Szczecin, 28 lutego 2025 r.

dr hab. inż. Dariusz Frejlichowski, prof. uczelni
Katedra Systemów Multimedialnych
Wydział Informatyki
Zachodniopomorski Uniwersytet
Technologiczny w Szczecnie
Ul. Żołnierska 49
71-210 Szczecin

RECENZJA ROZPRAWY DOKTORSKIEJ

DLA RADY NAUKOWEJ DYSCYPLINY

INFORMATYKA TECHNICZNA I TELEKOIMUNIKACJA

TYTUŁ ROZPRAWY: Data sample optimisation for network traffic classification using machine learning methods

AUTOR ROZPRAWY: mgr inż. Jacek Krupski

PROMOTOR: dr hab. inż. Marcin Iwanowski, prof. PW

PROMOTOR POMOCNICZY: dr inż. Waldemar Graniszewski

WSTĘP

Niniejsza recenzja dotyczy Rozprawy doktorskiej, autorstwa mgr. inż. Jacka Krupskiego, zatytułowanej „Data sample optimisation for network traffic classification using machine learning methods", napisanej w 2024 r., pod opieką promotorską dr. hab. inż. Marcina Iwanowskiego, prof. PW, przy wsparciu Promotora pomocniczego dr. inż. Waldemara Graniszewskiego. Niniejsza recenzja została przygotowana na zlecenie prof. dr. hab. inż. Jarosława Arabasa, Przewodniczącego Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja, zawarte w piśmie z dnia 9 grudnia 2024 r.

1. ZAKRES, CHARAKTER I STRUKTURA PRACY

Przedmiotem ocenianej Rozprawy są badania nad optymalizacją próbek ruchu sieciowego przy akceptowalnie wysokiej wydajności klasyfikacji, na dwóch najpopularniejszych formach zapisu ruchu sieciowego: ruchu surowym oraz ruchu zapisanym jako cechy ruchu sieciowego.

Prace w zakresie ruchu surowego obejmują analizę wpływu segmentów pliku PCAP na wynik klasyfikacji ruchu sieciowego w celu ustalenia tych segmentów, które prowadzą do wysokich wyników dla zbioru trenującego, ale nie sprawdzają się w przypadku nowych rzeczywistych danych testowych. Zagadnienie to jest istotne w kontekście wykorzystywania zarejestrowanego wcześniej ruchu sieciowego do trenowania modeli uczenia maszynowego, ponieważ kluczowym celem jest osiągnięcie uogólnionego rezultatu uczenia tak, aby możliwe było jego wykorzystanie w różnych środowiskach. Przeprowadzone dalsze prace w ramach Rozprawy pozwoliły na potwierdzenie, że możliwe jest skuteczne klasyfikowanie ruchu niebezpiecznego dla małych próbek ruchu sieciowego, a także optymalizacja zbioru cech ruchu sieciowego na podstawie autorskiej metody selekcji cech opartej na zespole klasyfikatorów.

Prowadzone przez Autora Rozprawy badania, jak również proponowane rozwiązania, mieszczą się w zakresie dwóch istotnych informatycznych dziedzin – uczenia maszynowego i cyberbezpieczeństwa, co potwierdza ich ścisły związek z Dyscypliną Informatyka Techniczna i Telekomunikacja.

Rozprawa składa się z dziewięciu rozdziałów, spośród których pierwszy poświęcony jest krótkiemu wprowadzeniu do poruszanych w Pracy zagadnień, przedstawieniu problemu badawczego, celu, tezy, planu Pracy, a także najważniejszych osiągnięć Autora Rozprawy.

W drugim rozdziale przedstawiono przegląd literatury z uwzględnieniem systemów wykrywania włamań opartych na uczeniu maszynowym, a także metod optymalizacji surowego ruchu i zbioru cech ruchu.

W rozdziale trzecim omówiono podstawy teorii sieci komputerowych, przedstawiono opis popularnych narzędzi do analizy ruchu sieciowego, jak również dwóch przywoływanych już form analizy ruchu sieciowego. Zaproponowano także nową taksonomię cech zgodnie z ich podobieństwem i zależnością od infrastruktury sieciowej, w której zostały przechwycone.

Rozdział czwarty dotyczy sztucznej inteligencji, a w szczególności modeli uczenia maszynowego, używanych w Rozprawie, wraz z ich oceną. Skoncentrowano się na podejściach bazujących na drzewach, modelach liniowych oraz metrykach przeznaczonych do oceny algorytmów uczenia maszynowego.

Rozdział piąty koncentruje się na przeglądzie zestawów danych wybranych do osiągnięcia celów Rozprawy, ze szczególnym uwzględnieniem ataków na ruch sieciowy, które są rejestrowane w tych zestawach danych.

W rozdziale szóstym w pierwszej kolejności skoncentrowano się na danych, które mogą negatywnie wpływać na proces analizy ruchu sieciowego przez modele uczenia maszynowego. Aby zapewnić, że dane uczące nie będą zaburzały późniejszego działania w warunkach rzeczywistych, dla surowego ruchu sieciowego zaproponowano i opisano trzy poziomy anonimizacji. Przeanalizowano również kwestię metod anonimizacji stosowanych w innych badaniach. W dalszej części opisano wyniki badań w zakresie wyboru cech ruchu sieciowego i zdolności do uogólniania zbioru danych.

W rozdziale siódmym przedstawiono wyniki badań w zakresie surowego ruchu sieciowego. Najpierw przedstawiono opis techniczny sposobu przetwarzania plików PCAP, a następnie testy na trzech indywidualnych zestawach danych. Wykorzystano trzy poziomy anonimizacji, które rozróżniają, które dane sieciowe są przetwarzane.

W ramach rozdziału ósmego skoncentrowano się na stworzeniu rankingu cech w celu ustalenia najistotniejszych cech dla wybranych klas ataków sieciowych. W tym celu przeprowadzono szereg obliczeń na różnych zestawach danych zawierających cechy ruchu sieciowego.

Rozdział dziewiąty zawiera podsumowanie Rozprawy, ze wskazaniem najważniejszych rezultatów badań oraz osiągnięć Autora.

Zawarta w rozdziale pierwszym teza została sformułowana przez Autora w sposób jasny. Tezę można uznać za poprawnie skonstruowaną, przedstawioną w postaci pozwalającej na jej potwierdzenie (bądź nie), odczuwalny jest jednak brak ściśle zdefiniowanych, mierzalnych parametrów, które w znaczący sposób potwierdzą Autorowi osiągnięcie celu procesu dowodzenia zaproponowanej tezy badawczej. Oczekiwane dane pojawiają się co prawda w dalszej, eksperymentalnej części Rozprawy, co jest niezbędne z perspektywy badawczego charakteru prac, jednakże określenie bezpośrednio po sformułowaniu celu i tezy wprost, jak mierzalnie Autor interpretuje takie pojęcia jak: „znacząco" („significantly"), „bez pogorszenia jakości klasyfikacji" („without worsening the quality of the classification") oraz „wysokiej jakości miary" („high-quality measures"), podkreśliłoby już w tym miejscu ważkość uzyskiwanych w Rozprawie rezultatów.

W pracy można odnaleźć zarówno elementy teoretyczne, jak i praktyczne. Do pierwszej grupy zaliczyć należy między innymi przeprowadzoną analizę literatury, powiązaną z umiejscowieniem prac Autora w kontekście stosowanych obecnie podejść, jak również sformułowane przez Autora metody i podejścia dla poszczególnych aspektów proponowanego rozwiązania. Jednakże jako dominujący należy wskazać doświadczalny charakter Rozprawy. Potwierdzeniem tego jest już sformułowana przez Autora teza, która w sposób wyraźny narzuca konieczność jej dowodzenia poprzez praktyczne eksperymenty badawcze. Również przyjęty przez Autora tok myślowy prowadzi poprzez kolejne uzasadnienia przyjętych założeń oraz szereg eksperymentów do potwierdzenia oczekiwań stawianych przez Autora względem opracowywanego podejścia, jak również warunków osiągnięcia założonego celu Rozprawy.

2. ANALIZA ŹRÓDEŁ, STANU WIEDZY I ZASTOSOWAŃ W PRZEMYŚLE

W Rozprawie odniesiono się do 169 pozycji literaturowych różnego rodzaju, w tym do pięciu publikacji we współautorstwie Autora, Promotora i Promotora pomocniczego. Warto podkreślić, że w czterech spośród nich Autor Rozprawy znajduje się na pierwszym miejscu, co podkreśla Jego wkład w powstanie tych publikacji.

Odejmując od całkowitej liczby pozycje w spisie bibliograficznym należące do Autora Rozprawy, uzyskujemy 164 obce publikacje, które mieszczą się w zakresie lat 1981 – 2024. Wśród tych publikacji zaledwie niewielką część stanowi materiał wydany przed rokiem 2000, co potwierdza silną koncentrację na najnowszych doniesieniach z dziedziny bliskiej tematyce Rozprawy. Na uwagę zasługuje także operowanie różnego rodzaju źródłami, nie tylko publikacjami naukowymi, ale także definicjami standardów, czy też raportami sięgającymi początkowych lat w rozwoju Internetu, jak również odwołaniami do witryn internetowych stosowanych narzędzi. Należy jednak pamiętać, że zagadnienia, którym poświęcona jest Rozprawa, a więc cyberbezpieczeństwo i uczenie maszynowe, to jedne z najpopularniejszych obecnie gałęzi informatyki, przez co uwzględnienie wszystkich ważnych publikacji w ich zakresie staje się praktycznie niemożliwe. Konieczne jest ograniczenie się do głównych nurtów, a zwłaszcza aspektów bliskich zakresowi pracy Autora. W związku z tym, przeprowadzoną przez Autora analizę światowego stanu wiedzy, która z konieczności jest wybiórcza, należy ocenić jako dobrą i wystarczającą, co wynika z koncentracji Autora na materiałach najbliższych zakresowi merytorycznemu Rozprawy. Należy jednak zauważyć, że pewną nieścisłość wprowadza przyjęta struktura rozdziałów poświęconych analizie materiałów źródłowych. Autor Rozprawy wprost deklaruje rozdział drugi jako poświęcony temu zagadnieniu (rozdział nosi tytuł „Literature review"). W rzeczywistości rozdział ten dotyczy jednakże tylko pewnej części wcześniejszych światowych osiągnięć zbieżnych z zakresem Rozprawy, podczas gdy w kolejnych rozdziałach nadal *de facto* odbywa się proces analizy publikacji, co szczególnie widoczne jest w rozdziale piątym (pt. „Machine Learning").

Pomimo powyżej wskazanej nieścisłości w konstrukcji struktury Pracy, zastosowany przez Autora ciąg myślowy jest czytelny i opiera się na analizie źródłowej przeprowadzonej w sposób odrębny dla kolejnych aspektów związanych z merytorycznym zakresem Rozprawy. Brakuje natomiast szerszego kontekstu wdrożeniowego dotyczącego obecnie stosowanych praktycznych rozwiązań, które mogłyby stanowić inspirację, bądź nawet wprost miejsce, w którym rezultat Rozprawy mógłby zostać zaimplementowany. Warto podkreślić bowiem, że jedną z istotnych zalet Pracy jest jej wysoce aplikacyjny charakter. Uzyskane rezultaty zasługują na wdrożenie do praktyki i wykorzystanie w systemach związanych z bezpieczeństwem sieci. Warto byłoby także ten aspekt przeanalizować i uwypuklić, chociażby w kontekście konkretnych funkcjonujących obecnie narzędzi dostępnych na rynku.

FR

3. METODA ROZWIĄZANIA POSTAWIONEGO ZAGADNIENIA, ORYGINALNOŚĆ ROZPRAWY

Autor Rozprawy osiągnął postawiony w pierwszym rozdziale cel, którym było znalezienie optymalnego zestawy cech dla ruchu sieciowego, dla którego wyniki klasyfikacji pozwolą osiągnąć akceptowalne metryki oceny. W sposób eksperymentalny potwierdzono także założenia zawarte w tezie, zgodnie z którymi wykazano, że możliwe jest ograniczenie zbioru potrzebnych cech bez straty w jakości klasyfikacji. Kluczowe znaczenie z perspektywy potwierdzenia rozwiązania postawionych w Rozprawie zagadnień miały liczne eksperymenty, których wyniki przedstawiono w rozdziałach szóstym, siódmym i ósmym. Przyjęta przez Autora Rozprawy metodologia jest prawidłowa, szczególnie w kontekście przyjętej kolejności opracowywania i potwierdzania wydajności autorskich pomysłów i rozwiązań.

Za kluczowe z perspektywy oceny oryginalności Rozprawy należy uznać zaproponowane podejście oparte na trzypoziomowej anonimizacji danych w trakcie analizy plików PCAP. Jest to ważne, ponieważ brak anonimizacji może zakłócać faktyczne, praktyczne działanie opracowanych metod, gdy nowe dane są znacząco inne od tych, których użyto w procesie uczenia.

Kolejnym ważnym osiągnięciem jest wykazanie, że oparcie się na cechach ruchu sieciowego niezależnych od infrastruktury sieciowej, w której ruch został przechwycony lub emulowany, jest wystarczające do skutecznej analizy ruchu sieciowego. Analogicznie, potwierdzono, że odpowiednia selekcja pozwala na uzyskanie zredukowanego zestawu cech skutecznego w identyfikacji znanych ataków sieciowych. Dodatkowo, wykazano że analizę cech można przeprowadzić używając metod rankingowych.

Autor Rozprawy potwierdził także, że aspekt możliwości uogólnienia w kontekście doboru zbiorów testowych ma zauważalny wpływ na praktyczne działanie metod. Potwierdził również że w przypadku surowego ruchu, jak również dla cech ruchu sieciowego wystarczające jest oparcie się na prostych klasyfikatorach.

W Rozprawie odniesiono się do pięciu prac opublikowanych przez Autora w toku realizowanej pracy naukowej. W przypadku czterech spośród nich Autor Rozprawy wystąpił na pierwszym miejscu listy Autorów. W niektórych interpretacjach w zakresie analizy dorobku naukowego przyjmuje się, że kolejność autorów jest decydująca i wskazuje wkład, w tym twórczy, w powstanie publikacji. W przywołanej sytuacji należy zatem mówić o istotnym wkładzie Autora Rozprawy w badania, które stanowiły przedmiot publikacji, jak i Dysertacji. Z drugiej strony, należy zauważyć, że na liście autorów wszystkich przywołanych publikacji znajdują się także Promotor i Promotor pomocniczy, co z kolei potwierdza, że Autor ma kompetencje do pracy w zespole badawczym. Jednocześnie, analiza zakresu Dysertacji nie pozostawia wątpliwości, że główne tezy badawcze zostały przez Doktoranta sformułowane i dowiedzione przy Jego znaczącym udziale w tym zespole.

Jak już wspomniano, rezultaty Rozprawy mają znaczący walor wdrożeniowy, aplikacyjny. Przedstawione w Rozprawie rozwiązania, jak również zawarte w niej wyniki

FR

eksperymentalne, potwierdzają wysoką przydatność uzyskanych przez Autora rezultatów prac dla nauk technicznych, w szeroko rozumianym pojęciowo zakresie cyberbezpieczeństwa. Ich wdrożenie nie powinno stanowić większego problemu i jest naturalnym kolejnym etapem prac nad zagadnieniem. Potwierdzają to liczne aspekty implementacyjne, które Autor poruszył w Dysertacji.

4. UWAGI REDAKCYJNE I TECHNICZNE

Poziom redakcyjny (napisanej w języku angielskim) Rozprawy należy uznać za dobry, a zastosowany język za czytelny dla odbiorcy. Niewątpliwą zaletą Autora jest łatwość w przekazywaniu własnych poglądów, idei oraz osiągniętych rezultatów. Uzyskane wyniki zostały przedstawione w sposób poprawny i przekonujący. Docenić należy przy tym w szczególności zwięzłość i jasność formułowanych treści. Stosowana przez Autora narracja prowadzi czytelnika poprzez kolejne fazy powstawania opracowanych podejść, niejako pozwalając na obserwację chronologiczną kolejnych zmian i rozwoju efektów pracy Autora.

Niewiele znaleźć można w pracy błędów i są to głównie błędy interpunkcyjne. Kilka jest też niepoprawnie użytych (bądź brakujących) przedimków. Przykładowo na stronie 11 w tytule podrozdziału „1.1 Description of research problem" brakuje „the" przed „research", podobnie jak w przypadku treści „The structure of decision tree, presented in Fig. 16", gdzie przed „decision" również powinno się znaleźć „the". Natomiast całościowo praca napisana jest poprawnym, zrozumiałym i czytelnym językiem angielskim.

Wśród innych drobnych niedociągnięć wskazać można brak rozwinięcia w Rozprawie jednego z kluczowych dla niej skrótów. Autor nigdzie w Pracy nie wskazał, że PCAP pochodzi od „Packet CAPture", mimo iż jednocześnie rozwija tak oczywiste skróty jak choćby URL czy HTTP.

W przypadku większości ilustracji w podpisie Autor nie wskazuje źródła. W konsekwencji nie zawsze wiadomo, czy zostały one opracowane wyłącznie przez Autora Rozprawy, zostały opracowane na podstawie określonego źródła, czy też może pochodzą z konkretnych publikacji. Przy tym, ilustracja góry lodowej na str. 47 (Fig. 9) nie wnosi istotnej treści do omawianych zagadnień, chyba że w sposób w pełni wiarygodny odzwierciedla udział procentowy indeksowanego Internetu („web pages" na przywołanym rysunku), sieci głębokiej („deep web") i tzw. ciemnej sieci („dark web").

Swobodną lekturę Rozprawy zaburzają w kilku miejscach wprowadzone wypunktowania w procesie przedstawiania przeglądu literatury, jak np. na str. 17 – 18 w podrozdziale „2.1 Machine learning-based intrusion detection systems", str. 18 – 19 w podrozdziale „2.2 Optimisation of sample of raw traffic", czy str. 20 – 21 w podrozdziale „2.3 Optimisation of set of traffic features". Tekst główny w pozostałej części Rozprawy jest pisany w sposób ciągły, stąd tego typu dłuższe wypunktowania zaburzają przyjęty schemat.

W pewnym stopniu odczuwalny jest także brak krótkich podsumowań w zakończeniu niektórych rozdziałów, które wskazywałyby, co istotnego stanowi rezultat dyskusji przeprowadzonej w danym rozdziale. Na przykład, w rozdziale trzecim takie podsumowanie pokrótce podkreśliłoby, co jest głównym efektem przeprowadzonej analizy. Obecnie, można spekulować, że kluczowy jest podrozdział „3.6 Proposed taxonomy of network traffic features" i że jest on konsekwencją dyskusji prowadzonej w podrozdziałach 3.1 – 3.6. Podobne wrażenie, dotyczące braku podsumowań, można mieć po lekturze rozdziału piątego i szóstego.

Można także polemizować z pewnymi mniej istotnymi deklaracjami zawartymi w Dysertacji – czy faktycznie smartfon Samsung Galaxy A53 jest równie popularny jak system operacyjny Windows 11 Home (str. 23)? Czy istnieją badania w tym zakresie?

5. UWAGI DYSKUSYJNE I KRYTYCZNE

Ogólna ocena Rozprawy jest bardzo dobra, mimo iż w niniejszej recenzji pojawiały się różne uwagi, które sprowadzają się do kilku elementów, które można rozpatrywać w kategoriach pewnych nieznacznych braków czy niedociągnięć Dysertacji. Stanowią one jednak raczej wstęp do dyskusji, rozważań, otwarte pytania, a nie krytykę i obejmują:

1. Brak podjęcia w szerszym zakresie wątku wdrożenia rezultatu Dysertacji. Praca ma znaczący praktyczny efekt, w związku z czym wyłania się pytanie, czy i – jeżeli tak – to w jakim zakresie zbadano możliwość transferu wyników badań do praktycznych rozwiązań, które mogłyby konkurować co najmniej na rynku polskim? Czy badano aspekt przewagi konkurencyjnej nad dostępnymi na rynku narzędziami?
2. Co prawda zbiory testowe użyte w badaniach mają adekwatny przekrój danych, aby było możliwe potwierdzenie przyjętych założeń badawczych, ale pojawia się pytanie, czy nie można było się oprzeć na nowszych danych benchmarkowych?, a także – co wydaje się być szczególnie kuszące w pracy badawczej – czy nie było możliwości przygotowania własnych danych testowych? Czy też może z jakiegoś powodu nie było to uzasadnione? Jaka była przy tym podstawa wyboru danych testowych?
3. W kontekście założonej tezy badawczej i jej opisu w pierwszym rozdziale – na jakim wstępnym poziomie Autor zakładał mierzalne parametry dla takich pojęć jak: „znacząco" („significantly"), czy „wysokiej jakości miary" („high-quality measures")? Czy też może te oczekiwania wykrystalizowały się na konkretnym poziomie dopiero w trakcie realizowanych przez Autora Rozprawy prac badawczych?
4. W rozdziale 4 pn. „Machine learning" Autor dokonał wyboru podejść z zakresu uczenia maszynowego, które wykorzystał w swoich badaniach. Na jakiej podstawie dokonano tego wyboru? Dlaczego zrezygnowano z innych popularnych rozwiązań w tej dziedzinie?
5. W podrozdziale 6.2 pn. „Selection of features" pojawiają się istotne założenia związane z wyborem cech. Najpierw wskazano, że szereg spośród nich, związanych z infrastrukturą, zostało odrzuconych na podstawie dyskusji w podrozdziale 6.1.,

FK

jednakże na zakończenie wskazano w bardzo ogólny sposób, że „The remaining 32 numerical features could be easily checked for any correlations using the Pearson correlation coefficient. Fig. 11 presents the achieved results in the form of a heat map". Ten fragment wydaje się być ograniczonym odniesieniem do bardzo istotnych badań nad ważnym konceptem Autora, związanym z usuwaniem zbędnych cech. Jakie były zatem szczegółowe badania w tym zakresie?

Powyżej przedstawione uwagi, podobnie jak i inne zastrzeżenia, które pojawiły się w niniejszej recenzji, nie obniżają ogólnej wartości naukowej Rozprawy, stanowiąc jedynie element dyskusji nad jej treścią.

6. PODSUMOWANIE

Podsumowując recenzję, jednoznacznie stwierdzam, że przedstawiona Dysertacja mgr. inż. Jacka Krupskiego pt. „Data sample optimisation for network traffic classification using machine learning methods" – pomimo przedstawionych uwag i kwestii dyskusyjnych – zawiera oryginalne rozwiązania problemów naukowych oraz potwierdza kompetencje Autora w zakresie prowadzenia badań o charakterze naukowym, w tym w kategoriach badań podstawowych, przemysłowych i prac rozwojowych, a zatem **spełnia wymagania** Ustawy o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki, stawiane rozprawom doktorskim i w związku z tym wnioskuję o dopuszczenie mgr. inż. Jacka Krupskiego do dalszych etapów przewodu doktorskiego, w tym wnoszę o dopuszczenie Rozprawy do publicznej obrony.

Frejlichowski

(Dariusz Frejlichowski)