu cechy te skupione są w poszczególnych grupach czy typach/rodzajach grup, nie przedyskutowano czy taki podział cech na grupy jest związany z finalnym rankingiem cech (np. przedstawionym na rysunkach 38 i 39), itd. Bez takiej analizy zaproponowany podział cech na grupy może wydawać się arbitralny.

7. Jaka jest przydatność rozprawy dla nauk technicznych?

Recenzent wysoko ocenia przydatność ocenianej rozprawy doktorskiej dla nauk technicznych. Jak już wspomniano – rozprawa dotyczy aktualnego i ważnego problemu z zakresu cyberbezpieczeństwa i sztucznej inteligencji: optymalizacji wielkości próbek ruchu sieciowego (zarówno w surowej postaci jak i zbioru wybranych cech) używanych przy opracowywaniu klasyfikatorów ruchu. Dzięki rozprawie nie tylko wiadomo, że możliwa jest znaczna redukcja wielkości tych próbek, ale także jakie cechy powinny być brane w pierwszej kolejności podczas procesu trenowania oraz jaka część surowego ruchu stanowiącego daną sesję może być pomijana bez degradacji jakości klasyfikatora. Nowa wiedza pozyskana w wyniku badań może pomóc w opracowywaniu wydajnych i skutecznych klasyfikatorów ruchu. Poza tym Autor udowodnił, że odpowiednio przeprowadzona anonimizacja ruchu ma pozytywny wpływ na jakość klasyfikacji i wykazał, że wyższy poziom anonimizacji pakietów pozwala jednocześnie na większą redukcję długości próbek ruchu. Wyniki rozprawy mogą posłużyć do budowania wydajnych i skutecznych rozwiązań służących do wykrywania ataków sieciowych.

8. **Konkluzja – do której z następujących kategorii Recenzent zalicza rozprawę:**
   a) niespełniająca wymagań stawianych rozprawom przez obowiązujące przepisy
   b) wymagająca wprowadzenia poprawek i ponownego recenzowania
   c) <u>spełniająca wymagania</u>
   d) spełniająca wymagania z wyraźnym nadmiarem
   e) wybitnie dobra, zasługująca na wyróżnienie

9. **Uzasadnienie**

Biorąc po uwagę wszystkie wspomniane aspekty rozprawy doktorskiej, moja końcowa konkluzja jest jednoznacznie pozytywna. Autor zaprezentował w rozprawie ogólną wiedzę teoretyczną w dyscyplinie i przedstawił oryginalne rozwiązanie problemu naukowego. Wykazał, że możliwa jest znaczna redukcja wielkości próbek ruchu sieciowego użytych podczas trenowania modeli uczenia maszynowego w celu opracowania skutecznych klasyfikatorów ruchu, przy jednoczesnym braku degradacji jakości klasyfikacji. Poza tym był w stanie opracować ranking cech ruchu sieciowego, które mają największy wpływ na jakość klasyfikatorów, a także zaproponował sposób grupowania cech ruchu oraz przedstawił trzy poziomy anonimizacji surowego ruchu i wykazał jaki mają one wpływ na działanie opracowanych modeli. Tym samym Autor rozprawy potwierdził umiejętność samodzielnego prowadzenia pracy naukowej.

Podsumowując stwierdzam, że rozprawa doktorska spełnia wymagania *Ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce* oraz wnioskuję o dopuszczenie Pana mgr. inż. Jacka Krupskiego do dalszych etapów przewodu doktorskiego.