Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich
Profesora Sylwestra Kaliskiego 7,
85-796 Bydgoszcz
Dyscyplina naukowa: **Informatyka Techniczna i Telekomunikacja**
za pośrednictwem:
Rady Doskonałości Naukowej
pl. Defilad 1
00-901 Warszawa
(Pałac Kultury i Nauki, p. XXIV, pok. 2401)

Marek Pawlicki
Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich
Wydział Telekomunikacji, Informatyki i Elektrotechniki
Profesora Sylwestra Kaliskiego 7,
85-796 Bydgoszcz

Wniosek

z dnia 02.01.2025

o przeprowadzenie postępowania w sprawie nadania stopnia doktora habilitowanego w dziedzinie **Nauk inżynieryjno-technicznych** w dyscyplinie **Informatyka Techniczna i Telekomunikacja.**

Określenie osiągnięcia naukowego będącego podstawą ubiegania się o nadanie stopnia doktora habilitowanego: **cykl powiązanych tematycznie artykułów naukowych, pt. "Zaawansowane metody uczenia maszynowego oraz wyjaśnialnej sztucznej inteligencji (xAI) dla wykrywania ataków sieciowych."**

Wnioskuję – na podstawie art. 221 ust. 10 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce – aby komisja habilitacyjna podejmowała uchwałę w sprawie nadania stopnia doktora habilitowanego w głosowaniu **jawnym**.

Zostałem poinformowany, że:

Administratorem w odniesieniu do danych osobowych pozyskanych w ramach postępowania w sprawie nadania stopnia doktora habilitowanego jest Przewodniczący Rady Doskonałości Naukowej z siedzibą w Warszawie (pl. Defilad 1, XXIV piętro, 00-901 Warszawa).

Kontakt za pośrednictwem e-mail: kancelaria@rdn.gov.pl , tel. 22 656 60 98 lub w siedzibie organu. Dane osobowe będą przetwarzane w oparciu o przesłankę wskazaną w art. 6 ust. 1 lit. c) Rozporządzenia UE 2016/679 z dnia z dnia 27 kwietnia 2016 r. w związku z art. 220 - 221 oraz art. 232 – 240 ustawy z dnia 20 lipca 2018 roku - Prawo o szkolnictwie wyższym i nauce, w celu przeprowadzenie postępowania o nadanie stopnia doktora habilitowanego oraz realizacji praw i obowiązków oraz środków odwoławczych przewidzianych w tym postępowaniu.

Szczegółowa informacja na temat przetwarzania danych osobowych w postępowaniu dostępna jest na stronie www.rdn.gov.pl/klauzula-informacyjna-rodo.html

PODPIS ZAUFANY
MAREK
PAWLICKI
02.01.2025 00:03:43 [GMT+1]
Dokument podpisany elektronicznie podpisem zaufanym

(podpis wnioskodawcy)

Załączniki:

1. Dane wnioskodawcy
2. Kopia dokumentu potwierdzającego posiadanie stopnia doktora
3. Autoreferat wnioskodawcy
4. Wykaz osiągnięć naukowych
5. Deklaracje współautorów dotyczące wkładu pracy, certyfikat o odbytym stażu i oświadczenie o udziale w projektach badawczych
6. Publikacje wchodzące w skład osiągnięcia naukowego

Autoreferat

1. Imię i nazwisko

Marek Pawlicki

2. Posiadane dyplomy, stopnie naukowe/artystyczne – z podaniem nazwy, miejsca i roku ich uzyskania oraz tytułu rozprawy doktorskiej.

- **2020** - Stopień naukowy doktora
    - Nadany uchwałą 3/447 Senatu Uniwersytetu Technologiczno-Przyrodniczego (UTP) im J. i J. Śniadeckich w Bydgoszczy
    - Tytuł rozprawy: „Zastosowanie metod uczenia maszynowego do wykrywania ataków sieciowych”
- **2011** - Tytuł magistra inżyniera - Wyższa Szkoła Informatyki w Łodzi

3. Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych/artystycznych

- **2020-obecnie**

  Adiunkt
    - Wydział Telekomunikacji, Informatyki i Elektrotechniki,
    - Politechnika Bydgoska im J. i J. Śniadeckich (PBS) (dawniej: Uniwersytet Technologiczno-Przyrodniczy (UTP) im J. i J. Śniadeckich i od 2021)
- **2018-2020**

  Asystent
    - Uniwersytet Technologiczno-Przyrodniczy im J. i J. Śniadeckich (UTP)

4. Omówienie osiągnięć, o których mowa w art. 219 ust. 1 pkt. 2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478 z późn. zm.)

4.1. Tytuł osiągnięcia naukowego

Zaawansowane metody uczenia maszynowego oraz wyjaśnialnej sztucznej inteligencji (xAI) dla wykrywania ataków sieciowych.

4.2. Osiągnięcie naukowe – jednotematyczny cykl publikacji naukowych

Poniżej przedstawiono, w ramach jednotematycznego cyklu naukowego, dziewięć publikacji. Publikacje ułożone są w kolejności chronologicznej, obejmują listę dziewięciu artykułów wydanych w latach 2020–2024. Wszystkie prezentowane wartości bibliometryczne odzwierciedlają stan na dzień 31. grudnia 2024 r., zgodnie z bazami danych naukowych:

- Web of Science (WoS)
    - https://www.webofscience.com/wos/author/record/ABC-3938-2021
- Scopus (Sco)
    - https://www.scopus.com/authid/detail.uri?authorId=57204352435
- Google Scholar (GSc)
    - https://scholar.google.com/citations?user=tsqBC7QAAAAJ&hl=en
- Baza Expertus Politechniki Bydgoskiej (xPBS)
    - https://bg.pbs.edu.pl/expertus/bib/

Publikacje te dokumentują zakres badań, które początkowo prowadziłem jako część zespołu badawczego, a z czasem coraz bardziej samodzielnie, w miarę rozwoju naukowego. W swoich pracach zajmowałem się rozwiązaniami zaawansowanych problemów w zakresie systemów wykrywania ataków sieciowych wykorzystujących sztuczną inteligencję. Rozwiązania te dotyczą następujących problemów:

- Propozycji metod uczenia transferowego i uczenia z małą ilością danych jako rozwiązanie istotnego problemu wysokiego kosztu pozyskiwania oznakowanych próbek ataków sieciowych do uczenia nadzorowanego w wykrywaniu ataków sieciowych.
- Propozycji zmian w procesie przetwarzania danych pozwalających na poprawienie metryk detekcji algorytmów uczenia maszynowego.
- Zaproponowanie nowych metod zapewniających przejrzystość działania algorytmów AI oraz sposobów oceny jakości uzyskanych wyjaśnień w zastosowaniach dotyczących wykrywania ataków sieciowych.

Wpisy bibliograficzne dla artykułów w czasopismach zawierają informacje o wartości czynnika Impact Factor (IF) wydawnictwa z roku publikacji, zgodnie z Journal Citation Reports

https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports

Informacje o poziomie konferencji, na których przedstawiano artykuły, są podane według rankingu CORE w roku publikacji.

Każdy wpis zawiera również informację o liczbie punktów przyznanych przez Ministerstwo w roku publikacji.

Do każdego wpisu z listy publikacji wchodzących w skład cyklu dodane zostały również informacje o moim wkładzie autorskim, zgodnie z wytycznymi wzorca CRediT (Contributor Roles Taxonomy)

https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement

Informacje te są zgodne z deklaracjami współautorów zawartymi w Załączniku do wniosku.

	DOI	**Publikacja**	**Punkty**	**IF/CORE**
[C01]	https://doi.org/10.1016/j.neucom.2020.07.138	Choraś, M., and **Pawlicki, M**. (2021) Intrusion detection approach based on optimised artificial neural network. In Neurocomputing, 452, (pp. 705–715).	140	IF 5.719

		CRediT: Konceptualizacja, Opracowanie danych, Analiza formalna, Badania, Metodologia, Oprogramowanie, Walidacja, Pisanie - oryginalny szkic, Pisanie - recenzja i edycja Cytowania WoS: 54; Sco: 73; Gsc: 112;		
[C02]	https://doi.org/10.1016/j.neucom.2022.06.002	**Pawlicki, M**., Kozik, R. and Choraś, M. (2022) A survey on neural networks for (cyber-) security and (cyber-) security of neural networks. In Neurocomputing, 500, (pp. 1075–1087). **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja Cytowania WoS: 31; Sco: 44; Gsc: 60;	140	IF 5.779
[C03]	https://doi.org/10.1007/978-3-031-42823-4_21	**Pawlicki, M**., Pawlicka, A., Kozik, R., and Choraś, M. (2023) How to Boost Machine Learning Network Intrusion Detection Performance with Encoding Schemes. In International Conference on Computer Information Systems and Industrial Management (pp. 283-297). Cham: Springer Nature Switzerland. **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja, nadzór, zarządzanie projektem, pozyskiwanie funduszy	40	CORE C

		Cytowania WoS: 0; Sco: 0; Gsc: 0;		
[C04]	https://doi.org/10.1145/3538969.3544428	**Pawlicki, M.**, Kozik, R., and Choraś, M. (2022) Towards Deployment Shift Inhibition Through Transfer Learning in Network Intrusion Detection. In Proceedings of the 17th International Conference on Availability, Reliability and Security (pp. 1-6). **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja, nadzór, zarządzanie projektem, pozyskiwanie funduszy Cytowania WoS: ; Sco: 1; Gsc: 2;	70	CORE B
[C05]	https://doi.org/10.1109/IJCNN54540.2023.10191496	**Pawlicki, M.,** Kozik, R., and Choraś, M. (2023), Improving Siamese Neural Networks with Border Extraction Sampling for the use in Real-Time Network Intrusion Detection. In International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, (pp. 1-8) **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja, nadzór, zarządzanie projektem, pozyskiwanie funduszy Cytowania WoS: 2; Sco: 2; Gsc: 3;	70	CORE B

[C06]	https://doi.org/10.1109/DSAA60987.2023.10302535	**Pawlicki**, **M.,** (2023) Towards Quality Measures for xAI algorithms: Explanation Stability. in IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA), (pp. 1–10) Cytowania WoS: 0; Sco: 0; Gsc: 0;	140	CORE B
[C07]	https://dl.acm.org/doi/10.1145/3665451.3665527	**Pawlicki**, **M.,** Pawlicka, A., Kozik, R. and Choraś, M., (2024) Explainability versus Security: The Unintended Consequences of xAI in Cybersecurity, in Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems, (pp. 1–7) **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja, nadzór, zarządzanie projektem, pozyskiwanie funduszy Cytowania WoS: 0; Sco: 0; Gsc: 2;	140	CORE A
[C08]	https://doi.org/10.1109/ACCESS.2024.3454084	**Pawlicki**, **M.,** (2024) ARIA, HaRIA, and GeRIA: Novel Metrics for Pre-Model Interpretability, in IEEE Access, vol. 12, (pp. 123561-123580) Cytowania WoS: 0; Sco: 0; Gsc: 0;	100	IF 3.4
[C09]	https://doi.org/10.1007/s10462-024-10972-3	**Pawlicki, M.**, Pawlicka, A., Kozik, R., and Choraś, M., (2024) The survey on the dual nature of xAI challenges in intrusion detection and their potential for AI innovation. In Artificial Intelligence Review, 57, (art. no. 330)	140	IF 10.7

		CRediT: konceptualizacja, metodologia, analiza formalna, badania, pisanie - oryginalny szkic Cytowania WoS: 0; Sco: 0; Gsc: 0;		

4.3. Sumaryczne dane naukometryczne dotyczące publikacji naukowych w cyklu

Poniższa tabela prezentuje sumaryczne dane dotyczące publikacji naukowych, które weszły w skład przedstawionego osiągnięcia. Wartości oddają stan na dzień 31. grudnia 2024, zgodnie z bazami Web of Science (WoS), Scopus (Sco), Google Scholar (GSc) i Expertus PBS (xPBS).

Liczba publikacji wchodzących w skład osiągnięcia	9		
Sumaryczna punktacja ministerialna	980		
Sumaryczny IF osiągnięcia	25,598		
Sumaryczna liczba cytowań osiągnięcia	WoS	Sco	GSc
	87	120	179

Poniższa tabela zawiera skonsolidowane dane na temat profilu naukowego autora, tak, jak zostały zarejestrowane w bazach naukowych: Web of Science (WoS), Scopus (Sco) oraz Google Scholar (GSc). Dane przedstawiają stan na dzień 31. grudnia 2024 roku.

	WoS	**Sco**	**GSc**
Index Hirscha autora	13	17	22
Liczba wszystkich cytowań autora	680	1124	1745

Bez autocytowań	538	853	-
Liczba dokumentów	67	104	112

4.4. Omówienie celu naukowego wyżej wymienionych prac i osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania

Do głównych osiągnięć naukowych habilitanta, będących wkładem w dziedzinę informatyki, zaliczam:

1. Opracowanie i analizę zastosowania sztucznych sieci neuronowych i innych metod uczenia maszynowego w kontekście wykrywania cyberataków w sieciach komputerowych.
2. Badanie przesunięcia wdrożeniowego w systemach wykrywania ataków sieciowych i zaproponowanie metod uczenia transferowego w celu adaptacji modeli uczenia maszynowego do nowych środowisk sieciowych.
3. Zaproponowanie nowej metody trenowania sieci neuronowych typu Siamese z wykorzystaniem ekstrakcji granic. Zaproponowana metoda może zwiększyć skuteczność SNN w wykrywaniu nowych typów ataków sieciowych.
4. Zaproponowanie nowej metody oceny technik xAI poprzez badanie stabilności wyjaśnień xAI w obliczu zakłóceń danych.
5. Zaproponowanie nowej metody wykorzystania algorytmów xAI do tworzenia ataków typu adversarial.
6. Zaproponowanie wyodrębnienia zagadnienia interpretowalności przed wytrenowaniem modelu (pre-model interpretability) jako nową dziedzinę xAI.
7. Opracowanie nowych metryk do oceny cech przed uczeniem w zagadnieniu wyjaśnialności przed modelem (pre-model interpretability).

4.4.1. Cyberprzestrzeń i ataki sieciowe

Cyberprzestrzeń, choć uznawana za jedno z największych osiągnięć ludzkości [1], jest również potencjalnym miejscem dla rozwijania cyberprzestępczości. Od momentu jej

powstania różnego rodzaju złośliwi aktorzy i podmioty o złych intencjach próbują wykorzystywać Internet dla własnych korzyści – zdobywania pieniędzy, informacji lub władzy [2].

W maju 2021 roku parlament Belgii, belgijskie uniwersytety oraz inne instytucje naukowe stały się celem ataku DDoS [3]. Nieznana grupa hakerów spowodowała przeciążenie usług belgijskiego dostawcy Internetu Belnet, odcinając jego użytkowników od dostępu do sieci, częściowo lub całkowicie. Atak dotknął ponad dwieście organizacji.

W październiku 2021 roku holenderska policja ostrzegła 29 osób, które korzystały z usług DDoS-for-hire dostępnych w Internecie. W ostrzeżeniach zawarto informacje o przestępczym charakterze ich działań. Ostrzeżenia były wynikiem zamknięcia przez policję strony internetowej oferującej usługi DDoS w 2020 roku [4]. Przykład ten podkreśla problem i powszechność ataków DDoS, które obecnie są dostępne dla każdego, kto jest gotów je wykorzystać.

Z kolei atak opisany w [5] dotknął zakład wodociągowy w mieście Oldsmar na Florydzie. Choć samo włamanie trwało nie dłużej niż pięć minut, spowodowało 111-krotne zwiększenie poziomu wodorotlenku sodu w wodzie dostarczanej do miasta, co udało się naprawić dopiero po niemal sześciu godzinach. Jak podano, wodorotlenek sodu może powodować wymioty, ból w klatce piersiowej i brzuchu, oparzenia skóry oraz utratę włosów.

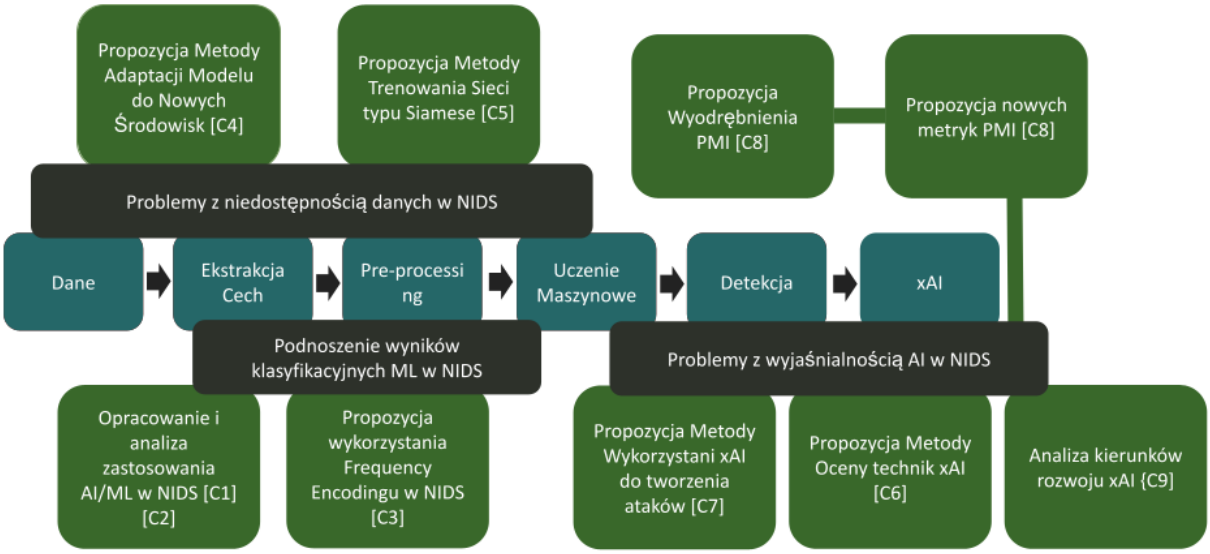
Jeszcze inny przykład dotyczy Universal Health Services, który jest jednym z największych systemów opieki zdrowotnej w Stanach Zjednoczonych. System padł ofiarą cyberataku [6]. Atak dotknął wiele szpitali, powodując awarie systemów komputerowych, usług telefonicznych, łączności internetowej, a w konsekwencji opóźnienia w realizacji procedur medycznych.

Szeroka gama naruszeń cyberbezpieczeństwa przez ataki sieciowe przyczyniła się do intensyfikacji badań nad różnorodnymi metodami wykrywania zagrożeń. W literaturze domenowej wyłoniły się dwa główne nurty badań i rozwoju: wykorzystujące sygnatury oraz wykrywające anomalie. Systemy IDS (Intrusion Detection Systems - systemy wykrywania ataków sieciowych) wykorzystujące sygnatury działają poprzez odnoszenie się do bazy rozpoznanych ataków, podczas gdy metody wykrywające anomalie tworzą model ruchu niewykazującego znamion ataków i generują alarm przy wykryciu odchylenia od tego modelu [7].

W tej dziedzinie, zastosowanie algorytmów uczenia maszynowego (Machine Learning-ML) oraz sztucznych sieci neuronowych (Artificial Neural Networks, ANN) nie jest nowym pomysłem. Już w 2009 roku przeprowadzano oceny koncepcji wykorzystania ANN do wspomagania systemów wykrywania ataków sieciowych (Network Intrusion Detection

Systems - NIDS) [8]. Wykorzystanie AI/ML w NIDS nadal jednak wiąże się z szeregiem wyzwań, które stoją na drodze szerokiej adopcji tej technologii w cyberbezpieczeństwie.

W swoich pracach rozwijam ten dotychczasowy stan wiedzy proponując metody rozwiązywania tych problemów. Schemat części istniejących problemów i propozycji ich rozwiązania znajduje się na Rys. Rysunek 1.



Rysunek 1. Proponowane w cyklu rozwiązania problemów z aplikacją AI w NIDS

4.4.2. Zagadnienia związane z wykorzystaniem ML i ANN w NIDS – zarys problemu

NIDS wykorzystujące ML stanowią szeroki obszar badań, w którym podejmowane są różne próby osiągnięcia lepszej efektywności w zastosowaniu algorytmów ML do wykrywania zagrożeń sieciowych.

Obszary badawcze obejmują pozyskiwanie odpowiednich zbiorów danych [9][10][11], różnorodne metody przetwarzania danych [12][13][14][15], a także zastosowanie licznych algorytmów ML i Deep Learning (DL) [16][17][18][19][20][21].

Badania przedstawione w [C01] i [C02] są bezpośrednią kontynuacją i rozszerzeniem prac, które prowadziłem w doktoracie. Przeprowadziłem kompleksowy przegląd metodologii mających na celu zwiększenie efektywności ANN w NIDS, takich jak dobór odpowiedniego paradygmatu ANN, kwestia balansowania danych, dostosowanie hiperparametrów. W badaniu zidentyfikowałem siedem szerokich kategorii paradygmatów ANN stosowanych

w zadaniu NIDS - to rekurencyjne sieci neuronowe (RNN), głębokie sieci neuronowe (DNN), ograniczone maszyny Boltzmanna (RBM), głębokie sieci przekonań (DBN), konwolucyjne sieci neuronowe (CNN), głębokie maszyny Boltzmanna (DBM) oraz głębokie autoenkodery (DAE). Poruszyłem też ważne kwestie związane z wykorzystaniem ANN, takie jak podatności na ataki typu adversarial oraz wyzwania związane z ich wyjaśnialnością. Przeprowadzono eksperymenty w celu identyfikacji optymalnych konfiguracji hiperparametrów, które mają dużą wagę w maksymalizacji zdolności detekcyjnych ANN w wykrywaniu zagrożeń sieciowych.

Jako podsumowanie i rozwinięcie eksperymentów zawartych w doktoracie badania te stanowią punkt wyjścia niniejszego cyklu w kierunku rozwijania możliwości ML i ANN w zadaniach związanych z wykrywaniem ataków sieciowych.

4.4.3. Uczenie Maszynowe w Wykrywaniu Ataków Sieciowych

Często badania dotyczące wykorzystania ANN w NIDS prezentują arbitralnie dobrane topologie i konfiguracje hiperparametrów, pomijając ogromny wpływ, jaki hiperparametry mogą mieć na osiąganą dokładność.

Pierwszym wynikiem badań jest podkreślenie niewykorzystanego potencjału tkwiącego w konfiguracjach hiperparametrów, które często pozostają nieprzetestowane (gdyż ML/AI traktowane jest przez wielu badaczy jak czarna skrzynka).

Odnotowałem konfiguracje topologii i hiperparametrów osiągające najwyższe wyniki detekcji oraz zbliżone do najlepszego zestawu, aby znaleźć najmniejsze topologie, oferujące najwyższe wyniki metryk detekcji [C01] [C01]. Zbadałem wpływ kolejnych faz przetwarzania danych na wyniki detekcji w NIDS [C01].

Jak zauważono w [22], optymalizacja hiperparametrów jest niezbędna do znalezienia najlepszego dopasowania dla danego zbioru danych. Istnieje wiele metod optymalizacji hiperparametrów, w tym metody wykorzystujące podejścia ewolucyjne, optymalizację rojem cząstek, czy losowe przeszukiwanie [22].

Aby osiągnąć wyżej wymienione cele zastosowałem gridsearch, który przeszukuje przestrzeń hiperparametrów, wyczerpująco generując kandydatów na ustawienia ANN i klasyfikując je według wybranej funkcji oceny.

4.4.3.1. Dane wykorzystane do uczenia ML/ANN w NIDS - format NetFlow

Zadania ML można podzielić na uczenie nadzorowane, uczenie nienadzorowane oraz uczenie przez wzmacnianie [23].

Główną ideą nadzorowanego ML jest wykorzystanie danych do tworzenia modeli, które potrafią określić coś *Y* na podstawie pewnych informacji *X*. Aby wykonać swoje zadanie, algorytm ML podsumowuje *X* w formie *T*. Najlepsze podsumowanie *T* powinno dostarczać nam jak najwięcej informacji o *Y*, jednocześnie zachowując jedynie istotne części *X* [24]. Skuteczność uczenia się algorytmu ML zależy od dwóch aspektów: złożoności modelu oraz próbek treningowych – zarówno ich ilości, jak i jakości zawartych w nich informacji. Metody ML typowo potrzebują odpowiedniej liczby próbek do trenowania modelu, aby uzyskać dobre wyniki.

Prace badawcze zawarte w tym cyklu wykorzystują uczenie nadzorowane i w związku z tym zajmują się rozwiązywaniem problemów wynikających z wykorzystania podejścia nadzorowanego w zadaniu wykrywania ataków sieciowych.

Charakterystyka uzyskanych danych stanowi jedno z wyzwań, które należy rozwiązać, w celu wykorzystania ML w systemach NIDS. Badacze często korzystają z zestawów danych referencyjnych (benchmarków), które zawierają cechy niemożliwe do uzyskania w czasie rzeczywistym [25][26].

Cechy wykorzystujące przepływy (flow-based features, NetFlow) dostarczają przydatnego wglądu w zachowanie sieci [27][28] i pozwalają na działanie w czasie rzeczywistym.

Według [29] zaletami wykorzystania danych przepływowych są ich przydatność w sieciach o dużej przepustowości, fakt, że są już szeroko wdrożone i wykorzystywane przez analityków bezpieczeństwa, oraz to, że cechy przypominające przepływy zużywają mniej przepustowości dzięki agregacji. Ponadto, ponieważ podejście to wykorzystuje agregacje, jest mniej wrażliwe na kwestie prywatności w porównaniu z głęboką inspekcją pakietów, co jest szczególnie pożądane z perspektywy przepisów dotyczących prywatności, takich jak RODO [30]. Niektóre z dostępnych formatów zapewniających te korzyści, wraz z narzędziami i eksporterami do przechwytywania przepływów, opisano w [31].

ML i ANN stają się coraz ważniejszym elementem w walce z cyberatakami w sieciach komputerowych, dzięki ich zdolności do rozpoznawania skomplikowanych wzorców. Wykorzystywanie ML i ANN w NIDS to zbiór wyzwań w którym każdy krok przetwarzania danych stanowi pewien problem, a samo wykorzystanie sztucznej inteligencji wiąże się z wyzwaniami takimi jak wyjaśnialność (explainability of AI - xAI), czy samo bezpieczeństwo AI [C01] [C02].

Niektóre z problemów zidentyfikowanych w [C01] i [C02], jak i w wyniku późniejszych badań, to:

- Wpływ hiperparametrów na ML/ANN wykorzystywane w NIDS
- Wpływ różnych elementów przetwarzania danych w procesie ML na wyniki klasyfikacji
- Wysoki koszt i trudność pozyskiwania danych do NIDS
- Brak transferowalności wyuczonych detektorów ML/ANN w NIDS z jednej sieci na drugą
- Trudności ML/ANN z wykrywaniem ataków, które nie są ujęte w zbiorze treningowym
- Brak łatwego sposobu zrozumienia mechanizmów klasyfikacji ataków w ML/ANN wykorzystywanych w NIDS
- Zagadnienie wykorzystania metod wyjaśnialności AI w kontekście bezpieczeństwa sieciowego (i związanych z tym niebezpieczeństw)
- Trudności w ocenie trafności wytłumaczeń mechanizmów klasyfikacji ataków w ML/ANN proponowanych przez mechanizmy wyjaśnialności AI
- Brak metod umożliwiających porównanie wyników metod wyjaśnialności AI do punktu odniesienia (ground truth)

W związku ze zidentyfikowanymi problemami, sformułowałem trzy ogólne kategorie problematyczne wykorzystania ML w NIDS:

- Podnoszenie wyników klasyfikacyjnych ML/ANN w NIDS
- Rozwiązywanie problemów niedostępności danych w NIDS
- Rozwiązywanie problemów z wyjaśnialnością AI w NIDS

W kolejnych sekcjach opiszę zaproponowane przeze mnie metody oraz towarzyszące im publikacje, wpisujące się w opisana tematykę i będące podstawą osiągnięcia naukowego.

4.4.3.2. Kodowanie kategorii w wykrywaniu ataków sieciowych – zarys problemu

Niektóre cechy danych nie są ilościowe i są reprezentowane w formie nominalnej, często czytelnej dla człowieka. W zestawie cech wykorzystującym NetFlow występują

wartości kategoryczne, które przenoszą ważne informacje. Te cechy zawierają informacje o, na przykład, używanym protokole lub wykorzystywanym porcie docelowym. Inny niż oczekiwany numer portu docelowego lub nietypowe wykorzystanie protokołu może wskazywać na nadchodzący atak. Wprowadza to pojęcie właściwego kodowania tych wartości w dziedzinie NIDS.

Temat kodowania cech kategorycznych nie jest często poruszany w najnowszych badaniach IDS, jest jednak tematem znanym w dziedzinie badań ML, także w innych zastosowaniach.

Autorzy [32] oceniają sześć różnych schematów kodowania cech kategorycznych w dziedzinie przewidywania chorób serca. Autorzy zauważają, jak ważne jest kodowanie cech kategorycznych jako krok przetwarzania wstępnego w ML, gdyż adekwatne kodowanie może podkreślić ekspresywność zakodowanej cechy. W [33] oceniano siedem schematów kodowania, eksperymenty wykonano za pomocą ANN w zadaniu klasyfikacji z użyciem zestawu danych oceny samochodów.

W swoim przeglądzie dotyczącym wykorzystania wartości kategorycznych w ANN, autorzy [34] identyfikują szeroki zakres różnych schematów kodowania. Autorzy dzielą schematy na trzy kategorie: kodowanie ustalone, kodowanie automatyczne i kodowanie algorytmiczne. Niektóre z opisanych technik to liczenie kodów, kodowanie jednoznakowe, kodowanie etykiet, kodowanie opuszczające jedno wyjście oraz kodowanie przez hashowanie.

W [35] oceniono wpływ kodowania cech kategorycznych. Testowane schematy kodowania to 'Kodowanie Etykiet', 'Kodowanie Jednoznakowe' oraz 'Kodowanie Binarne'. Metody te są testowane z sześcioma różnymi algorytmami uczenia maszynowego.

4.4.3.3. Kodowanie kategorii w wykrywaniu ataków sieciowych – kontrybucja

W swoich pracach zaproponowałem wykorzystanie kodowania danych kategorycznych przez frequency encoding oraz zbadałem wpływ sposobu kodowania danych kategorycznych w NIDS wykorzystujących algorytmy ML [C03]

Głównym celem jest ocena różnych strategii kodowania, w celu określenia ich wpływu na dokładność modeli ML oraz możliwość zastosowania tych cech w czasie rzeczywistym. Wykorzystując algorytmy Random Forest (RF), k-Nearest Neighbours (k-NN) i ANN, ewaluuję skuteczność kodowania etykiet (label encoding), kodowania „1 z n” (One Hot

Encoding) oraz kodowania częstotliwościowego (frequency encoding) na rzeczywistym zbiorze danych NetFlow.

Kodowanie częstotliwości zastępuje kategorie częstotliwością ich występowania w zbiorze uczącym, zachowując niższą wymiarowość niż OHE. Zapewnia kompaktową reprezentację danych, może jednak wpłynąć na preferencje (bias) modelu w stronę kategorii częściej występujących, co jest ważnym aspektem do rozważenia przy jego zastosowaniu.

Wyniki eksperymentalne pokazują, że kodowanie częstotliwości konsekwentnie przewyższało inne metody pod względem efektywności obliczeniowej i dokładności detekcji. Zachowując niski koszt obliczeniowy, a jednocześnie skutecznie ujmując istotne informacje z danych kategorycznych, kodowanie częstotliwości lepiej niż inne metody wspiera wymagania przetwarzania czasu rzeczywistego w NIDS; widać to w Tabela 1, Tabela 2 i Tabela 3.

Tabela 1. Wyniki stratyfikowanej walidacji krzyżowej 10-krotnej, Random Forest

Encoding:	**Frequency**	**Label**	**Dropping**	**One-Hot**
ACC Max:	0.979	0.966	0.966	0.969
ACC Min:	0.979	0.964	0.964	0.968
ACC Mean:	0.979	0.965	0.965	0.969
BACC Max:	0.778	0.719	0.719	0.666
BACC Min:	0.696	0.639	0.639	0.642
BACC Mean:	0.740	0.683	0.683	0.650
MCC Max:	0.906	0.841	0.841	0.860
MCC Min:	0.905	0.834	0.834	0.853
MCC Mean:	0.905	0.837	0.837	0.854

Tabela 2. Wyniki stratyfikowanej walidacji krzyżowej 10-krotnej, ANN

Encoding:	**Frequency**	**Label**	**Dropping**
ACC Max:	0.978	0.976	0.940
ACC Min:	0.977	0.974	0.936
ACC Mean:	0.978	0.976	0.938
BACC Max:	0.690	0.688	0.440
BACC Min:	0.535	0.577	0.340
BACC Mean:	0.615	0.658	0.398
MCC Max:	0.902	0.894	0.700
MCC Min:	0.897	0.884	0.673
MCC Mean:	0.900	0.891	0.686

Tabela 3. Wyniki stratyfikowanej walidacji krzyżowej 10-krotnej, k-NN

Encoding:	**Frequency**	**Label**	**Dropping**

ACC Max:	0.980	0.980	0.966
ACC Min:	0.969	0.969	0.955
ACC Mean:	0.977	0.978	0.961
BACC Max:	0.780	0.779	0.718
BACC Min:	0.702	0.700	0.652
BACC Mean:	0.734	0.736	0.682
MCC Max:	0.908	0.908	0.843
MCC Min:	0.862	0.861	0.790
MCC Mean:	0.898	0.898	0.820

Implikacje tych wyników są znaczące dla systemów NIDS wykorzystujących ML. Kodowanie częstotliwości nie tylko zwiększa osiągnięcia detekcyjne modelu, ale również zapewnia, że systemy są zdolne do efektywnego działania w czasie rzeczywistym, co jest ważnym wymogiem w detekcji ataków. Osiąga optymalny balans między zachowaniem treści informacyjnej a wymaganiami obliczeniowymi, co jest niezbędne do utrzymania dokładności NIDS działających w czasie rzeczywistym.

4.4.4. Problem przesunięcia wdrożeniowego i kosztownych danych

Badania nad systemami wykrywania ataków sieciowych wykorzystującymi ML cieszą się dużym zainteresowaniem w społeczności naukowej oraz w praktyce przemysłowej [36][37][38]. Mimo to, ich wykorzystanie nadal jest dość problematyczne. Jedną z głównych przeszkód w szerokim wdrożeniu systemów NIDS wykorzystujących ML jest konieczność zbierania oznakowanych danych w obrębie chronionej sieci [39], ponieważ komponenty ML muszą nauczyć się specyficznych rozkładów danych. Zbieranie tego typu danych wymaga specjalistycznej wiedzy i odpowiednich zasobów [9].

Kolejnym niewygodnym założeniem wdrożonych komponentów ML jest przypuszczenie, że przyszłe dane będą miały taki sam rozkład jak dane użyte do trenowania modelu [40]. Założenie to jest naruszone w kontekście NIDS, gdyż charakterystyka ruchu zmienia się z biegiem czasu.

4.4.4.1. Brak odpowiednich danych potrzebnych do wykrywania ataków w nowej sieci komputerowej – zarys problemu

Brak odpowiednich danych stanowi przeszkodę w formułowaniu skutecznego modelu za pomocą standardowego procesu ML. Z tego samego powodu modele trenowane na różnej dystrybucji danych osiągają gorsze wyniki, gdy są testowane na danych pochodzących z pokrewnej, ale innej dziedziny. W kontekście NIDS ten efekt stanowi problem, gdy detektor wykorzystujący ML osiągający adekwatne wyniki wykrywania ataków próbuje się przenieść na nową sieć, ponieważ w nowej sieci komponent ten działa gorzej.

Istnieje szereg technik, które pomagają algorytmom uczącym się zredukować różnice między dziedzinami źródłową a docelową [41][42][43][44][45].

Jednym z możliwych sposobów na zmniejszenie kosztów koniecznych do wdrożenia systemów NIDS wykorzystujących ML w nowej sieci jest wykorzystanie wiedzy uzyskanej z zestawów danych referencyjnych i innych wcześniej zebranych danych. Można to osiągnąć za pomocą technik uczenia transferowego (transfer learning) [40].

Uczenie transferowe jest zdolne do wykorzystywania już istniejącej wiedzy, która została wygenerowana w innych, podobnych lub powiązanych przypadkach w celu uczenia modelu [41]; w założeniu, dzięki uczeniu transferowemu można nauczyć model wykonywać podobne zadanie na nowych danych przy użyciu znacznie mniejszego zestawu danych [46][42]. Uczenie transferowe jest szczególnie pomocne, gdy w dziedzinie źródłowej istnieje obfitość oznakowanych danych, ale w dziedzinie docelowej danych jest niewystarczająco [44].

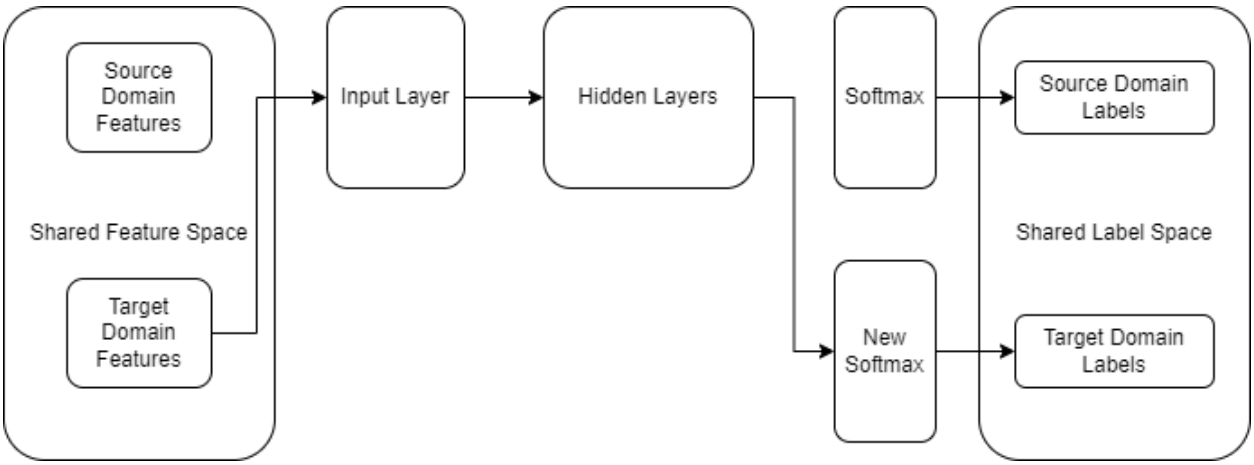
Zbieranie i oznaczanie danych do zastosowań w ML w systemach NIDS, jest procesem kosztownym i czasochłonnym, który wymaga intensywnych zasobów ludzkich i technicznych. Wykonywanie i oznaczanie ataków w sieci, którą ma się chronić jest mocno utrudnione, jeśli nie niemożliwe. Mimo to, wykorzystanie ML w NIDS zapewnia wysoką skuteczność w wykrywaniu znanych rodzajów ataków.

4.4.4.2. Proponowana metoda: minimalizacja ilości danych potrzebnych do wykrywania ataków w nowej sieci komputerowej

Zbadałem zagadnienie przesunięcia wdrożeniowego w NIDS i zaproponowałem metody uczenia transferowego w celu adaptacji modeli ML do nowych środowisk sieciowych. Zaproponowałem metodę pozwalającą na minimalizację potrzebnych danych przy wykorzystywaniu NIDS działających przy pomocy ML, poprzez stosowanie technik uczenia transferowego. Zaproponowałem metodę w której, jak dowiedziono eksperymentalnie, przy użyciu minimalnej ilości danych z domeny docelowej można znacznie poprawić skuteczność sieci neuronowych [C04].

W eksperymencie oszacowano minimalną ilość danych potrzebną do dostrojenia wcześniej wytrenowanej ANN, tak aby zdolność wykrywania ataków nowej ANN zastosowanej w nowym środowisku pozostała adekwatna. Zbadana technika polega na usunięciu warstwy softmax z w pełni wytrenowanego klasyfikatora, zamrożeniu jego warstw oraz szkoleniu nowej warstwy softmax na danych z dziedziny docelowej.

Zadanie klasyfikacji wykorzystane w tym badaniu rozróżnia między dwoma etykietami docelowymi: 'Benign' (bezpieczny) i 'Attack' (atak).



Rysunek 2. Wykorzystanie wiedzy ANN z dziedziny źródłowej do minimalizacji potrzebnych danych

Jak pokazano na Rysunek 2 wykorzystywana jest sieć wstępnie nauczona na danych z dziedziny źródłowej. Przy stałej przestrzeni cech i przestrzeni etykiet można założyć, że wystąpi znaczące pokrycie parametrów sieci neuronowej nauczanej do działania w dziedzinie źródłowej i docelowej. Jednakże samo użycie wstępnie nauczonej sieci w niezmienionej formie prowadzi do wyników poniżej oczekiwań.

Kolejnym krokiem w testowaniu tego założenia jest użycie wstępnie nauczonej sieci jako ekstraktora cech, mającego na celu zminimalizowanie liczby punktów danych potrzebnych do dostrojenia sieci. Warstwa softmax wstępnie nauczonej sieci została usunięta, ukryte warstwy zamrożone, a nowa warstwa softmax została ponownie połączona z przedostatnią warstwą ANN. Nowa sieć ANN została poddana treningowi (tylko warstwa softmax) na zestawie danych z dziedziny docelowej.

Tabela 4. Wyniki ANN na zbiorze testowym domeny źródłowej

	precision	**recall**	**f1-score**
Attack	0.95	1.00	0.98
Benign	1.00	0.90	0.95
BACC	0.95		
MCC	0.92		

Tabela 5. Wyniki ANN na zbiorze testowym domeny docelowej

	precision	**recall**	**f1-score**
Attack	0.10	0.00	0.01
Benign	0.09	0.78	0.16
BACC	0.39		
MCC	-0.42		

Tabela 6. Wyniki ANN na zbiorze testowym domeny docelowej, dostrojonej dwoma próbkami

	precision	recall	f1-score
Attack	0.81	0.36	0.50
Benign	0.05	0.31	0.09
BACC	0.33		
MCC	-0.21		

Tabela 7. Wyniki ANN na zbiorze testowym domeny docelowej, dostrojonej dwudziestoma próbkami

	precision	recall	f1-score
Attack	0.99	0.95	0.96
Benign	0.66	0.88	0.75
BACC	0.91		
MCC	0.73		

Tabela 8. Wyniki ANN na zbiorze testowym domeny docelowej, dostrojonej dwustoma próbkami

	precision	recall	f1-score
Attack	1.00	0.98	0.99
Benign	0.86	0.97	0.91
BACC	0.97		
MCC	0.90		

Tabela 9. Wyniki ANN na zbiorze testowym domeny docelowej, dostrojonej całym zbiorem treningowym domeny docelowej

	precision	recall	f1-score
Attack	1.00	0.98	0.99
Benign	0.86	0.97	0.91
BACC	0.97		
MCC	0.90		

Jak można stwierdzić na podstawie analizy tabel, wstępnie nauczona sieć osiągnęła dobre wyniki na zestawie testowym dziedziny źródłowej (Tabela 4), ale nie osiągnęła zadowalających wyników na zestawie testowym w dziedzinie docelowej (Tabela 5). Ponowne nauczenie nowej warstwy softmax tylko z dwoma próbkami - jedną z klasy 'Benign' i jedną z klasy 'Attack' przez 1000 epok - nie poprawiło metryk klasyfikacji (Tabela 6), osiągając Balanced Accuracy (BACC) na poziomie tylko 0,33 i Matthews correlation coefficient (MCC) -0,21. Jednakże zwiększenie liczby próbek do tylko dziesięciu z każdej klasy - przy utrzymaniu liczby epok i niższym współczynniku uczenia - znacząco poprawiło zdolność detekcji klasyfikatora (Tabela 7), osiągając 0,91 BACC i 0,73 MCC. Zwiększenie liczby oznakowanych próbek do 100 na każdą klasę jeszcze bardziej poprawiło wyniki

(Tabela 8), osiągając BACC 0,97 i MCC 0,90, a także zwiększając Recall klasy ataku do 0,98. Wyniki te były prawie identyczne z wynikami sieci nauczonych z pełnym zestawem treningowym dziedziny docelowej, który zawierał 63768 próbek (Tabela 9).

Te wyniki sugerują, że uczenie transferowe jest ważnym narzędziem do łagodzenia efektów przesunięć wdrożeniowych w NIDS. Dzięki efektywnemu wykorzystaniu wcześniej wytrenowanych modeli i małej ilości danych z domeny docelowej, organizacje mogą wdrażać skuteczne systemy wykrywania ataków, które dostosowują się do nowych środowisk bez konieczności intensywnego ponownego trenowania lub zbierania ogromnych ilości danych treningowych.

4.4.5. Problem niewykrywania ataków niezawartych w zbiorze treningowym

Możliwość zminimalizowania ilości danych do trenowania klasyfikatora ML zaprezentowana w [C04] nie sprawia jednak, że problem pozyskiwania znika całkowicie - nadal istnieje potrzeba symulacji i oznaczania nawet małej liczby ataków w sieci docelowej. Pokrewnym problemem jest zagadnienie wykrywania ataków, które nie znajdują się w zbiorze treningowym. Problemy te prowadzą to do następnej zaproponowanej metody [C05], która pozwala na wykrywanie ataków wcześniej niewidzianych.

4.4.5.1. Uczenie z małą ilością danych i kontrastywna funkcja straty (contrast loss) – zarys problematyki

Nazwa „Sieć neuronowa typu Siamese" (SNN) odnosi się do specjalizowanego rodzaju ANN. Termin „Siamese" (syjamska) pochodzi od struktury sieci, która wykorzystuje dwie (lub więcej) identyczne kopie tej samej sieci do przetwarzania próbek wejściowych. Głównym celem SNN jest nauczenie rozpoznawania podobieństw i różnic między próbkami wejściowymi.

SNN są zdolne do rozróżniania między dwoma klasami, nawet gdy dostępna jest ograniczona liczba próbek z nowej klasy. Wynik działania SNN można postrzegać jako semantyczne podobieństwo między danymi wejściowymi [47]. Do trenowania SNN używa się zbioru par próbek wejściowych, z których każda para zawiera próbki do siebie podobne (należące do tej samej klasy), albo różne (należące do różnych klas). Sieć uczy się budować reprezentację (embedding) tak, by minimalizować odległość między podobnymi parami i maksymalizować odległość między różnymi parami. W konsekwencji SNN staje się zdolna do oceny podobieństwa między nowymi, wcześniej nieobserwowanymi próbkami na podstawie ich reprezentacji. Ta właściwość pozwala na wykorzystanie SNN w scenariuszach z małą ilością danych (few-shot learning) [48].

W podejściach uczenia nadzorowanego do ANN używana funkcja straty to zazwyczaj albo binarna, albo kategoryczna entropia krzyżowa (znana również jako strata logistyczna lub wielomianowa strata logistyczna). Funkcja Straty (loss function) ta jest używana do

obliczania reprezentacji różnic między etykietami próbek treningowych a ich klasyfikacjami wykonywanymi przez uczoną ANN [49].

Kontrastywna funkcja straty generuje miarę odległości między próbkami przetwarzanymi przez obie połówki SNN, minimalizując stratę, gdy podobne wejścia są blisko siebie, a różne wejścia znajdują się w dużej odległości od siebie, co pozwala na modelowanie relacji między podobnymi i różnymi parami wejściowymi. Powoduje to, że SNN formułuje reprezentację wejść w przestrzeni cech. Reprezentacje te można następnie wykorzystać aby zmierzyć odległość między znanymi klasami a próbką wejściową lub porównać próbkę wejściową z klasą, która nie była uwzględniona w procesie treningu SNN. Używanie par próbek wejściowych, które wymaga kontrastywna funkcja straty, daje szereg unikalnych korzyści, takich jak polepszona generalizacja uczenia, zwiększona odporność na nierównowagę klas oraz zdolność radzenia sobie ze zmienną liczbą klas [50][51][52][53].

4.4.5.2. Zaproponowane podejście: nowa metoda uczenia dla kontrastywnej funkcji straty polegająca na próbkowaniu próbek granicznych.

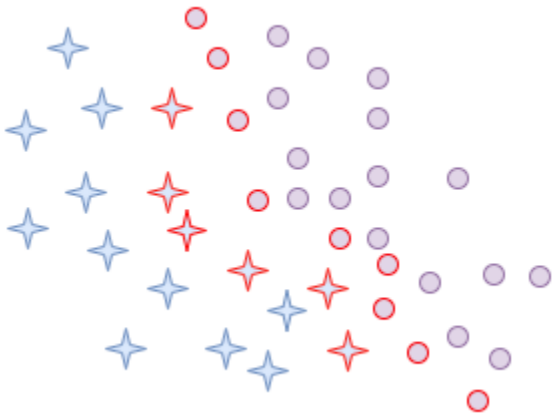
Zaproponowałem metodę rozwoju ANN typu Siamese poprzez wprowadzenie nowej techniki próbkowania danych treningowych wykorzystującej ekstrakcję granic. Metoda ta ma na celu zwiększenie skuteczności NIDS w identyfikowaniu wcześniej niewidzianych ataków [C05].

W badaniu [C05] przedstawiłem technikę próbkowania z ekstrakcją granic, nowe podejście, które wykorzystuje algorytm k-NN do efektywniejszego trenowania SNN poprzez skoncentrowanie się na najbardziej informatywnych przykładach, dobranych specjalnie do uczenia kontrastowego wykorzystywanego w sieciach SNN. Są to próbki, które znajdują się blisko granicy decyzji między klasami. Takie ukierunkowane podejście zwiększa zdolności wykrywania nowych ataków przez SNN.

Aby sformułować NIDS z wykorzystaniem SNN, na początku formułowany jest zestaw treningowy składający się z par pozytywnych i negatywnych. Pary negatywne generowane są przez połączenie próbek ze wszystkich różnych klas. Pary pozytywne są formułowane przez połączenie dwóch próbek z tej samej klasy. Celem treningu jest sformułowanie reprezentacji, która zbliża do siebie reprezentacje próbek z tych samych klas, a reprezentacje próbek z różnych klas oddala od siebie. W kontekście NIDS, do trenowania SNN używany jest zestaw danych NetFlow. Sformułowany detektor może być używany w dwóch trybach: klasyfikacji binarnej i klasyfikacji wieloklasowej. W tym badaniu rozważana była klasyfikacja binarna. Próbkę referencyjną która jest przedstawiana jednej z bliźniaczych sieci neuronowych do porównania z drugą, nieznaną próbką nazywamy

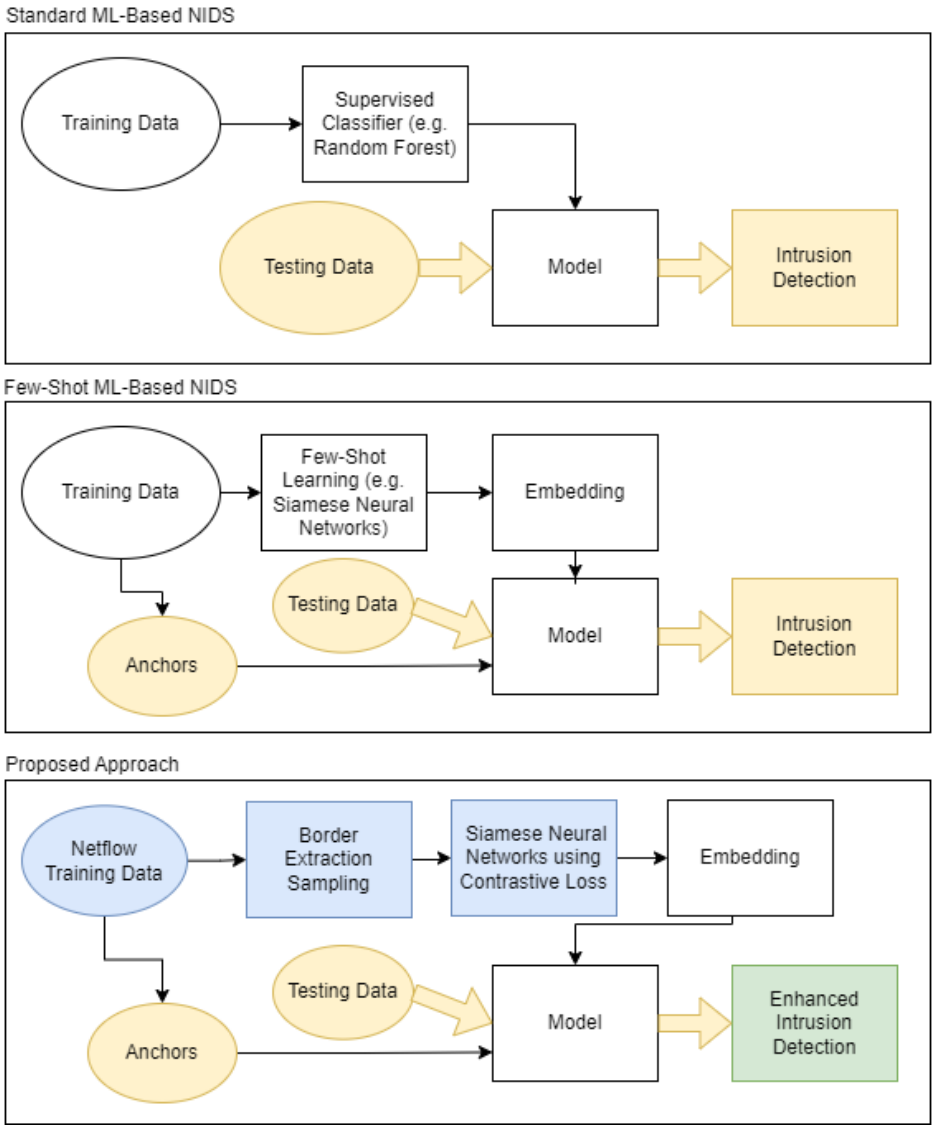
„kotwicą”. Klasyfikator binarny można uzyskać, używając klasy „benign” jako kotwicy w czasie testowania, a następnie porównując klasy ataków z kotwicą.

Zaproponowałem podejście wykorzystuje zdolność algorytmu k-NN do identyfikacji próbek sąsiednich oraz możliwości formułowania reprezentacji za pomocą kontrastywnej funkcji straty używanej w SNN.



Rysunek 3. Granice klas wyodrębnione przy użyciu algorytmu k-NN (czerwony kontur). Próbki bliskie jakiejkolwiek innej klasie są włączone do zestawu treningowego.

Algorytm k-NN z k=5 jest używany do znalezienia wszystkich próbek w zbiorze treningowym, które mają przynajmniej jedną próbkę innej klasy w sąsiedztwie. Jest to zilustrowane na Rysunek 3. W ten sposób identyfikowane są wszystkie próbki, które znajdują się na granicach klas w przestrzeni cech zestawu treningowego. Ponieważ te próbki są stosunkowo blisko siebie w przestrzeni cech, wykorzystanie tych próbek w procesie trenowania SNN z użyciem kontrastywnej funkcji straty zmusza SNN do stworzenia reprezentacji, które oddalając od siebie próbki pochodzące z różnych klas i jednocześnie zbliżając do siebie próbki pochodzące z tej samej klasy, uwypukla granice między klasami.



Rysunek 4. Proponowane podejście zestawione z podejściem standardowym, i uczenia z małą ilością próbek (Few-Shot).

Jak przedstawiono na Rysunek 4, próbki ruchu z zestawu treningowego są najpierw poddawane procedurze ekstrakcji granic. Przetworzony zestaw treningowy jest następnie używany do trenowania SNN-ów w celu sformułowania reprezentacji ruchu, która może być następnie używana do wykrywania ataków - znanych i nieznanych.

Wyniki badania (Tab. Tabela 10 i Tabela 11) wskazują, że SNN trenowane z użyciem próbkowania z ekstrakcją granic wykazują lepsze zdolności wykrywania wcześniej niespotkanych przez SNN ataków sieciowych (Only Border, Unknown Attacks) w porównaniu z tradycyjnymi podejściami, takimi jak RF. Na znanych atakach SNN wytrenowany proponowaną metodą odznaczył się najwyższą zdolnością wykrywania ataków pośród eksperymentów SNN (Only Border, Known Attacks, Recall).

Tabela 10. Metryki wydajności dla klasyfikatorów wykorzystujących sieci neuronowe typu Siamese.

Problem	Model	Type	Precision	Recall	f1-score
Border-Removed, Known Attacks	SNN	Attack	1.00	0.79	0.88
Border-Removed, Known Attacks	SNN	Benign	0.28	0.97	0.43
Border-Removed, Unknown Attacks	SNN	Attack	0.77	0.81	0.79
Border-Removed, Unknown Attacks	SNN	Benign	0.80	0.76	0.78
Only Border, Known Attacks	SNN	Attack	0.97	0.93	0.95
Only Border, Known Attacks	SNN	Benign	0.43	0.68	0.53
Only Border, Unknown Attacks	SNN	Attack	0.77	0.81	0.79
Only Border, Unknown Attacks	SNN	Benign	0.80	0.76	0.78
Whole set, Known Attacks	SNN	Attack	1.00	0.69	0.82
Whole set, Known Attacks	SNN	Benign	0.24	0.97	0.38
Whole set, Unknown Attacks	SNN	Attack	0.77	0.81	0.79
Whole set, Unknown Attacks	SNN	Benign	0.80	0.76	0.78

Tabela 11. Metryki wydajności dla klasyfikatorów wykorzystujących algorytm Random Forest.

Problem	Model	Type	Precision	Recall	f1-score
Border-Removed, Known Attacks	RF	Attack	0.81	0.86	0.83
Border-Removed, Known Attacks	RF	Benign	0.98	1.00	0.99
Border-Removed, Unknown Attacks	RF	Benign	0.50	0.77	0.61
Border-Removed, Unknown Attacks	RF	Attack	0.00	0.00	0.00
Only Border, Known Attacks	RF	Attack	0.05	0.03	0.04
Only Border, Known Attacks	RF	Benign	0.87	0.77	0.82
Only Border, Unknown Attacks	RF	Benign	0.50	0.77	0.61
Only Border, Unknown Attacks	RF	Attack	0.00	0.00	0.00
Whole set, Known Attacks	RF	Attack	0.81	0.86	0.83
Whole set, Known Attacks	RF	Benign	0.98	1.00	0.99
Whole set, Unknown Attacks	RF	Benign	0.50	0.77	0.61
Whole set, Unknown Attacks	RF	Attack	0.00	0.00	0.00

Wprowadzenie próbkowania z ekstrakcją granic do treningu SNN dla NIDS stanowi znaczący postęp w dziedzinie cyberbezpieczeństwa, pozwalając na wykrycie nieznanych ataków. Skupiając się na najbardziej informatywnych próbkach, SNN nie tylko poprawiają swoją dokładność wykrywania, ale także działają efektywnie w środowiskach czasu rzeczywistego. Podejście to jest szczególnie wartościowe w scenariuszach, w których ciągle pojawiają się nowe typy ataków sieciowych, co wymaga systemów zdolnych do szybkiego dostosowania się do nowych zagrożeń bez konieczności ponownego trenowania.

Badanie to podkreśla potencjał zaawansowanych technik, takich jak SNN, do znacznego zwiększenia możliwości systemów wykrywania intruzji w sieci.

Możliwość skutecznego wykrywania ataków niezawartych w zbiorze treningowym jest bardzo ważna dla realnego zastosowania ML w NIDS.

4.4.6. Problem braku wyjaśnialności algorytmów ML

Wiele metod AI/ML funkcjonuje jako czarne skrzynki (black box), co oznacza, że ich wewnętrzny sposób działania nie jest łatwy do zrozumienia. Rodzi to obawy etyczne dotyczące zastosowań tych modeli w wielu dziedzinach, tworzy problemy z zaufaniem i staje się barierą dla wielu korzystnych zastosowań [54], również w kontekście cyberbezpieczeństwa.

Powoduje to powstanie potrzeby na metody interpretowalności, które ułatwiłyby operatorowi zrozumienie procesu decyzyjnego modelu AI. Metody wyjaśnialności AI (xAI - explainability of AI) są potrzebne, aby zapewnić etyczne użycie AI i wzmocnić zaufanie użytkowników. xAI jest obecnie intensywnie badanym obszarem, w którym ciągle pojawiają się nowe podejścia, a niektóre metody stopniowo zyskują na znaczeniu [55].

Pojawienie się metod xAI pozwala zajrzeć do wnętrza czarnej skrzynki; jednak, w przeciwieństwie do oceny wydajności modelu nadzorowanego, nie ma jeszcze ustalonych metod na określenie, czy wyjaśnienia proponowane przez metody xAI są trafne i prawdziwe.

4.4.6.1. Problem oceny algorytmów wyjaśnialności

Jak wyjaśnia [56], potrzeba metryk oceniających techniki xAI jest dobrze określona, ale nie ma prostego narzędzia pozwalającego na osiągnięcie kwantyfikacji ważności czy jakości xAI. Co więcej, ocena xAI wymaga solidnych podstaw naukowych oraz uwzględnienia potrzeb użytkownika końcowego, a także faktu, że metody xAI wywodzą się z wielu różnych dziedzin wiedzy [57].

Jedną z najbardziej znaczących metod xAI jest SHapley Additive exPlanations (SHAP) [58], podejście zaczerpnięte z teorii gier, zapewniające stosunkowo łatwe do zrozumienia i interpretacji spojrzenie na to, jak model poddany ocenie przydziela wagę cechom wejściowym.

4.4.6.2. Propozycja mierzenia stabilności xAI przy pomocy korelacji

Zaproponowałem pomiar stabilności wyjaśnień jako jedną z możliwych metod oceny xAI [C06].

Podejście to wydaje się rozsądne, ponieważ:

- Stabilna interpretacja zapewnia spójne interpretacje na wielu przykładach wyników modelu. Ta spójność interpretacji jest ważna w budowaniu zaufania użytkowników, gdyż wyjaśnienia, które zmieniają się bez wyraźnego powodu, mogą prowadzić do zamieszania.

- Rzeczywiste dane często zawierają szumy i brakujące wartości; dobre wyjaśnienie powinno pozostać stabilne lub zmieniać się w sposób przewidywalny. Mała perturbacja nie powinna powodować drastycznych zmian w wyjaśnieniu.

- Wnioski wynikające ze stabilnych wyjaśnień można zastosować do szerszej gamy przypadków poza konkretnymi instancjami, co zapewnia lepszą generalizację.

- Stabilne wyjaśnienia pozwalają na większą pewność w procesie decyzyjnym modelu AI. Jeśli wyjaśnienia ulegają dramatycznym zmianom na skutek niewielkich perturbacji, może to podważyć tę pewność.

- Poza dziedziną AI, dobre wyjaśnienia zazwyczaj są stabilne. Na przykład, lekarz konsekwentnie wyjaśnia podobne przypadki medyczne podobnym uzasadnieniem, bank konsekwentnie wyjaśnia podobne decyzje dotyczące zatwierdzeń kredytowych itd. Jako ludzie uważamy te wyjaśnienia za godne zaufania, ponieważ są zgodne z naszymi oczekiwaniami, i oczekujemy, że wyjaśnienia będą stabilne w podobnych przypadkach.

Zrozumienie stabilności wyjaśnień w różnych scenariuszach danych jest zatem niezbędne dla wykorzystania xAI w praktyce. Celem badania przedstawionego w [C06] jest ewaluacja stabilności wartości SHAP pod wpływem trzech rodzajów perturbacji danych: przetasowania cech, brakujących wartości oraz wprowadzenia szumu Gaussa, na różnych poziomach intensywności. Podejście obejmuje zastosowanie tych perturbacji do czterech różnorodnych zestawów danych (Breast Cancer, Iris, Wine i CIC-IDS2018) i ocenę zmian wartości SHAP przy użyciu różnych metryk, w tym miar korelacji i błędu, a także testów statystycznych.

Zaproponowałem nowy sposób badania stabilności metod xAI na przykładzie SHapley Additive exPlanations (SHAP) w celu konstrukcji metryk wartości wyjaśnień proponowanych przez xAI. W tym celu badam odporność SHAP na różne rodzaje zakłóceń w danych [C06].

Badanie ma na celu określenie stabilności xAI w scenariuszach coraz bardziej niestabilnych danych, biorąc pod uwagę różne rodzaje szumów - jak przetasowanie cech, brakujące wartości oraz szum Gaussa. Metody typu SHAP kwantyfikują wkład poszczególnych cech w przewidywania modeli ML. Zrozumienie wpływu takich zakłóceń jest ważne dla oceny niezawodności narzędzi xAI w praktycznych zastosowaniach, gdzie integralność danych może się znacznie różnić.

4.4.6.3. Relacja do innych metryk stabilności xAI

Do najważniejszych prac nad stabilnością jako metryką jakości w xAI należą [59] i [60]. W [59] autorzy kwantyfikują stabilność wyjaśnienia, używając punktowej, lokalnej ciągłości Lipschitza wykorzystującej sąsiedztwo. Sami autorzy zauważają, że przed takim podejściem stoi szereg wyzwań, w tym fakt, że większość metod interpretowalności nie jest różniczkowalna od początku do końca, a szacowanie stałej L jest kosztowne obliczeniowo. W [60] autorzy zauważają, że [59] zakłada, że model zachowuje się tak samo przy różnych danych wejściowych, co może być prawdą, gdy model sam w sobie jest stabilny i odporny, ale nie można tego uważać za oczywistość. W celu zaradzenia temu autorzy proponują między innymi Względną Stabilność Wejściową (Relative Input Stability), która mierzy względną odległość między wyjaśnieniami względem względnej różnicy między danymi wejściowymi. Te metryki znalazły swoje zastosowanie w [61] i [62]. W [63], autorzy również używają współczynnika Lipschitza jako miary stabilności. W naszej pracy, w przeciwieństwie do wspomnianych opracowań, unikamy używania współczynnika Lipschitza, z wielu powodów.

Lipschitz oferuje maksymalne ograniczenie tego jak funkcja może się zmienić względem jednostkowej zmiany w jej parametrach wejściowych, podczas gdy inne miary, użyte w propozycji oddają różne aspekty stabilności. Na przykład, Korelacja Pearsona ocenia liniowy związek między dwoma zmiennymi, Korelacja Rang Spearmana i Tau Kendalla oceniają związek monotoniczny, a Podobieństwo Kosinusowe ocenia kąt między dwoma wektorami w przestrzeni wielowymiarowej. Chociaż wcześniejsze metryki stabilności zakładają liniowy związek między zmianami, Tau Kendalla i Korelacja Rang Spearmana mogą uchwycić związki nieliniowe. Użycie różnorodności tych miar może dać bardziej kompleksowy obraz stabilności.

4.4.6.4. Opis metody mierzenia stabilności xAI

Zaproponowane podejście do oceny wyjaśnień SHAP polega na analizie wyjaśnień SHAP wytworzonych z zakłóconych próbek, ponieważ tego rodzaju analiza może dostarczyć wglądu w to, jak odporny jest algorytm na niewielkie zmiany w danych wejściowych. Istnieje wiele sposobów wprowadzania zakłóceń do próbek. Do pomiaru zachowania wartości SHAP pod wpływem różnych rodzajów szumów wybrano trzy i każda z tych metod została zastosowana na czterech różnych poziomach intensywności. Te szumy to: szum gaussowski, tworzenie brakujących wartości oraz przetasowanie wartości w cesze.

Aby ocenić niezmienność lub wrażliwość wartości SHAP na szum, wybrany został zestaw miar statystycznych.

Wyniki eksperymentów ujawniają, że wszystkie rodzaje zakłóceń wpływają na stabilność wyjaśnień SHAP, choć w różnym stopniu. Przetasowanie cech i szum Gaussa zazwyczaj wywołują bardziej znaczące zakłócenia w wartościach SHAP, co prowadzi do zmniejszenia stabilności wyjaśnień. Z kolei brakujące wartości wykazują stosunkowo umiarkowany wpływ, co sugeruje, że wyjaśnienia SHAP są bardziej odporne na niekompletne dane. Wyniki te są kwantyfikowane za pomocą miar korelacji, pokazują tendencję spadku stabilności wraz ze wzrostem intensywności zakłóceń.

Tabela 12. Średnie metryki dla walidacji krzyżowej 10-krotnej - porównanie między oryginalnymi a zakłóconymi wartościami SHAP; zestaw danych NF-CICIDS2018.

Średnia z 10-fold CV - porównanie między oryginalnymi a zakłóconymi wartościami SHAP					
	Kendall's Tau	Spearman's Rank Correlation	Cosine Similarity	Pearson Correlation	RMSE
shuffled_05	0.769	0.844	0.407	0.407	0.025
shuffled_10	0.769	0.844	0.407	0.407	0.025
shuffled_20	0.769	0.844	0.407	0.407	0.025
shuffled_50	0.406	0.474	0.140	0.140	0.028
missing_05	0.932	0.957	0.949	0.949	0.007
missing_10	0.868	0.913	0.890	0.890	0.008
missing_20	0.763	0.829	0.777	0.777	0.013
missing_50	0.513	0.591	0.454	0.454	0.021
noise_05	0.417	0.492	0.475	0.475	0.005
noise_10	0.378	0.447	0.405	0.405	0.007
noise_20	0.339	0.402	0.332	0.332	0.013
noise_50	0.281	0.334	0.230	0.230	0.028

Na podstawie wyników przedstawionych w Tabela 12 można wnioskować, że w wypadku zbioru NF-CICIDS2018 metryki dla danych z przetasowaniem są stabilne przy mieszaniu do

20% danych, podczas gdy przy 50% następuje wyraźny spadek tych metryk. Może to sugerować, że wartości SHAP są stabilne przy przetasowaniu do 20%, ale również wskazywać, że model skupia się na jednej, decydującej cesze, która odpowiada za ponad 20% całkowitego wkładu w klasyfikacje, a różne poziomy szumu w tej kategorii wpływają jedynie na przetasowanie najważniejszej cechy.

Stabilność wartości SHAP pogarsza się stopniowo wraz ze wzrostem poziomu brakujących danych. Nawet przy brakach na poziomie 5% obserwuje się niewielki spadek metryk, który staje się bardziej wyraźny wraz ze wzrostem poziomu brakujących danych. Sugeruje to, że brakujące dane mają bardziej natychmiastowy wpływ na wartości SHAP w porównaniu do przetasowywania w przypadku zbioru danych NIDS, NF-CICIDS2018.

Wprowadzenie szumu Gaussa do danych ma poważny wpływ na wartości SHAP. Nawet przy szumie na poziomie 5% metryki są znacznie niższe niż w przypadku przetasowanych lub brakujących danych na tym samym poziomie. Podobnie jak w przypadku brakujących danych, stabilność wartości SHAP pogarsza się stopniowo wraz ze wzrostem poziomu szumu.

We wszystkich przypadkach stabilność maleje wraz ze wzrostem poziomu zakłóceń. Pokazuje to, że integralność danych wejściowych ma istotny wpływ na interpretację klasyfikacji modelu przy użyciu metody SHAP.

Badanie podkreśla, że stabilność wartości SHAP zmienia się w zależności od wprowadzonych zakłóceń (przetasowywanie, brakujące wartości i szum) na różnych poziomach intensywności. Ma to istotne implikacje dla wykorzystania wartości SHAP w rzeczywistych scenariuszach, podkreślając znaczenie jakości i kompletności danych przy interpretacji klasyfikacji modelu. Należy jednak zauważyć, że wpływ tych zakłóceń na stabilność SHAP może się różnić w zależności od specyficznych cech zbioru danych, takich jak liczba cech i ich względne znaczenie dla procesu klasyfikacyjnego wykonywanego przez model.

Badanie wskazuje na potrzebę testowania stabilności narzędzi xAI i wskazuje możliwość budowy jednolitej metryki stabilności xAI wykorzystującej perturbacje danych i miary korelacji.

4.4.7. Problem niebezpieczeństw związanych z udostępnianiem algorytmów xAI - możliwość ataków na klasyfikator

4.4.7.1. Zarys problemu

Nacisk na wyjaśnialność komponentów AI w kontekście wykrywania ataków sieciowych prowadzi to do paradoksalnej sytuacji, w której dążenie do transparentności może odbywać się kosztem samego bezpieczeństwa [64].

W szczególności mechanizmy umożliwiające interpretację decyzji AI, takie jak wyjaśnienia kontrfaktyczne (Counterfactual Examples - CE), mogą dostarczyć atakującym informacji na temat możliwości manipulowania modelami AI.

Konieczność wyboru między wyjaśnialnością a bezpieczeństwem to bardzo ważny dylemat, który jest jednak niedostatecznie omawiany w społeczności badawczej zajmującej się cyberbezpieczeństwem.

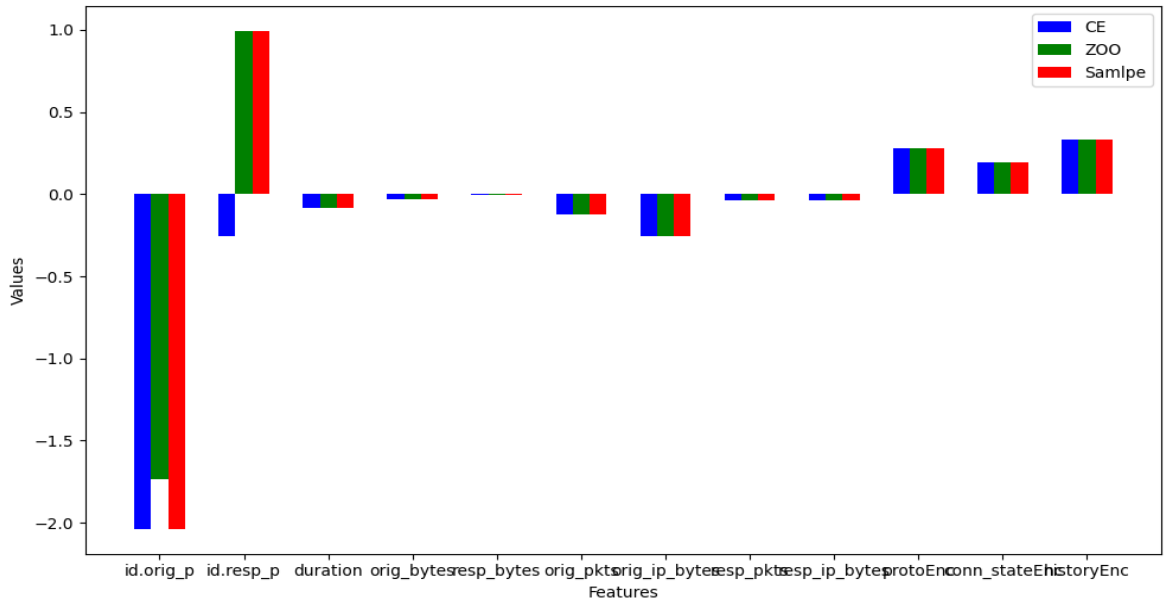
4.4.7.2. Propozycja wykorzystania algorytmu DiCE w celu tworzenia ataków na klasyfikator

Zaproponowałem nowy sposób formułowania ataków typu adversarial wykorzystując algorytm DiCE (Diverse Counterfactual Explanations), zaprojektowany w celu generowania wyjaśnień kontrfaktycznych decyzji AI, wykorzystując go do skutecznego zmieniania klasyfikacji modelu i w ten sposób ukrywania ataków przed detekcją [C07].

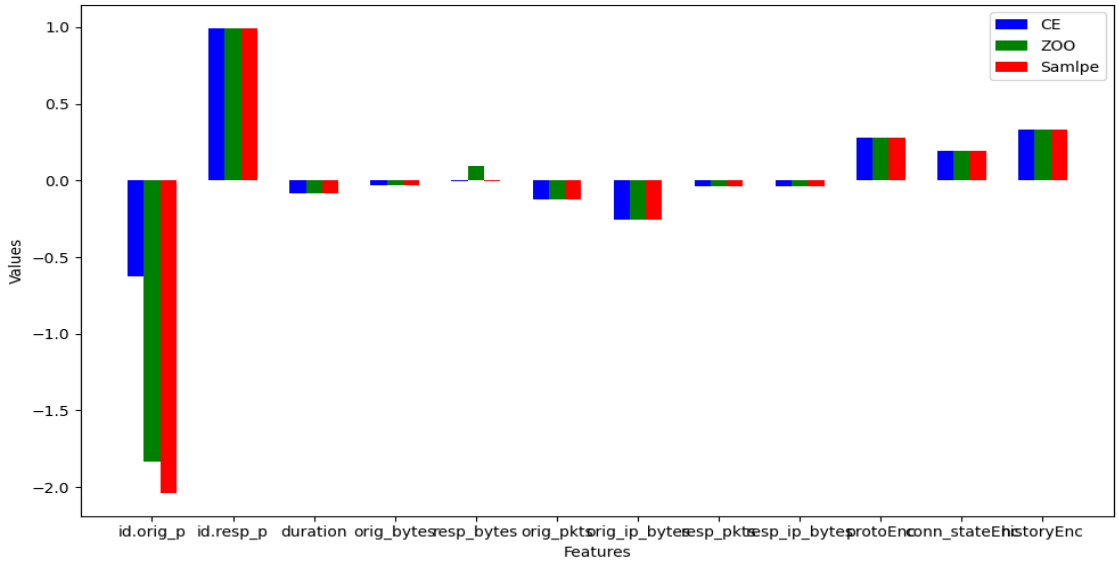
W tym celu proponuję nowy sposób wykorzystania procedury optymalizacyjnej DiCE do stworzenia próbek będących atakami na klasyfikator wykorzystany w wykrywaniu ataków sieciowych. Jest to osiągnięte poprzez tworzenie specyficznych próbek, które mogą odwracać etykiety detektora wykorzystującego ML.

Badanie to demonstruje istotne naruszenie integralności NIDS, kwestionując tradycyjne podejście do transparentności AI w kontekście cyberbezpieczeństwa. Zjawisko to zestawiono z jednym z szeroko znanych ataków typu adversarial, Zeroth Order Optimisation (ZOO), aby ukazać podobieństwa między tymi dwoma podejściami [C07].

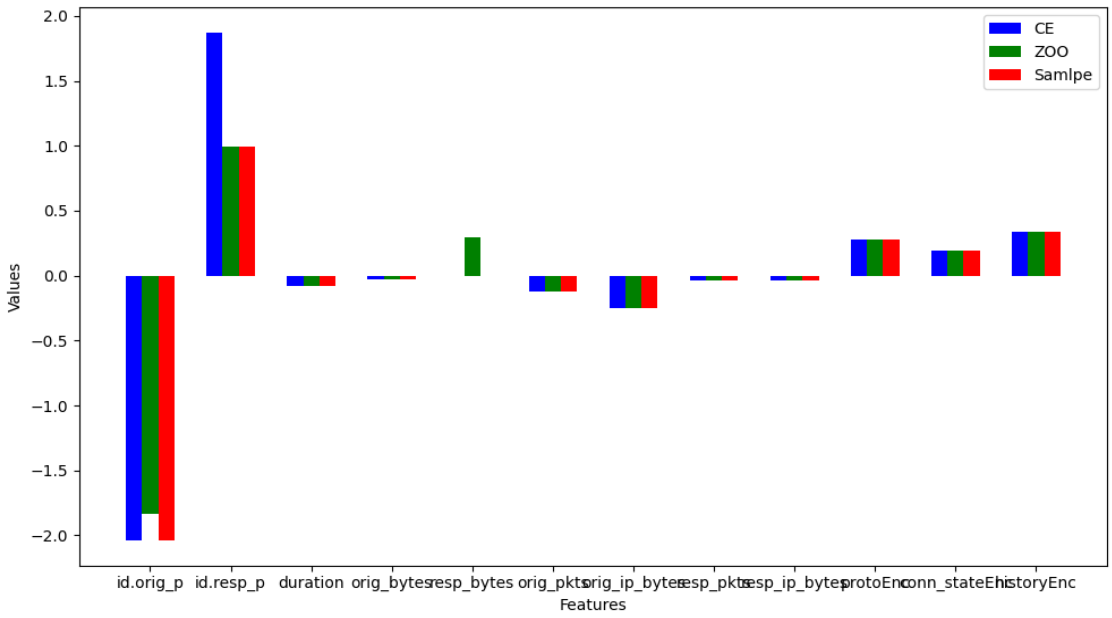
Eksperyment polega na zastosowaniu algorytmu DiCE na zbiorze danych IoT-23, ze szczególnym uwzględnieniem złośliwego oprogramowania Okiru, znanego ze swoich zdolności do unikania wykrycia. Poprzez generowanie próbek kontrfaktualnych przy pomocy DiCE, które mogą skutecznie zmienić klasyfikację ataku na „bezpieczny" (Benign), uwypukla się istotna luka w systemach bezpieczeństwa wykorzystujących AI i xAI: mimo że mają one na celu zwiększenie przejrzystości i zaufania, CE mogą umożliwić atakującym manipulowanie decyzjami AI w celu unikania wykrycia. Wyniki badania pokazują, że chociaż ataki typu adversarial wykonane przy pomocy algorytmu DiCE były mniej skuteczne niż ZOO w zmienianiu wyników klasyfikacji - osiągając 42% skuteczności w porównaniu do prawie 100% w przypadku ZOO - stanowią jednak znaczące zagrożenie dla bezpieczeństwa.



Rysunek 5. Względne zmiany cech próbki – porównanie pomiędzy atakiem ZOO (zielony), atakiem wykonanym przy pomocy algorytmu DiCE (CE – niebieski) i pierwotna próbka (czerwony). Oś x przedstawia cechy, oś Y przedstawia standaryzowane wartości cech.



Rysunek 6. Względne zmiany cech próbki – porównanie pomiędzy atakiem ZOO (zielony), atakiem wykonanym przy pomocy algorytmu DiCE (CE – niebieski) i pierwotna próbka (czerwony). Oś x przedstawia cechy, oś Y przedstawia standaryzowane wartości cech.



Rysunek 7. Względne zmiany cech próbki – porównanie pomiędzy atakiem ZOO (zielony), atakiem wykonanym przy pomocy algorytmu DiCE (CE – niebieski) i pierwotna próbka (czerwony). Oś x przedstawia cechy, oś Y przedstawia standaryzowane wartości cech.

Tabela 13. Porównanie współczynników sukcesu różnych typów ataków

Typ Ataku	**Success Rate**
Counterfactuals	42%
Adversarials	99%

Wyniki badania skłaniają do krytycznego spojrzenia na strategie wdrażania xAI w cyberbezpieczeństwie. Podkreślają potrzebę zrównoważonego podejścia, które uwzględnia nie tylko korzyści płynące z uczynienia systemów AI bardziej zrozumiałymi i odpowiedzialnymi, ale także potencjalne ryzyko, jakie te metody mogą stanowić dla bezpieczeństwa systemów. Szersza społeczność zajmująca się cyberbezpieczeństwem musi rozważyć korzyści związane z wyjaśnialnością w kontekście potencjalnych zagrożeń, jakie te narzędzia mogą stwarzać jako wektory ataków.

Wnioski płynące z badania sugerują potrzebę rozwoju metod wykrywania i przeciwdziałania wykorzystaniu wyników xAI do celów adwersarialnych, zwłaszcza w aplikacjach krytycznych dla bezpieczeństwa.

4.4.8. Problem oceny algorytmów xAI - nowe podejście, PMI

4.4.8.1. Zarys problemu

Wyjaśnialność jest ważna z kilku powodów [65][66]:

- Pomaga identyfikować i eliminować modele typu „Clever Hans", w których, gdy widoczne są jedynie wyniki, a nie proces, okazuje się, że proces jest jedynie obejściem, odbywającym się z pominięciem intencji twórcy modelu AI.
- xAI zapewnia niezawodność poprzez umożliwienie regularnej weryfikacji modeli i danych, co pozwala rozwiązywać problemy takie jak dryf koncepcji (Concept Drift) czy starzenie się danych (Data Decay).
- Buduje zaufanie do systemów AI, szczególnie w decyzjach wysokiej stawki, takich jak diagnozy medyczne czy cyberbezpieczeństwo, gdzie zrozumienie procesu decyzyjnego jest niezbędne.
- xAI redukuje uprzedzenia i promuje sprawiedliwość poprzez ujawnianie niechcianych uprzedzeń w zbiorach danych, co pozwala dostosować modele AI do współczesnych standardów etycznych.
- Dostarcza głębszych wglądów w analizowaną dziedzinę, odkrywając nieznane wcześniej zależności w danych i umożliwiając nowe odkrycia, co czyni ją nieocenioną dla społeczności naukowej.

Nie sposób nie zauważyć, że wszystkie te kwestie są równie ważne jeszcze przed rozpoczęciem treningu jakiegokolwiek modelu.

Większość czynników, które przyczyniają się do znaczenia xAI, zachowuje swoją istotność już na początkowych etapach rozwoju modelu, zanim rozpocznie się jakiekolwiek uczenie.

- Ważne jest zidentyfikowanie cech w danych, które mogą prowadzić model do podejmowania decyzji na podstawie fałszywych korelacji, zamiast rzeczywistych zależności. Wymaga to analizy istotności i wkładu każdej cechy w potencjalne wyniki.
- Przed rozpoczęciem szkolenia konieczna jest weryfikacja integralności danych i założeń leżących u podstaw podejścia modelowania. Obejmuje to sprawdzenie jakości danych, ich spójności oraz zgodności rozkładów z oczekiwanymi wzorcami.

- Przed wdrożeniem modeli ML należy skorygować wszelkie niechciane uprzedzenia w zbiorze danych.
- Analiza relacji między cechami a ich wpływem na zmienne docelowe może ujawnić złożone interakcje i zależności, które są nieocenione dla zrozumienia domeny problemu.
- Wreszcie, ustanowienie mechanizmów transparentności zaczyna się od jasnego zrozumienia charakterystyki danych i ich oczekiwanego wpływu na zachowanie modelu.

Zjawisko możliwości odpowiedzi na część tych pytań przed uczeniem modelu określono mianem interpretowalności przedmodelowej (PMI, pre-model interpretability) – metod pozwalających na zrozumienie charakterystyki danych i potencjalnego zachowania modeli ML bez ponoszenia kosztów obliczeniowych związanych z ich uczeniem.

Tradycyjne podejścia post-hoc, takie jak SHAP czy wskaźniki ważności cech (feature importance scores), wymagają wyuczonego modelu. Analizują one model dopasowany do danych. Jednak bez zrozumienia, jakie informacje były dostępne w danych, nie ma możliwości porównania tych miar z rzeczywistym punktem odniesienia (ground truth).

4.4.8.2. Proponowane metody: trzy nowe metryki PMI

W celu zaadresowania powyższej problematyki:

Zaproponowałem nowe podejście do zagadnienia xAI - Interpretowalność Przed Modelem (PMI: Pre-Model Interpretability), oraz trzy nowe metryki PMI - ARIA HaRIA GeRIA [C08].

- Zaproponowałem trzy metryki PMI, aby skutecznie oszacować wpływ cech na model przed jego wytrenowaniem. Ich użyteczność w kontekście PMI jest oceniana i porównywana z wynikami znaczenia cech Random Forest (Random Forest Importance Score) oraz wartościami SHAP.
- Zaproponowane metryki przetestowałem i zwalidowałem na różnorodnych zbiorach danych, demonstrując jak osiągają wyniki porównywalne do tradycyjnych metryk potrzebujących wytrenowanego modelu ML, ale bez kosztów związanych z uczeniem.
- Przeprowadziłem analizę statystyczną, pokazując, jak proponowane metryki przedmodelowe korelują z tradycyjnymi wynikami znaczenia cech i SHAP, na wielu zbiorach danych.

- Nakreśliłem potencjalne przyszłe badania, które miałyby na celu zawarcie w metrykach synergicznych relacji między cechami. Mogłoby to dotyczyć scenariuszy, w których wspólna wiedza o dwóch lub więcej cechach znacząco zwiększa przewidywalność zmiennej celu, ponad to, co mogłoby być wywnioskowane z każdej cechy indywidualnie .

Zaproponowane metryki oferują ocenę potencjalnego wykorzystania cech przed rozpoczęciem procesu uczenia. Robiąc to, dostarczają one podstawowej informacji o tym, które cechy są ważne, na podstawie ich zawartości informacyjnej i wariancji, które są niezależnie od jakichkolwiek charakterystyk modelu. Metryki PMI mogą służyć jako punkt odniesienia dla dalszych ewaluacji metod xAI. Dostarcza to informacji czy wyjaśnienia generowane przez metody xAI faktycznie odzwierciedlają dynamikę zawartą w danych, czy są artefaktami procesu modelowania.

Wgląd w znaczenie cech przed procesem uczenia może pomóc w ustaleniu, czy modelowanie przebiega zgodnie z intuicją branżową i pomaga zauważyć rozbieżności w samych danych.

Zaproponowane metryki - HaRIA, ARIA i GeRIA - oceniają względną dostępność informacji (Relative Information Availability, RIA) cech danych, wykorzystując informację wzajemną (Mutual Information) oraz wartości F z analizy ANOVA, skalowane za pomocą skalowania bezwzględnych wartości maksymalnych (maximum absolute scaling). Podejście to umożliwia efektywną ocenę potencjału cech bez kosztów obliczeniowych związanych z treningiem modelu, oferując nowy sposób oceny cech w ML [C08].

Zaproponowane metryki „Względnej Dostępności Informacji" to zestaw trzech złożonych metryk, które łączą różne aspekty cech oferując zrozumienie ich możliwych wkładów w proces uczenia się, przed właściwym trenowaniem modelu. Metryki te wykorzystują redukcję niepewności, jaką cecha dostarcza odnośnie wartości celu, oraz kwantyfikację wpływu cechy na cel kategoryczny, mierzoną poprzez ocenę zmienności wartości cech w kategoriach określonych przez cel. Wartości te są uzyskiwane względem zbioru danych - elementy metryk są skalowane przy użyciu maksymalnej wartości bezwzględnej uzyskanej na cechę. Części proponowanej metryki są łączone ze sobą na jeden z trzech sposobów: przy użyciu średnich arytmetycznej, harmonicznej lub geometrycznej.

Proponowane metryki łączą zalety metod filtrujących (szybkość i prostota) z potencjałem do uchwycenia interakcji poprzez używane metody agregacji (średnie arytmetyczna, harmoniczna, geometryczna). Dostarczają bardziej zaokrągloną ocenę znaczenia cech, która uwzględnia różne właściwości.

Metryki łączą względną zawartość informacji cechy z jej względną zmiennością, dostarczając zrównoważonej perspektywy na to, co czyni cechę prawdopodobnie istotną w treningu modelu. Poprzez integrację kwantyfikacji różnych aspektów cech, metryki zyskują na wyrazistości uchwyconych złożoności.

Podczas gdy MI może uchwycić relacje nieliniowe, wartość F może wzmacniać znaczenie cech z istotnymi efektami zmienności między klasami.

Metryki dostarczają istotnych wglądów przedmodelowych, pozwalając użytkownikowi sprawdzić, czy intuicja informacji oczekiwanych od cech zgadza się z tym, co faktycznie jest obecne w danych i dostępne dla modeli AI/ML, kształtując oczekiwania użytkownika końcowego co do zachowania modelu.

Ponieważ te metryki nie wykorzystują konkretnego modelu ML, mogą dostarczyć bardziej obiektywny punkt odniesienia, który nie jest obciążony specyfiką konkretnego algorytmu modelowania.

$$\mathrm{ARIA}(X_f, Y) = \frac{\dfrac{\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\dfrac{p(x, y)}{p(x)p(y)}\right)}{\max\limits_{f \in F}\left(\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\dfrac{p(x, y)}{p(x)p(y)}\right)\right)} + \dfrac{\dfrac{\sum_{i=1}^{k} n_i \left(\overline{X_{i,f}} - \overline{X_f}\right)^2}{k-1}}{\max\limits_{f \in F}\left(\dfrac{\sum_{i=1}^{k} n_i \left(\overline{X_{i,f}} - \overline{X_f}\right)^2}{k-1}\right)}}{2}$$

Średnia arytmetyczna używana w formule ARIA (Arithmetic Relative Information Availability) oferuje prostą średnią, dającą równowagę między dwoma elementami równania, zakładając równą wagę obu elementów.

$$\mathrm{HaRIA}(X_f, Y) = \frac{2\left(\dfrac{\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\dfrac{p(x, y)}{p(x)p(y)}\right)}{\max\limits_{f \in F}\left(\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\dfrac{p(x, y)}{p(x)p(y)}\right)\right)}\right)\left(\dfrac{\dfrac{\sum_{i=1}^{k} n_i \left(\overline{X_{i,f}} - \overline{X_f}\right)^2}{k-1}}{\max\limits_{f \in F}\left(\dfrac{\sum_{i=1}^{k} n_i \left(\overline{X_{i,f}} - \overline{X_f}\right)^2}{k-1}\right)}\right)}{\left(\dfrac{\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\dfrac{p(x, y)}{p(x)p(y)}\right)}{\max\limits_{f \in F}\left(\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\dfrac{p(x, y)}{p(x)p(y)}\right)\right)}\right) + \left(\dfrac{\dfrac{\sum_{i=1}^{k} n_i \left(\overline{X_{i,f}} - \overline{X_f}\right)^2}{k-1}}{\max\limits_{f \in F}\left(\dfrac{\sum_{i=1}^{k} n_i \left(\overline{X_{i,f}} - \overline{X_f}\right)^2}{k-1}\right)}\right)}$$

Średnia harmoniczna obecna we wzorze HaRIA w równaniu faworyzuje mniejszy z dwóch elementów równania, co może być szczególnie przydatne w scenariuszach, gdzie oba wyniki muszą być wystarczająco wysokie, aby uznać cechę za naprawdę istotną. Pozwala to wskazać cechy, które są informacyjne i prawdopodobnie będą wykorzystywane przez model.

$$\mathrm{GeRIA}(X_f, Y) = \sqrt{\frac{\sum_{y\in Y}\sum_{x\in X_f} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)}{\max\limits_{f\in F}\left(\sum_{y\in Y}\sum_{x\in X_f} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right)} \cdot \frac{\frac{\sum_{i=1}^{k} n_i\left(\overline{X_{i,f}} - \overline{X_f}\right)^2}{k-1}}{\max\limits_{f\in F}\left(\frac{\sum_{i=1}^{k} n_i\left(\overline{X_{i,f}} - \overline{X_f}\right)^2}{k-1}\right)}}$$

Średnia geometryczna we wzorze GeRIA w równaniu zapewnia równowagę, która jest mniej wrażliwa na wartości ekstremalne niż ARIA, co może być przydatne podczas łączenia elementów, które mogą mieć różne rozkłady. Naturalnie, czyni to ją podobną do ARIA w tym kontekście, ponieważ dzięki zastosowaniu skalowania MaxAbs możliwość napotkania wartości ekstremalnych jest minimalna.

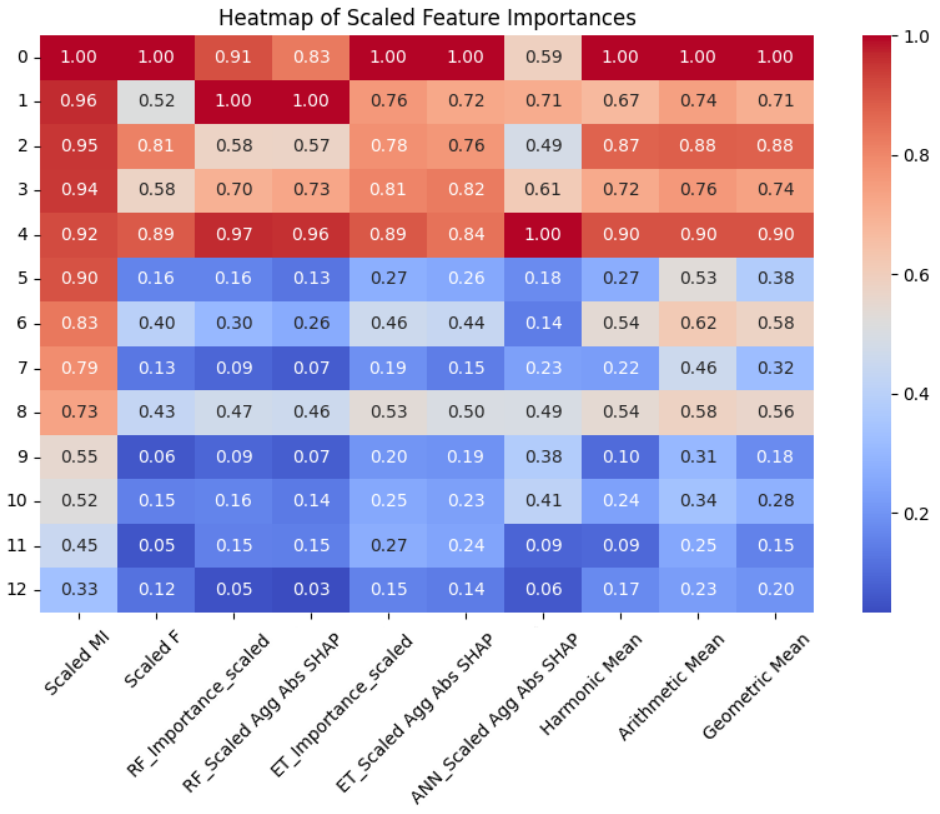
W ramach serii eksperymentów zweryfikowałem zaproponowane metryki interpretowalności przed modelem na różnych zestawach danych, wykazując ich silną korelację z tradycyjnymi ocenami po treningu modelu, takimi jak oceny ważności cech i wartości SHAP, co jest ukazane w Tabela 14. Ta korelacja podkreśla zdolność metryk do przybliżania złożonych ocen modelu, które tradycyjnie są dostępne dopiero po intensywnych obliczeniach. Wyniki te wskazują na potencjał zaproponowanych metryk do dostarczania wczesnych informacji na temat ważności cech, zbieżnych z wynikami bardziej kosztownych procesów wyjaśniania modelu [C08].

Tabela 14. Korelacja Pearsona współczynnika ważności cech mierzonego różnymi metodami z zaproponowanymi metrykami, w różnych zbiorach danych.

Feature	**Correlation with HaRIA**	**Correlation with ARIA**	**Correlation with GeRIA**	**Dataset**
RF_Importance_scaled	0.87444	0.90712	0.87707	ToN
RF_Scaled Agg Abs SHAP	0.92545	0.93913	0.93472	ToN
ET_Importance_scaled	0.86113	0.90014	0.85972	ToN
ET_Scaled Agg Abs SHAP	0.96045	0.96453	0.97619	ToN
ANN_Scaled Agg Abs SHAP	0.66161	0.57921	0.69730	ToN
RF_Importance_scaled	0.90385	0.88366	0.90164	Wine
RF_Scaled Agg Abs SHAP	0.89099	0.86769	0.88768	Wine
ET_Importance_scaled	0.97213	0.94811	0.96811	Wine

ET_Scaled Agg Abs SHAP	0.96841	0.94348	0.96401	Wine
ANN_Scaled Agg Abs SHAP	0.75080	0.73585	0.74796	Wine
RF_Importance_scaled	0.99999	0.99687	0.99942	Iris
RF_Scaled Agg Abs SHAP	0.99948	0.99424	0.99803	Iris
ET_Importance_scaled	0.99630	0.99185	0.99530	Iris
ET_Scaled Agg Abs SHAP	0.99627	0.98975	0.99440	Iris
ANN_Scaled Agg Abs SHAP	0.97791	0.96465	0.97340	Iris
RF_Importance_scaled	0.85429	0.81331	0.84491	Diabetes
RF_Scaled Agg Abs SHAP	0.97634	0.96515	0.97482	Diabetes
ET_Importance_scaled	0.85597	0.80228	0.84351	Diabetes
ET_Scaled Agg Abs SHAP	0.08882	0.10342	0.08298	Diabetes
ANN_Scaled Agg Abs SHAP	0.19900	0.14118	0.16582	Diabetes
RF_Importance_scaled	0.91715	0.97498	0.91272	Rice
RF_Scaled Agg Abs SHAP	0.90529	0.96790	0.90759	Rice
ET_Importance_scaled	0.98419	0.98702	0.98160	Rice
ET_Scaled Agg Abs SHAP	0.94445	0.98375	0.94470	Rice
ANN_Scaled Agg Abs SHAP	0.63238	0.60907	0.63846	Rice
RF_Importance_scaled	0.69840	0.79211	0.72810	Glass
RF_Scaled Agg Abs SHAP	0.79213	0.81322	0.81130	Glass
ET_Importance_scaled	0.70721	0.81783	0.70117	Glass
ET_Scaled Agg Abs SHAP	0.72324	0.84801	0.75600	Glass
ANN_Scaled Agg Abs SHAP	0.51669	0.80122	0.62040	Glass
RF_Importance_scaled	0.55572	0.87330	0.62040	Drybean
RF_Scaled Agg Abs SHAP	0.72617	0.85478	0.75994	Drybean
ET_Importance_scaled	0.53811	0.82810	0.60219	Drybean
ET_Scaled Agg Abs SHAP	0.71731	0.80643	0.75229	Drybean
ANN_Scaled Agg Abs SHAP	0.51669	0.80122	0.56820	Drybean

W zbiorze danych Wine, jak widać na Rysunek 8, występują znaczące korelacje między metrykami wykorzystującymi model (Importance Score z RF i ET, wartości SHAP) a proponowanymi metrykami RIA. Rysunek ukazuje, że niższa wariancja wyrażona przez wartość F ogranicza użyteczność cech, które inaczej mogłyby być bardzo silne, jak wskazuje MI (np. w cesze 5). Jednak cecha 1 i cecha 2 wykazują przeciwne zachowanie, gdzie stosunkowo niska wartość F jest równoważona przez wysoką zawartość informacji (MI), gdzie większość tradycyjnych metryk i proponowanych metryk wyraźnie wskazują na znaczenie tych cech. Cechy 0 i 4 utrzymują wysoką ważność niemal we wszystkich metrykach, sugerując, że są kluczowe dla modeli w tym zbiorze danych. Jest to uchwycone i wyrażone przez proponowane metryki.



Rysunek 8. Mapa ciepła zbioru danych Wine.

4.4.9. Kierunki dalszych badań

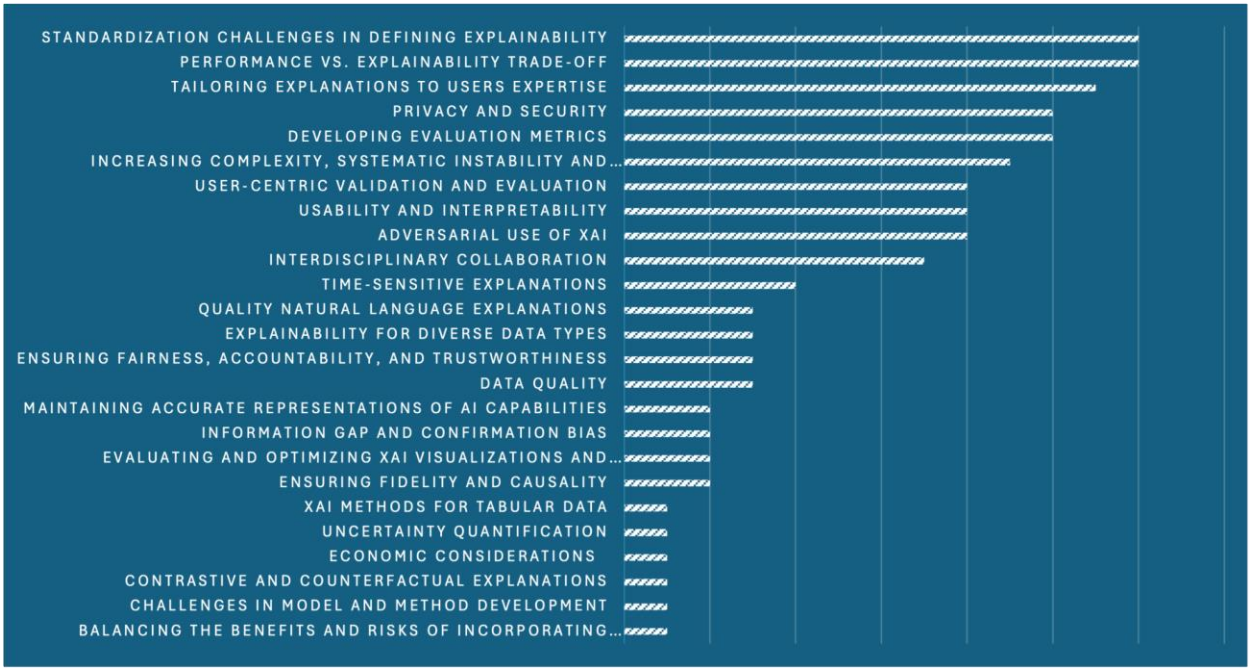
W temacie metryk RIA sugeruję również obszary do dalszych badań, zwłaszcza w zrozumieniu, jak różne modele przetwarzają i priorytetyzują informacje. Jednym z bezpośrednich kierunków rozwoju jest opracowanie metryk pozwalających wychwycić synergiczne relacje między cechami - na co nie pozwalają metryki RIA w obecnej formule.

4.4.9.1. Wyzwania xAI

Ostatnią publikację w cyklu poświęcam wyznaczeniu kierunków dalszych badań.

W formie przeglądu literatury zidentyfikowałem 25 konkretnych wyzwań w adopcji i rozwoju xAI, tak w NIDS jak i poza NIDS. Od obaw dotyczących prywatności danych po techniczne ograniczenia w tworzeniu modeli wyjaśnialnych [C09].

Każde wyzwanie jest szansą na dalsze badania i ulepszenia w tej dziedzinie. Zidentyfikowane wyzwania przytaczam na poniższej liście, oraz na Rysunek 9.



Rysunek 9. Proporcjonalna reprezentacja 25 wyzwań na podstawie ich częstości występowania w analizowanej literaturze.

1. Wyzwania standaryzacyjne w definiowaniu wyjaśnialności
2. Kompromis między wydajnością a wyjaśnialnością
3. Dostosowywanie wyjaśnień do wiedzy użytkowników
4. Prywatność i bezpieczeństwo
5. Opracowywanie metryk oceny
6. Zwiększona złożoność, systemowa niestabilność i wielość wyjaśnień
7. Wykorzystanie xAI w calu konstrukcji ataków typu adversarial
8. Walidacja i ocena zorientowane na użytkownika
9. Użyteczność i interpretowalność
10. Współpraca interdyscyplinarna
11. Wyjaśnienia wrażliwe na czas
12. Zapewnienie uczciwości, odpowiedzialności i wiarygodności
13. Jakość naturalnych wyjaśnień językowych

14. Jakość danych
15. Wyjaśnialność dla różnych typów danych
16. Luka informacyjna i uprzedzenia potwierdzające
17. Ocena i optymalizacja wizualizacji xAI oraz metod interpretowalności
18. Zapewnienie wierności i przyczynowości
19. Utrzymanie dokładnych reprezentacji możliwości AI
20. Zrównoważenie korzyści i ryzyk związanych z włączeniem wyjaśnialności do systemów AI
21. Metody xAI dla danych tabelarycznych
22. Wyzwania w rozwoju modeli i metod
23. Kontrastowe i kontrfaktyczne wyjaśnienia
24. Kwantyfikacja niepewności
25. Rozważania ekonomiczne

Wyzwania te mają dwoisty charakter. Każde z zidentyfikowanych wyzwań wymienionych może być również postrzegane jako potencjalne możliwości. Ta perspektywa obejmuje dynamiczną i ewoluującą naturę krajobrazu cyberbezpieczeństwa, gdzie przeszkody napotkane przez organizacje i badaczy często stanowią możliwości dla innowacji, poprawy i zdobycia strategicznej przewagi.

Staranne zbadanie zidentyfikowanych różnorodnych wyzwań xAI pozwala podkreślić pilną potrzebę kontynuacji badań i rozwoju w dziedzinie wyjaśnialnego wykrywania ataków sieciowych, ponieważ osiągnięcie przejrzystości i interpretowalności tych systemów jest ważne w celu zapewnienia bezpieczeństwa infrastruktury cyfrowej

Dwoistość wyzwań jako możliwości podkreśla dynamiczną naturę dziedziny xAI. Każda przeszkoda może być przekształcona w kamień milowy na drodze do bardziej solidnego, elastycznego i efektywnego ekosystemu wykrywania ataków sieciowych.

4.4.10. Podsumowanie osiągnięcia naukowego

Podsumowując, najważniejsze aspekty przedstawionego cyklu i ich naukowe znaczenie są następujące:

Nazwa osiągnięcia	Opis	Znaczenie naukowe
Optymalizacja metod AI i ML w NIDS	Systematycznie przeanalizowałem i zoptymalizowałem topologie i hiperparametry ANN dla NIDS, adresując problem arbitralnych konfiguracji w istniejących badaniach.	Praca ta dostarcza metodologicznych podstaw do rozwijania dokładniejszych i bardziej niezawodnych NIDS. Podkreślając znaczący wpływ hiperparametrów na dokładność wykrywania, kieruje przyszłe badania i praktyczne implementacje w kierunku najlepszych praktyk w konfiguracji modelu.
Rozwiązywanie problemu przesunięcia wdrożeniowego za pomocą uczenia transferowego	Zaproponowałem metody wykorzystujące uczenie transferowe do adaptacji modeli ML do nowych środowisk sieciowych przy minimalnych danych z sieci docelowej.	Zmaganie się z problemem „przesunięcia wdrożeniowego" jest ważne dla rzeczywistej użyteczności NIDS. Podejście to redukuje potrzebę zbierania obszernych danych i ponownego szkolenia przy wdrażaniu NIDS w różnych ustawieniach sieciowych, czyniąc zaawansowane środki cyberbezpieczeństwa bardziej praktycznymi i opłacalnymi.
Zwiększenie zdolności wykrywania nieznanych ataków za pomocą sieci neuronowych Siamese	Opracowałem nową metodę uczenia dla sieci neuronowych Siamese z wykorzystaniem ekstrakcji granic za pomocą algorytmu k-NN w celu wykrywania wcześniej nieobserwowanych ataków sieciowych (zero-day).	Metoda ta poprawia zdolności NIDS, koncentrując się na informatywnych próbkach blisko granic klas. Wykrywanie nieznanych ataków jest ważne dla proaktywnego cyberbezpieczeństwa; sposoby poprawiające wykrywalność nieznanych ataków stanowią znaczący postęp w technikach ML stosowanych w NIDS.
Ocena technik xAI poprzez analizę stabilności	Zaproponowałem nową metodę oceny technik xAI poprzez ocenę stabilności wyjaśnień w obliczu zakłóceń danych, wykorzystując metryki wykorzystujące miary korelacji	Zapewnienie wiarygodności wyjaśnień AI jest ważne dla zwiększenia zaufania do AI. Praca ta pomaga w ustaleniu metryk do oceny wiarygodności metod xAI, co jest ważne dla ich adopcji w aplikacjach cyberbezpieczeństwa
Ujawnianie ryzyka wykorzystania xAI w atakach typu adversarial	Wykazałem, że algorytmy xAI, takie jak DiCE do generowania wyjaśnień kontrfaktualnych, można wykorzystać do tworzenia ataków typu adversarial, które omijają NIDS.	Podkreślenie tej podatności zwraca uwagę na potrzebę równoważenia xAI z bezpieczeństwem w systemach AI. Powinno to mieć wpływ na dalszy rozwój metod xAI, które muszą zarówno dostarczać wyjaśnialności jak i brać pod uwagę odporność na możliwe wykorzystanie ich w złej intencji. Świadomość ta przyczynia się do bezpieczniejszych wdrożeń AI.
Wprowadzenie interpretowalności przedmodelowej (PMI) i nowych metryk	Wprowadziłem koncepcję Interpretowalności Przedmodelowej (PMI) i opracowałem nowe metryki (ARIA, HaRIA,	PMI pozwala na zrozumienie charakterystyk danych i potencjalnego zachowania modelu bez kosztów uczenia. Te metryki pomagają wykrywać potencjalne problemy wcześnie, poprawiają niezawodność modelu

	GeRIA) do oceny znaczenia cech przed uczeniem modelu.	i stanowią bazę do oceny metod xAI.
Identyfikacja przyszłych kierunków badań w xAI	Na podstawie kompleksowego przeglądu literatury zidentyfikowano 25 wyzwań i możliwości w xAI, w tym standaryzacji, obaw dotyczących prywatności i kompromisu między wydajnością a wyjaśnialnością.	Praca ta służy jako pierwszy krok dla przyszłych badań, zachęcając do interdyscyplinarnej współpracy i kierując wysiłki na rozwiązanie ważnych problemów w xAI.

Cyberbezpieczeństwo to dziedzina, w której niezawodność, solidność i zaufanie są bardzo ważne. Poprzez optymalizację wydajności NIDS, adaptację modeli do nowych środowisk, zwiększenie wykrywania nieznanych ataków oraz rozwijanie zrozumienia i zastosowania xAI, przedstawione badania rozwijają zarówno teoretyczną wiedzę, jak i praktyczne narzędzia do zabezpieczania sieci komputerowych. Te postępy są ważne w budowaniu bezpiecznych, przejrzystych i godnych zaufania systemów AI zdolnych do obrony przed ewoluującymi zagrożeniami cybernetycznymi, mając tym samym wpływ na szersze przyjęcie AI w zastosowaniach cyberbezpieczeństwa.

4.5. Omówienie pozostałych osiągnięć naukowo-badawczych

4.5.1. Kolejne Osiągnięcie (poza głównym cyklem): Propozycje mechanizmów defensywnych przed atakami typu adversarial

Moje drugie osiągnięcie dotyczy opracowania zaawansowanych mechanizmów obronnych przeciwko atakom typu adversarial na systemy sztucznej inteligencji. Ataki typu adversarial to wynik procedury optymalizacyjnej, która maksymalizuje wpływ na model AI przy minimalizacji zmian w cechach próbki danych. Efektem tej procedury jest bardzo specyficzny rodzaj drobnego szumu, który dodaje się do oryginalnej próbki. W efekcie powstaje próbka, która dla człowieka wygląda identycznie jak próbka oryginalna, ale model ML klasyfikuje ją w inny sposób. W kontekście NIDS, można wykorzystać ten mechanizm w celu ukrycia złośliwego oprogramowania przed detektorami wykorzystującymi ML/AI. W obliczu rosnących zagrożeń związanych z manipulacją danymi wejściowymi oraz próbami obejścia zabezpieczeń, ważne jest stworzenie rozwiązań, które zwiększą odporność systemów na tego typu ataki. Prace badawcze **[DA01]** oraz **[DA02]** stanowią rozpatrzenie różnych podejść do zabezpieczenia systemów sztucznej inteligencji; w każdej z nich proponuję innowacyjne metody defensywne.

W badaniu **[DA01]** skupiam się na obronie systemów detekcji włamań sieciowych przed atakami typu adversarial z rodziny ataków ewazyjnych. Propozycja to budowa detektora obserwującego aktywacje neuronów w ANN działającej w roli detektora w NIDS (ANN1).

Kolejna sieć (ANN2), która obserwuje aktywacje neuronów sieci ANN1 jest wytrenowana na zestawach aktywacji pojawiających się gdy ANN1 jest atakowana i gdy ANN2 nie jest atakowana. Dzięki temu ANN2 jest w stanie rozpoznawać próby ataków typu adversarial. Integracja tej metody pozwala na stworzenie bardziej odpornego i niezawodnego systemu detekcji.

Artykuł **[DA02]** rozszerza tematykę obrony przed atakami typu adversarial na obszar przetwarzania obrazów, szczególnie w kontekście reidentyfikacji. Wprowadzam w nich zaawansowane strumienie przetwarzania wstępnego, w tym konwolucyjną sieć neuronową do dopasowywania bloków (block-matching CNN) stosowaną do odszumiania (denoising) obrazów. Takie podejście oczyszcza dane wejściowe z części szumów typu adversarial, czyniąc systemy wykorzystujące AI bardziej odporne na manipulacje takie jak ataki typu adversarial.

[DA01]	https://doi.org/10.1016/j.future.2020.04.013	Defending network intrusion detection systems against adversarial evasion attacks WoS: 82; Sco: 111; GSc: 161;	Future Generation Computer Systems	Marek Pawlicki, Michał Choraś, Rafał Kozik
[DA02]	https://doi.org/10.3390/e23101304	Preprocessing Pipelines Including Block-Matching Convolutional Neural Network for Image Denoising to Robustify Deep Reidentification against Evasion Attacks WoS: 3; Sco: 4; GSc: 6;	ENTROPY	Marek Pawlicki, Ryszard S. Choraś

4.5.2. Kolejne Osiągnięcie: Pozyskiwanie zbiorów danych i formułowanie cech dla ML w zadaniu wykrywania ataków sieciowych

Moja rola jako eksperta w dziedzinie uczenia maszynowego i cyberbezpieczeństwa w różnych projektach finansowanych przez Komisję Europejską pozwoliła mi zdobyć unikalne doświadczenie. W projektach unijnych miałem okazję współpracować przy tworzeniu i udoskonalaniu realnych zbiorów danych służących do wykrywania ataków sieciowych. Przykładem może być zbiór danych pozyskanych wraz z partnerami ze Słowenii [DS01]. Ta współpraca pozwoliła mi na głębsze zrozumienie wyzwań związanych z detekcją ataków w czasie rzeczywistym.

Podobnie w [DS02] mogłem bezpośrednio przyczynić się do propozycji, formułowania i ewaluacji nowego zbioru danych, co było istotne dla rozwoju efektywnych systemów detekcji ataków sieciowych w sieci RoEduNet. Moje doświadczenie w analizie

i przetwarzaniu danych sieciowych [DS04] okazało się bardzo ważne dla sukcesu tego projektu. Pozyskane zbiory danych [DS01][DS02][DS03] upubliczniono do użytku społeczeństwu naukowemu pracującemu nad wykorzystaniem ML w NIDS.

[DS01]	https://dl.acm.org/doi/abs/10.1145/3538969.3544486	M. Komisarek, **M. Pawlicki**, M-E Mihailescu, D. Mihai, M. Carabas, R. Kozik, and M. Choraś, (2022). A novel, refined dataset for real-time Network Intrusion Detection. In Proceedings of the 17th International Conference on Availability, Reliability and Security (ARES '22). Association for Computing Machinery, New York, NY, USA, Article 43, 1–8. https://doi.org/10.1145/3538969.3544486	CORE B
[DS02]	http://dx.doi.org/10.3390/s21134319	Mihailescu, M. -E., Mihai, D., Carabas, M., Komisarek, **M., Pawlicki**, M., Hołubowicz, W., & Kozik, R. (2021). The Proposition and Evaluation of the RoEduNet-SIMARGL2021 Network Intrusion Detection Dataset. Sensors, 21(13), 4319. https://doi.org/10.3390/s21134319	IF 3.847
[DS03]	https://dl.acm.org/doi/10.1145/3600160.3605094	M. Komisarek, **M. Pawlicki**, T. Simic, D. Kavcnik, R. Kozik, and M. Choraś, (2023), Modern NetFlow network dataset with labeled attacks and detection methods, in Proceedings of the 18th International Conference on Availability, Reliability and Security, pp. 1–8.	CORE B
[DS04]	https://doi.org/10.3390/e23111532	M. Komisarek, **M. Pawlicki**, R. Kozik, W. Hołubowicz and M. Choraś (2021). How to Effectively Collect and Process Network Data for Intrusion Detection? in Entropy 23 (11): 1532. https://doi.org/10.3390/e23111532.	IF 2.738

4.6. Zestawienie metryk naukometrycznych przed i po doktoracie

4.6.1. Zestawienie danych naukometrycznych przed doktoratem i po doktoracie

	Przed doktoratem	**Po doktoracie (od lipca 2020)**	**Razem**	**Źródło danych:**
Sumaryczny IF	7.078	178.053	185.131	system expertus Politechniki Bydgoskiej
Liczba Publikacji	19	99	118	system expertus Politechniki Bydgoskiej

Publikacji w WoS	14	53	67	WoS
Cytowań w WoS	144	536	680	WoS
Bez cytowań własnych	143	395	538	WoS
Liczba punktów ministerialnych	1134	8100	9234	system expertus Politechniki Bydgoskiej
H-index WoS	6	13	13	WoS

4.6.2. Sumaryczne zestawienie wskaźników bibliometrycznych (stan na 31.12.2024)

		źródło:
Sumaryczny IF	185.131	system expertus Politechniki Bydgoskiej
łączna liczba prac	118	system expertus Politechniki Bydgoskiej
łączna liczba prac z IF	35	system expertus Politechniki Bydgoskiej
łączna liczba prac z punktacją MNiSW	115	system expertus Politechniki Bydgoskiej
łączna wartość punktacji MNiSW	9234	system expertus Politechniki Bydgoskiej
Publikacji w Scopus	104	Scopus
Publikacji w WoS	67	WoS
Publikacji w Google Scholar	112	Google Scholar
Cytowań w Scopus	1124	Scopus
Cytowań w WoS	680	WoS
Cytowań w Google Scholar	1745	Google Scholar
Bez cytowań własnych	853	Scopus
Bez cytowań własnych	538	WoS
H-index Scopus	17	Scopus
H-index WoS	13	WoS
H-index Google Scholar	22	Google Scholar

Wykorzystana literatura pomocnicza:

[1] H. Zhao, *Cyberspace & Sovereignty*. WORLD SCIENTIFIC, 2022. doi: 10.1142/12027.

[2] A. Pawlicka, M. Choraś, i M. Pawlicki, „The stray sheep of cyberspace a.k.a. the actors who claim they break the law for the greater good”, *Pers Ubiquitous Comput*, t. 25, nr 5, s. 843–852, paź. 2021, doi: 10.1007/s00779-021-01568-7.

[3] „Belgium’s parliament and universities hit by cyber attack | Euronews”, maj 2022. [Online]. Dostępne na: https://www.euronews.com/2021/05/05/belgium-s-parliament-and-universities-hit-by-cyber-attack [dostęp: 15.12.2024]

[4] „Dutch police warn DDoS-for-hire customers to desist or face prosecution | The Daily Swig”, 2022. [Online]. Dostępne na: https://portswigger.net/daily-swig/dutch-police-warn-ddos-for-hire-customers-to-desist-or-face-prosecution [dostęp: 15.12.2024]

[5] „Lessons Learned from Oldsmar Water Plant Hack -- Security Today”, 2021. [Online]. Dostępne na: https://securitytoday.com/articles/2021/04/05/lessons-learned-from-oldsmar-water-plant-hack.aspx [dostęp: 15.12.2024]

[6] „UPDATE: UHS Health System Confirms All US Sites Affected by Ransomware Attack”, 2020. [Online]. Dostępne na: https://healthitsecurity.com/news/uhs-health-system-confirms-all-us-sites-affected-by-ransomware-attack [dostęp: 15.12.2024]

[7] N. Idika i A. Mathur, „A survey of malware detection techniques”, *Purdue University*, 2007.

[8] Y. Sani, A. Mohamedou, K. Ali, A. Farjamfar, M. Azman, i S. Shamsuddin, „An overview of neural networks use in anomaly Intrusion Detection Systems”, w *2009 IEEE Student Conference on Research and Development (SCOReD)*, lis. 2009, s. 89–92. doi: 10.1109/SCORED.2009.5443289.

[9] M.-E. Mihailescu *i in.*, „The Proposition and Evaluation of the RoEduNet-SIMARGL2021 Network Intrusion Detection Dataset”, *Sensors*, t. 21, nr 13, s. 4319, cze. 2021, doi: 10.3390/s21134319.

[10] M. Sarhan, S. Layeghy, N. Moustafa, i M. Portmann, „Netflow datasets for machine learning-based network intrusion detection systems”, w *Big Data Technologies and Applications: 10th EAI International Conference, BDTA 2020, and 13th EAI International Conference on Wireless Internet, WiCON 2020, Virtual Event, December 11, 2020, Proceedings 10*, 2021, s. 117–135.

[11] P. Szumelda, N. Orzechowski, M. Rawski, i A. Janicki, „VHS-22 – A Very Heterogeneous Set of Network Traffic Data for Threat Detection”, w *EICC 2022: Proccedings of the European Interdisciplinary Cybersecurity Conference*, New York, NY, USA: ACM, cze. 2022, s. 72–78. doi: 10.1145/3528580.3532843.

[12] T. Ahmad i M. N. Aziz, „Data preprocessing and feature selection for machine learning intrusion detection systems”, *ICIC Express Lett*, t. 13, nr 2, s. 93–101, 2019.

[13] J. J. Davis i A. J. Clark, „Data preprocessing for anomaly based network intrusion detection: A review”, *Comput Secur*, t. 30, nr 6–7, s. 353–375, 2011.

[14] V. Dutta, M. Choraś, M. Pawlicki, i R. Kozik, „Detection of Cyberattacks Traces in IoT Data", *Journal of Universal Computer Science*, t. 26, nr 11, s. 1422–1434, 2020, [Online]. Dostępne na: https://www.jucs.org/jucs_26_11/detection_of_cyberattacks_traces/jucs_26_11_1422_1434_dutta.pdf [dostęp: 15.12.2024]

[15] W. Jo, S. Kim, C. Lee, i T. Shon, „Packet preprocessing in CNN-based network intrusion detection system", *Electronics (Basel)*, t. 9, nr 7, s. 1151, 2020.

[16] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, i F. Ahmad, „Network intrusion detection system: A systematic study of machine learning and deep learning approaches", *Transactions on Emerging Telecommunications Technologies*, t. 32, nr 1, sty. 2021, doi: 10.1002/ett.4150.

[17] M. N. Chowdhury, K. Ferens, i M. Ferens, „Network intrusion detection using machine learning", w *Proceedings of the International Conference on Security and Management (SAM)*, 2016, s. 30.

[18] J. Li, Y. Qu, F. Chao, H. P. H. Shum, E. S. L. Ho, i L. Yang, „Machine Learning Algorithms for Network Intrusion Detection", w *AI in Cybersecurity*, L. F. Sikos, Red., Cham: Springer International Publishing, 2019, s. 151–179.

[19] M. Pawlicki, R. Kozik, i M. Choraś, „A survey on neural networks for (cyber-) security and (cyber-) security of neural networks", *Neurocomputing*, t. 500, s. 1075–1087, sie. 2022, doi: 10.1016/j.neucom.2022.06.002.

[20] C. Sinclair, L. Pierce, i S. Matzner, „An application of machine learning to network intrusion detection", w *Proceedings 15th Annual Computer Security Applications Conference (ACSAC'99)*, 1999, s. 371–377. doi: 10.1109/CSAC.1999.816048.

[21] M. Zaman i C.-H. Lung, „Evaluation of machine learning techniques for network intrusion detection", w *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, 2018, s. 1–5. doi: 10.1109/NOMS.2018.8406212.

[22] M. Feurer i F. Hutter, „Hyperparameter Optimization", w *Automated Machine Learning: Methods, Systems, Challenges*, F. Hutter, L. Kotthoff, i J. Vanschoren, Red., Cham: Springer International Publishing, 2019, s. 3–33. doi: 10.1007/978-3-030-05318-5_1.

[23] D. Sharma i N. Kumar, „A review on machine learning algorithms, tasks and applications", *International Journal of Advanced Research in Computer Engineering \& Technology (IJARCET)*, t. 6, nr 10, s. 1323–2278, 2017.

[24] Z. Goldfeld i Y. Polyanskiy, „The Information Bottleneck Problem and its Applications in Machine Learning", *IEEE Journal on Selected Areas in Information Theory*, t. 1, nr 1, s. 19–38, 2020, doi: 10.1109/JSAIT.2020.2991561.

[25] M. F. Umer, M. Sher, i Y. Bi, „Flow-based intrusion detection: Techniques and challenges", *Comput Secur*, t. 70, s. 238–254, 2017.

[26] J. Guerra, C. Catania, i E. Veas, „Datasets are not Enough: Challenges in Labeling Network Traffic", 2021.

[27] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, i A. Hotho, „A survey of network-based intrusion detection data sets", *Comput Secur*, t. 86, s. 147–167, 2019, doi: https://doi.org/10.1016/j.cose.2019.06.005.

[28] I. Ghafir, V. Prenosil, J. Svoboda, i M. Hammoudeh, „A Survey on Network Security Monitoring Systems", *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, s. 77–82, 2016.

[29] R. Hofstede *i in.*, „Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX", *IEEE Communications Surveys \& Tutorials*, t. 16, s. 2037–2064, 2014.

[30] A. Pawlicka, D. Jaroszewska-Choras, M. Choras, i M. Pawlicki, „Guidelines for Stego/Malware Detection Tools: Achieving GDPR Compliance", *IEEE Technology and Society Magazine*, t. 39, nr 4, s. 60–70, grudz. 2020, doi: 10.1109/MTS.2020.3031848.

[31] G. Vormayr, J. Fabini, i T. Zseby, „Why are My Flows Different? A Tutorial on Flow Exporters", *IEEE Communications Surveys \& Tutorials*, t. 22, s. 2064–2103, 2020.

[32] N. Kosaraju, S. R. Sankepally, i K. Mallikharjuna Rao, „Categorical Data: Need, Encoding, Selection of Encoding Method and Its Emergence in Machine Learning Models—A Practical Review Study on Heart Disease Prediction Dataset Using Pearson Correlation", w *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 1*, 2023, s. 369–382.

[33] K. Potdar, T. S. Pardawala, i C. D. Pai, „A comparative study of categorical variable encoding techniques for neural network classifiers", *Int J Comput Appl*, t. 175, nr 4, s. 7–9, 2017.

[34] J. T. Hancock i T. M. Khoshgoftaar, „Survey on categorical data for neural networks", *J Big Data*, t. 7, nr 1, s. 1–41, 2020.

[35] E. Jackson i R. Agrawal, „Performance Evaluation of Different Feature Encoding Schemes on Cybersecurity Logs", w *2019 SoutheastCon*, 2019, s. 1–9. doi: 10.1109/SoutheastCon42311.2019.9020560.

[36] C.-H. Lee, Y.-Y. Su, Y.-C. Lin, i S.-J. Lee, „Machine learning based network intrusion detection", w *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*, IEEE, wrz. 2017, s. 79–83. doi: 10.1109/CIAPP.2017.8167184.

[37] N. Shone, T. N. Ngoc, V. D. Phai, i Q. Shi, „A Deep Learning Approach to Network Intrusion Detection", *IEEE Trans Emerg Top Comput Intell*, t. 2, nr 1, s. 41–50, 2018, doi: 10.1109/TETCI.2017.2772792.

[38] S. Gamage i J. Samarabandu, „Deep learning methods in network intrusion detection: A survey and an objective comparison", *Journal of Network and Computer Applications*, t. 169, s. 102767, 2020, doi: https://doi.org/10.1016/j.jnca.2020.102767.

[39] M. Komisarek, M. Pawlicki, R. Kozik, W. Hołubowicz, i M. Choraś, „How to Effectively Collect and Process Network Data for Intrusion Detection?", *Entropy*, t. 23, nr 11, s. 1532, lis. 2021, doi: 10.3390/e23111532.

[40] S. J. Pan i Q. Yang, „A survey on transfer learning", *IEEE Trans Knowl Data Eng*, t. 22, nr 10, s. 1345–1359, 2009.

[41] F. Zhuang *i in.*, „A Comprehensive Survey on Transfer Learning", *Proceedings of the IEEE*, t. 109, nr 1, s. 43–76, 2021, doi: 10.1109/JPROC.2020.3004555.

[42] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, i C. Liu, „A survey on deep transfer learning", w *International conference on artificial neural networks*, 2018, s. 270–279.

[43] S. Dai i F. Meng, „Addressing modern and practical challenges in machine learning: A survey of online federated and transfer learning", *arXiv preprint arXiv:2202.03070*, 2022.

[44] N. Agarwal, A. Sondhi, K. Chopra, i G. Singh, „Transfer learning: Survey and classification", w *Smart innovations in communication and computational sciences*, Springer, 2021, s. 145–155.

[45] O. Day i T. M. Khoshgoftaar, „A survey on heterogeneous transfer learning", *J Big Data*, t. 4, nr 1, s. 1–42, 2017.

[46] C. Cai *i in.*, „Transfer Learning for Drug Discovery", *J Med Chem*, t. 63, nr 16, s. 8683–8694, sie. 2020, doi: 10.1021/acs.jmedchem.9b02147.

[47] D. Chicco, „Siamese neural networks: An overview", *Artificial neural networks*, s. 73–94, 2021.

[48] X. Li, X. Yang, K. Wei, C. Deng, i M. Yang, „Siamese contrastive embedding network for compositional zero-shot learning", w *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, s. 9326–9335.

[49] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, i R. Fergus, „Training convolutional networks with noisy labels", *arXiv preprint arXiv:1406.2080*, 2014.

[50] M. Shorfuzzaman i M. S. Hossain, „MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients", *Pattern Recognit*, t. 113, s. 107700, 2021.

[51] Z. Lian, Y. Li, J. Tao, i J. Huang, „Speech emotion recognition via contrastive loss under siamese networks", w *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, s. 21–26.

[52] M. Jin, Y. Zheng, Y.-F. Li, C. Gong, C. Zhou, i S. Pan, „Multi-scale contrastive siamese networks for self-supervised graph representation learning", *arXiv preprint arXiv:2105.05682*, 2021.

[53] R. Hadsell, S. Chopra, i Y. LeCun, „Dimensionality reduction by learning an invariant mapping", w *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, s. 1735–1742.

[54] M. Choraś Michałand Pawlicki, D. Puchalski, i R. Kozik, „Machine Learning--the results are not the only thing that matters! What about security, explainability and fairness?", w *Computational Science--ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3--5, 2020, Proceedings, Part IV 20*, 2020, s. 615–628.

[55] M. Nauta *i in.*, „From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI", *ACM Comput. Surv.*, luty 2023, doi: 10.1145/3583558.

[56] C. Molnar, G. Casalicchio, i B. Bischl, „Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges", *J Biomed Inform*, t. 113, s. 103655, paź. 2020, doi: 10.1007/978-3-030-65965-3_28.

[57] P. Lopes, E. Silva, C. Braga, T. Oliveira, i L. Rosado, „XAI Systems Evaluation: A Review of Human and Computer-Centred Methods", *Applied Sciences*, t. 12, nr 19, s. 9423, wrz. 2022, doi: 10.3390/app12199423.

[58] S. M. Lundberg i S.-I. Lee, „A Unified Approach to Interpreting Model Predictions", w *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, i R. Garnett, Red., Curran Associates, Inc., 2017. [Online]. Dostępne na: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[59] D. Alvarez-Melis i T. S. Jaakkola, „On the robustness of interpretability methods", *arXiv preprint arXiv:1806.08049*, 2018.

[60] C. Agarwal *i in.*, „Rethinking Stability for Attribution-based Explanations", mar. 2022, [Online]. Dostępne na: http://arxiv.org/abs/2203.06877

[61] C. Agarwal *i in.*, „OpenXAI: Towards a Transparent Evaluation of Model Explanations", cze. 2022, [Online]. Dostępne na: http://arxiv.org/abs/2206.11104

[62] A. Hedström *i in.*, „Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond", *Journal of Machine Learning Research*, t. 24, nr 34, s. 1–11, 2023, [Online]. Dostępne na: http://jmlr.org/papers/v24/22-0142.html

[63] S. Bobek i G. J. Bałaga Pawełand Nalepa, „Towards model-agnostic ensemble explanations", w *International Conference on Computational Science*, 2021, s. 39–51.

[64] K. Rafał, M. Ficco, A. Pawlicka, M. Pawlicki, F. Palmieri, i M. Choraś, „When explainability turns into a threat - using xAI to fool a fake news detection method”, *Comput Secur*, s. 103599, lis. 2023, doi: 10.1016/j.cose.2023.103599.

[65] M. Szczepański, M. Choraś, M. Pawlicki, i A. Pawlicka, „The Methods and Approaches of Explainable Artificial Intelligence”, 2021, s. 3–17. doi: 10.1007/978-3-030-77970-2_1.

[66] A. Adadi i M. Berrada, „Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”, *IEEE Access*, t. 6, s. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

5. Informacja o wykazywaniu się istotną aktywnością naukową albo artystyczną realizowaną w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej.

5.1. Wydział Telekomunikacji, Informatyki i Elektrotechniki Politechniki Bydgoskiej im. Jana i Jędrzeja Śniadeckich

Jestem pracownikiem Politechniki Bydgoskiej. W swoich pracach koncentruję się na wykorzystaniu metod uczenia maszynowego w różnych dziedzinach, w tym w cyberbezpieczeństwie.

5.2. Współpraca z University of Naples PARTHENOPE (staż naukowy)

Od 30 kwietnia 2024 roku do 1 sierpnia 2024 roku (trzy miesiące) odbyłem staż naukowy na University of Naples „Parthenope” w Neapolu, we Włoszech. Zakres zainteresowań badawczych grupy naukowej z Parthenope w dużej mierze pokrywa się z moimi zainteresowaniami badawczymi. W trakcie stażu skoncentrowałem się więc na intensywnej pracy naukowej z zakresu wykorzystania sztucznej inteligencji w cyberbezpieczeństwie.

Ważnym aspektem mojego pobytu było badanie zastosowania technik granular computing w analizie ruchu sieciowego (NetFlow). Wyniki eksperymentów przeprowadzonych w tym temacie przedstawiono na konferencji ARES2024 (CORE B) [N1].

W trakcie stażu opracowałem nową metodę poprawy odporności systemów wykrywania włamań na ataki typu adversarial. Metoda ta polega na pomijaniu wybranych cech oraz zastosowaniu komitetu klasyfikatorów. Praca ta została przyjęta na konferencję ESORICS2024 (CORE A) [N4].

Również w kontekście wykorzystania ML w NIDS w trakcie współpracy z University of Naples PARTHENOPE przeprowadziłem badanie metryk sprawdzających trafność algorytmów xAI [N3].

Podczas pobytu w Neapolu nawiązałem nowe kontakty, które z czasem będą owocować dalszymi publikacjami naukowymi. Zorganizowałem oraz brałem udział w warsztatach dotyczących nowoczesnych metod w cyberbezpieczeństwie oraz sztucznej inteligencji. Umożliwiło to wymianę wiedzy i doświadczeń z innymi badaczami oraz praktykami z branży. Zdobyta w czasie współpracy z włoskim zespołem wiedza pozwoliła mi na zastosowanie nowych metod w moich badaniach oraz podniesienie jakości mojej pracy naukowej. Przeprowadziłem prezentacje dla kadry akademickiej i studentów University of Naples „Parthenope", prezentując wyniki moich badań oraz dzieląc się wiedzą z zakresu sztucznej inteligencji, wyjaśnialności i cyberbezpieczeństwa.

Publikacje wynikające ze stażu z University of Naples PARTHENOPE:

- [N1] Komisarek, M., **Pawlicki, M.**, D'Antonio, S., Kozik, R., Pawlicka, A., & Choraś, M. (2024, July). Enhancing Network Security Through Granular Computing: A Clustering-by-Time Approach to NetFlow Traffic Analysis. In Proceedings of the 19th International Conference on Availability, Reliability and Security (pp. 1-8). **(70 pkt, CORE B)**

- [N2] Uccello, F., **Pawlicki, M.**, D'Antonio, S., Kozik, R., & Choraś, M. (2024, July). Comparative Analysis of AI-Based Methods for Enhancing Cybersecurity Monitoring Systems. In International Conference on Computational Science and Its Applications (pp. 100-112). Cham: Springer Nature Switzerland. (20 pkt)

- [N3] **Pawlicki, M.**, Pawlicka, A., Uccello, F., Szelest, S., D'Antonio, S., Kozik, R., & Choraś, M. (2024). Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. Neurocomputing, 128282. **(140 pkt, IF: 5.5)**

- [N4] **Pawlicki, M.**, Uccello, F., D'Antonio, S., Kozik, R., & Choraś, M. A Novel Method of Improving Intrusion Detection Systems Robustness Against Adversarial Attacks, through Feature Omission and a Committee of Classifiers, (przyjęte na ESORICS Workshops 2024), **(140 pkt, CORE A)**

5.3. Współpraca z innymi uczelniami w Polsce

Po uzyskaniu stopnia doktora zaangażowałem się we współpracę z zespołami akademickimi z Politechniki Warszawskiej i Politechniki Wrocławskiej. Niniejsza podsekcja opisuje charakter tej współpracy wraz z przykładowymi wspólnymi publikacjami.

5.3.1. Politechnika Warszawska

Do wspólnych badań z zespołem z Politechniki Warszawskiej należą studia nad zastosowaniem uczenia głębokiego do zwalczania dezinformacji, kompleksowe przeglądy zagrożeń związanych z malware oraz innowacyjne podejścia do wykrywania fałszywych

informacji za pomocą AI neuro-symbolicznej i modeli rozumienia języka naturalnego. Z badań tych wnikają poniższe wspólne publikacje:

- Kozik, R., Mazurczyk, W., Cabaj, K., Pawlicka, A., **Pawlicki, M.**, & Choraś, M. (2023). Deep Learning for Combating Misinformation in Multicategorical Text Contents. Sensors, 23(24), 9666.
- Caviglione, L., Choraś, M., Corona, I., Janicki, A., Mazurczyk, W., **Pawlicki, M.**, & Wasielewska, K. (2020). Tight arms race: Overview of current malware threats and trends in their detection. IEEE Access, 9, 5371-5396.
- Kozik, R., Pawlicka, A., **Pawlicki, M.**, Choraś, M., Mazurczyk, W., & Cabaj, K. (2023). A meta-analysis of state-of-the-art automated fake news detection methods. IEEE Transactions on Computational Social Systems.
- Kozik, R., Mazurczyk, W., Cabaj, K., Pawlicka, A., **Pawlicki, M**., & Choraś, M. (2023, October). Combating Disinformation with Holistic Architecture, Neuro-symbolic AI and NLU Models. In 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-9). IEEE.

5.3.2. Politechnika Wrocławska

Współpraca z zespołem Politechniki Wrocławskiej koncentrowała się głównie na wyzwaniach związanych z realizacją projektu Infostrateg SWAROG, w którym Politechnika Wrocławska i Politechnika Bydgoska są partnerami w konsorcjum. Poza projektem SWAROG pracowaliśmy nad zastosowaniem AI w cyberbezpieczeństwie urządzeń Internet of Things (IoT). Poniżej zamieszczam wspólne publikacje:

- Zyblewski, P., **Pawlicki, M.**, Kozik, R., & Choraś, M. (2022). Cyber-attack detection from iot benchmark considered as data streams. In Progress in Image Processing, Pattern Recognition and Communication Systems: Proceedings of the Conference (CORES, IP&C, ACS)-June 28-30 2021 12 (pp. 230-239). Springer International Publishing.
- Kozik, R., Komorniczak, J., Ksieniewicz, P., Pawlicka, A., **Pawlicki, M.**, & Choraś, M. (2023, August). SWAROG Project Approach to Fake News Detection Problem. In Computational Intelligence in Security for Information Systems Conference (pp. 79-88). Cham: Springer Nature Switzerland.

5.4. Udział w projektach krajowych i międzynarodowych

Realizacja projektów europejskich wymaga współpracy z różnorodnymi podmiotami, zarówno naukowymi, jak i technologicznymi. Każde doświadczenie w takich przedsięwzięciach było dla mnie unikalne i przyczyniło się do tworzenia nowych, innowacyjnych rozwiązań oraz kształtowania przyszłości badań. Moja rola w tych

projektach obejmowała zarządzanie, koordynowanie działań oraz zaangażowanie jako ekspert AI, co zaowocowało publikacjami i współpracą z wieloma instytucjami z całej Europy.

5.4.1. Pełniłem/pełnię rolę Kierownika Projektu po stronie polskiego partnera (ITTI. sp. z o.o.), pełniąc w nich również rolę Eksperta AI, w następujących projektach (po uzyskaniu doktoratu):

1. Horizon Europe AI4CYBER od 2022-09-01 (trwa)
2. H2020 ELEGANT w okresie od 2021-01-01 do 2023-12-31
3. H2020 APPRAISE w okresie od 2021-09-01 do 2024-02-29
4. H2020 ENSURESEC w okresie od 2020-06-01 do 2022-05-31

Akronim	**HE AI4CYBER**	
Tytuł	Trustworthy Artificial Intelligence for Cybersecurity Reinforcement and System Resilience	
Identyfikator umowy o grant	GA 101070450	
Lata realizacji	2022-2025	
Typ/finansowanie	HORIZON.2.3 - Civil Security for Society HORIZON.2.3.3 - Cybersecurity	
Opis projektu	Projekt dotyczy zastosowań sztucznej inteligencji w cyberbezpieczeństwie	
Rola	● Główny wykonawca i kierownik projektu po stronie polskiego partnera ● Koordynator zadania odpowiedzialnego za stworzenie rozwiązania zapewniającego wyjaśnialność (xAI) komponentów AI użytych w projekcie	

Akronim	**H2020 ELEGANT**	
Tytuł	Secure and Seamless Edge-to-Cloud Analytics	

Identyfikator umowy o grant	GA 957286	
Lata realizacji	2021-2023	
Typ/finansowanie	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT)	
Opis projektu	Projekt dotyczył wyzwań stojących przed urządzeniami IoT i domeną Big Data	
Rola	● Główny wykonawca i kierownik projektu po stronie polskiego partnera ● Koordynator pakietu prac odpowiedzialnego za stworzenie rozwiązania do analizy ruchu sieciowego w czasie rzeczywistym, dostosowanego do pracy z urządzeniami IoT i charakterystycznymi wymaganiami projektu	

Akronim	**H2020 APPRAISE**	
Tytuł	fAcilitating Public & Private secuRity operAtors to mitigate terrorIsm Scenarios against soft targEts	
Identyfikator umowy o grant	GA 101021981	
Lata realizacji	2021-2024	
Typ/finansowanie	SU-FCT03-2018-2019-2020 - Information and data stream management to fight against (cyber)crime and terrorism Funded under H2020-EU.3.7. H2020-EU.3.7.1. H2020-EU.3.7.8.	
Opis projektu	Projekt dotyczy bezpieczeństwa cyber-fizycznego obiektów typu soft target	
Rola	● Główny wykonawca i kierownik projektu po stronie polskiego partnera ● Koordynacja prac mających na celu wytworzenie i wdrożenie komponentów w zakresie detekcji ataków cybernetycznych na obiekty typu soft target	

Akronim	**H2020 ENSURESEC**	
Tytuł	End-to-end Security of the Digital Single Market's E-commerce and Delivery Service Ecosystem	
Identyfikator umowy o grant	883242	
Lata realizacji	2020-2022	
Typ/finansowanie	H2020 INFRA	
Opis projektu	Projekt dotyczy bezpieczeństwa cyber-fizycznego usług z obszaru e-commerce/handlu elektronicznego.	
Rola	● Główny wykonawca i kierownik projektu po stronie polskiego partnera ● Koordynacja prac mających na celu wytworzenie komponentów w zakresie reagowania, łagodzenia skutków i odbudowy systemu narażonego na atak cyber-fizyczny ● Lider zadań mających na celu wytworzenie narzędzi wspomagających: o Zarządzanie reakcją i łagodzeniem skutków ataku z użyciem mechanizmów sztucznej inteligencji o Analizę i audyt po incydencie	

5.4.2. Jako wykonawca brałem udział w poniższych projektach:

a. Projekty kończące się po uzyskaniu stopnia doktora:

1. **Infrostrateg SWARÓG**
   a. Źródło finansowania: NCBR
   b. Budżet: 8 657 006,25 zł
   c. Okres realizacji: od 2021-12-01 (trwa)
   d. Rola w projekcie: Ekspert AI
   e. Nr Projektu: INFOSTRATEG-I/0019/2021

2. **Horyzont 2020 STARLIGHT**
    a. Źródło finansowania: Horizon 2020 - Secure societies
    b. Budżet: € 18 947 200,63
    c. Okres realizacji: od 2021-10-01 (trwa)
    d. Rola w projekcie: Ekspert AI i xAI
    e. Grant Agreement: 101021797

3. **Horizon Europe ULTIMATE**
    a. Źródło finansowania:HORIZON-CL4-2021
    b. Budżet: € 3 529 112,50
    c. Okres realizacji: od 2022-10-01 (trwa)
    d. Rola w projekcie: Ekspert AI i xAI
    e. Grant Agreement: 101070162

4. **ATENA: Artificial inTElligence traiNing progrAmme (2021)**
    a. Źródło finansowania: Spinaker NAWA
    b. Budżet: 369 490,00 PLN
    c. Okres realizacji: 2021-05-01 - 2023-08-31
    d. Rola w projekcie: Wykładowca, trener - Ekspert AI
    e. Nr umowy o dofinansowanie: PPI/SPI/2020/1/00033/U/00001

5. **Horyzont 2020 SPARTA**
    a. Źródło finansowania: Horizon 2020 - INDUSTRIAL LEADERSHIP
    b. Budżet: € 15 999 913,00
    c. Okres realizacji: od 2019-02-01 do 2022-06-30
    d. Rola w projekcie: Ekspert AI i xAI

e. Grant Agreement: 830892

6. **Horyzont 2020 SIMARGL**
    a. Źródło finansowania: Horizon 2020 - INDUSTRIAL LEADERSHIP
    b. Budżet: € 6 076 050,00
    c. Okres realizacji: od 2019-05-01 do 2022-04-30
    d. Rola w projekcie: Ekspert AI i cyberbezpieczeństwa
    e. Grant Agreement: 833042

7. **Horyzont 2020 PREVISION**
    a. Źródło finansowania: Horizon 2020 - Secure societies
    b. Budżet: € 9 040 230,00
    c. Okres realizacji: od 2019-09-01 do 2021-12-31
    d. Rola w projekcie: Ekspert AI i cyberbezpieczeństwa
    e. Grant Agreement: 833115

8. **Horyzont 2020 InfraStress**
    a. Źródło finansowania: Horizon 2020 - Secure societies
    b. Budżet: € 10 137 673,75
    c. Okres realizacji: od 2019-06-01 do 2021-09-30
    d. Rola w projekcie: Ekspert AI i cyberbezpieczeństwa
    e. Grant Agreement: 833088

9. **Horyzont 2020 MAGNETO**
    a. Źródło finansowania: Horizon 2020 - Secure societies
    b. Budżet: € 5 320 475,00
    c. Okres realizacji: od 2018-05-01 do 2021-04-30

d. Rola w projekcie: Ekspert AI

e. Grant Agreement: 786629

10. Horyzont 2020 SocialTruth

a. Źródło finansowania: Horizon 2020 - INDUSTRIAL LEADERSHIP

b. Budżet: € 3 178 110,00

c. Okres realizacji: od 2018-12-01 do 2021-11-30

d. Rola w projekcie: Ekspert AI

e. Grant Agreement: 825477

b. Projekty kończące się przed uzyskaniem stopnia doktora

11. **Horyzont 2020 Q-Rapids**

a. Źródło finansowania: Horizon 2020 - INDUSTRIAL LEADERSHIP

b. Budżet: € 4 997 590,00

c. Okres realizacji: od 2018-03-01 do 2019-10-31

d. Rola w projekcie: Ekspert AI

e. Grant Agreement: 732253

W projektach SPARTA i STARLIGHT zajmowałem się opracowaniem systemów wyjaśnialności sztucznej inteligencji, SPARTA była projektem dotyczącym cyberbezpieczeństwa, STARLIGHT polega na przybliżaniu narzędzi wykorzystujących sztuczną inteligencję personelowi służb porządkowych. SWAROG i SocialTruth dotyczą wykorzystania sztucznej inteligencji do wykrywania fake newsów. Projekt SIMARGL dotyczył wykorzystania sztucznej inteligencji do przeciwdziałania atakom cybernetycznym, PREVISION i InfraStress dotyczyły wykorzystywania sztucznej inteligencji w ochronie infrastruktury krytycznej (InfraStress) i infrastruktury typu '*soft targets*' (PREVISION), Q-Rapids polegał na wykorzystywaniu sztucznej inteligencji do poprawiania jakości oprogramowania.

5.5. Współpraca międzynarodowa i z przemysłem

Według bazy SCOPUS na podstawie prac do końca 2023 roku, 28.8% prac w których jestem współautorem powstała przy współpracy z naukowcami z innych państw/regionów (International collaboration), 67.5% prac zawiera współautorów z przemysłu (Academic-Corporate collaboration). Jest to doświadczenie pochodzące z pracy w projektach unijnych polegających między innymi na zastosowaniu badań we współpracy z przemysłem. Współpraca ta umożliwia dostęp do nowoczesnych technologii i infrastruktury. Pozwala również na realne zastosowanie wyników badań, walidując ich przydatność w świecie rzeczywistym. Poniżej zamieszczam niektóre z publikacji wynikających z mojej współpracy z przemysłem:

Orange Polska (operator sieci)

- Komisarek, M., **Pawlicki, M.,** Kowalski, M., Marzecki, A., Kozik, R., & Choraś, M. (2021, August). Network Intrusion Detection in the Wild-the Orange use case in the SIMARGL project. In Proceedings of the 16th International Conference on Availability, Reliability and Security (pp. 1-7).

CESNET, Praga, Czechy (operator sieci)

- **Pawlicki, M**., Zadnik, M., Kozik, R., & Choraś, M. (2022, June). Analysis and Detection of DDoS Backscatter Using NetFlow Data, Hyperband-Optimised Deep Learning and Explainability Techniques. In International Conference on Artificial Intelligence and Soft Computing (pp. 82-92). Cham: Springer International Publishing.

Telprom, informacijske tehnologije d.o.o., Słowenia (operator sieci)

- Komisarek, M., **Pawlicki**, **M.**, Simic, T., Kavcnik, D., Kozik, R., & Choraś, M. (2023, August). Modern NetFlow network dataset with labeled attacks and detection methods. In Proceedings of the 18th International Conference on Availability, Reliability and Security (pp. 1-8).

RoEduNet, Bukareszt, Rumunia (operator sieci akademickiej)

- Mihailescu, M. E., Mihai, D., Carabas, M., Komisarek, M., **Pawlicki, M.,** Hołubowicz, W., & Kozik, R. (2021). The proposition and evaluation of the roedunet-simargl2021 network intrusion detection dataset. Sensors, 21(13), 4319.

Collins Aerospace, Cork, Irlandia (obsługa teleinformatyczna sieci dla Johnson & Johnson)

- Komisarek, M., **Pawlicki, M**., Soboński, P., Pawlicka, A., Kozik, R., & Choraś, M. (2022). Extending machine learning-based intrusion detection with the imputation method. In Progress in Image Processing, Pattern Recognition and Communication Systems: Proceedings of the Conference (CORES, IP&C, ACS)-June 28-30 2021 12 (pp. 284-292). Springer International Publishing.

6. Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę lub sztukę.

6.1. Nagrody

- W 2024r. otrzymałem „Stypendium Ministra Nauki dla wybitnych młodych naukowców”

- W 2021r. byłem częścią 3-osobowego zespołu nagrodzonego „Nagrodą Naukową Prezydenta Miasta w dziedzinie nauk ścisłych, medycznych i nauk o zdrowiu” wraz z Prof. dr hab. inż. Michałem Chorasiem i dr hab. inż. Rafałem Kozikiem

- Zostałem ośmiokrotnie (6x indywidualnie, 2x zespołowo) nagrodzony przez Rektora PBŚ/UTP za osiągnięcia naukowe

6.2. Popularyzacja nauki

6.2.1. Keynote

- Przeprowadziłem keynote na konferencji **ARES2024 (CORE B)** (pt. „Enhancing Network Cybersecurity with Novel Trustworthy AI Solutions”) https://www.ares-conference.eu/persons/marek-pawlicki

- Przeprowadziłem keynote na konferencji **ESORICS2024 (CORE A)** - (pt. „AI Explainability, Explained”) https://disa-iccs.github.io/

6.2.2. Wystąpienia i wykłady

- Byłem jednym z ekspertów w międzynarodowym panelu dotyczącym bezpieczeństwa AI „CLAIRE AQUA" (pt. „Cybersecurity of AI and AI for Cybersecurity" https://www.youtube.com/live/u44CiZhkbnY?si=_2bcGdT8ah8UoUj2)

- Byłem prelegentem w ramach seminarium „Nowoczesne rozwiązania i technologie sztucznej inteligencji" organizowanego przez Komisję Informatyki i Automatyki oddział PAN w Poznaniu, 22 listopada 2024, prezentacją pt. „Zagadnienie xAI w wykrywaniu ataków sieciowych"

- Byłem prelegentem w ramach seminarium „Informatyka i automatyka wobec wyzwań cyfrowej transformacji przemysłu" organizowanego przez Komisję Informatyki i Automatyki oddział PAN w Poznaniu, 8 kwietnia 2022, prezentacją pt. „Zastosowanie sztucznej inteligencji w informatycznych systemach bezpieczeństwa" https://kiiapoznan.pan.pl/wp-content/uploads/2023/03/KIA_Program_ONLINE_08042022.pdf

- Byłem jednym z ekspertów w debacie plenarnej „Advances in ICT & the society: Threading the thin line between progress, development and mental health" w ramach Research & Innovation Forum (13.04.2023) https://rii-forum.org/wp-content/uploads/2023/04/Rii2023-Program-FULL-01.04.2023-v3.pdf

- Przeprowadziłem wykład „Machine Learning: A Primer" w ramach projektu „ATENA: Artificial inTElligence traiNing prograMme", finansowanego w ramach programu SPINAKER (nabór 2020) przez Narodową Agencję Wymiany Akademickiej ze środków Unii Europejskiej. https://wtie.pbs.edu.pl/pl/nauka/projekty/atena

6.2.3. Wystąpienia w mediach

- Byłem gościem programu „Rozmowa Dnia" w Radiu PiK (jako ekspert od AI i cyberbezpieczeństwa) 16.09.2024 https://www.radiopik.pl/2,123336,polska-jest-najmocniej-atakowanym-krajem-na-swie

- Byłem gościem programu „Pomysł Innowacja Kariera" w Radiu PiK (jako ekspert od AI i cyberbezpieczeństwa) https://archiwum.radiopik.pl/service.go?action=details&show.Record.id=5072101&new=1

- Byłem gościem programu „Regionalny Punkt Widzenia" w Radiu PiK (jako ekspert od AI i cyberbezpieczeństwa) http://www.new.archiwum.radiopik.pl/service.go?action=details&show.Record.id=5063173&new=1

- Byłem gościem programu „Pytać Każdy Może” w Radiu PiK (30 listopada 2022) https://www.radiopik.pl/index.php?idp=145&idx=2&szukaj=&s=27 (jako ekspert od cyberbezpieczeństwa)

6.3. Członkostwa

- Członkostwo w Association for Computing Machinery (ACM) (2024)
- Członkostwo w Association for Information Systems (AIS) (2023)
- Członkostwo w International Neural Network Society (INNS) (2023)
- Jestem członkiem Rady Dyscypliny PBŚ (Informatyka Techniczna i Telekomunikacja)

6.4. Współorganizacja konferencji

6.4.1. Współorganizacja konferencji (Conference Chair) po uzyskaniu stopnia doktora:

- ESORICS 2024 - European Symposium on Research in Computer Security (Workshop Chair)

  https://esorics2024.org/organizing_team-en

- IP&C 2023, 2021 - International Conference on Image Processing & Communications - członek komitetu programowego, członek komitetu organizacyjnego (Program Committee, Organising Committee)

  https://ipc.pwr.edu.pl/

- ICCCI International Conference on Computational Collective Intelligence 2024

  - Special Session: Special Session on Security and reliability of information, networks and social media (SIRENE) 2024 - współorganizator sesji specjalnej

    https://iccci.pwr.edu.pl/2024/session.php?acronym=SIRENE

- ARES2024, 2023, 2022 - International Conference on Availability, Reliability and Security

  - International Workshop on Cyber Use of Information Hiding (CUING) - członek komitetu programowego (Program Committee)

    - 2024: https://2024.ares-conference.eu/cuing
    - 2023: https://2023.ares-conference.eu/workshops/cuing-2023/index.html

        - 2022: https://2022.ares-conference.eu/workshops/cuing-2022/index.html

    - International Workshop on Emerging Network Security (ENS) - członek komitetu programowego (Program Committee)

        - 2024: https://2024.ares-conference.eu/ens

        - 2023: https://2023.ares-conference.eu/workshops-eu-symposium/ens-2023/index.html

        - 2022: https://2022.ares-conference.eu/workshops-eu-symposium/ens-2022/index.html

- ICIC 2022, 2021 International Conference On Intelligent Computing

    - 2022 - Special Session on Intelligent Computing for Cybersecurity: Detection, Mitigation and Response in the protection of Soft Targets and Smart Cities - współorganizator sesji specjalnej

      http://www.ic-icc.cn/2022/Special%20Session.htm

    - 2021 - Special Session on Intelligent Computing for Cybersecurity: Detection, Mitigation and Response - współorganizator sesji specjalnej

      http://www.ic-icc.cn/2021/Special%20Session.htm#ss7

- EICC 2021 - European Interdisciplinary Cybersecurity Conference

    - Special Session on Advanced and Reliable Solutions to Counter Malware and Stegomalware - DETONATOR 2021 - członek komitetu programowego sesji specjalnej (Program committee)

      https://www.fvv.um.si/eicc2021/detonator.html

6.4.2. Przewodniczyłem sesjom (Session Chair) po uzyskaniu stopnia doktora:

- ICIC2021 - International Conference on Intelligent Computing 2021
- ICIC2022 - International Conference on Intelligent Computing 2022
- ISD2022 - International Conference on Information Systems Development 2022
- IJCNN2023 - International Joint Conference on Neural Networks 2023
- DSAA2023 - International Conference on Data Science and Advanced Analytics 2023

- ESORICS2024 - European Symposium on Research in Computer Security (workshop DisA)

6.4.3. Przewodniczyłem sesjom (Session Chair) przed uzyskaniem stopnia doktora:

- ICIC2019 - International Conference on Intelligent Computing 2019

6.5. Publikacje ze studentami

W ramach mojej pracy na stanowisku badawczo-dydaktycznym zajmuję się nie tylko standardowymi obowiązkami takimi jak promowanie prac magisterskich i inżynierskich oraz ich recenzowanie, ale także stawiam duży nacisk na publikowanie wspólnych prac ze studentami. Współautorstwo ze studentami jest dla mnie ważne, ponieważ pozwala im ono zdobyć cenne doświadczenie naukowe i zwiększa ich widoczność w środowisku akademickim. Poniżej znajduje się lista niektórych publikacji, które wspólnie opracowaliśmy w celu zaangażowania studentów w ich rozwój akademicki.

- Kuchczyński, M., Pawlicka, A., **Pawlicki, M.**, & Choraś, M. (2022). Using Machine Learning to Detect the Signs of Radicalization and Hate Speech on Twitter. In Progress in Image Processing, Pattern Recognition and Communication Systems: Proceedings of the Conference (CORES, IP&C, ACS)-June 28-30 2021 12 (pp. 210-218). Springer International Publishing.
- Szczepański, M.**, Pawlicki, M.**, Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. Scientific reports, 11(1), 23705. **(IF2021 = 4.380)**
- Szczepański, M., **Pawlicki, M.**, Kozik, R., & Choraś, M. (2022). Fast hybrid oracle-explainer approach to explainability using optimized search of comprehensible decision trees. In 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-10). IEEE. **(CORE A)**
- Komisarek, M., Choraś, M., Kozik, R., & **Pawlicki, M.** (2020). Real-time stream processing tool for detecting suspicious network patterns using machine learning. In Proceedings of the 15th International Conference on Availability, Reliability and Security (pp. 1-7). **(CORE B)**

6.6. Praca dydaktyczna

Na podstawie przygotowanych przez siebie materiałów prowadzę poniższe zajęcia:

- Wykład z przedmiotu „Podstawy przedsiębiorczości”
- Wykład z przedmiotu „Komunikacja Społeczna i Praca w Grupie”
- Wykład z przedmiotu „Współpraca w Środowisku Wielokulturowym”
- Wykład z przedmiotu „Cyberspace protection methods” (ERASMUS)
- Laboratoria z języka Python („Skryptowe Języki Programowania”)
- Laboratoria z języka JavaScript („Skryptowe Języki Programowania”)
- Tematy projektów do przedmiotu „Projektowanie serwisów sieciowych”

6.7. Wykonane recenzje

Wybrane Czasopisma Naukowe:

Telecommunication Systems (Springer)
Print ISSN: 1018-4864
Electronic ISSN: 1572-9451
Impact Factor (IF): 1.7

Soft Computing (Springer)
Print ISSN: 1432-7643
Electronic ISSN: 1433-7479
Impact Factor (IF): 3.1

Bulletin of the Polish Academy of Sciences: Technical Sciences
ISSN: 2300-1917
Impact Factor (IF): 1.2

Future Generation Computer Systems (Elsevier)
Print ISSN: 0167-739X
Online ISSN: 1872-7115
Impact Factor (IF): 6.2

Artificial Intelligence Review (Springer)
Print ISSN: 0269-2821
Electronic ISSN: 1573-7462
Impact Factor (IF): 10.7

Applied Soft Computing (Elsevier)
Print ISSN: 1568-4946
Online ISSN: 1872-9681
Impact Factor (IF): 7.2

Foundations of Computing and Decision Sciences
Print ISSN: 0867-6356
Electronic ISSN: 2300-3405
Impact Factor (IF): 1.8

Applied Artificial Intelligence (Taylor & Francis)
Print ISSN: 0883-9514
Online ISSN: 1087-6545
Impact Factor (IF): 2.9

Machine Learning (Springer)
Print ISSN: 0885-6125
Electronic ISSN: 1573-0565
Impact Factor (IF): 4.3

Neurocomputing (Elsevier)
Print ISSN: 0925-2312
Online ISSN: 1872-8286
Impact Factor (IF): 5.5

Applied Sciences (MDPI)
ISSN: 2076-3417
Impact Factor (IF): 2.5

The Journal of Universal Computer Science (J.UCS)
Print ISSN: 0948-695X
Online ISSN: 0948-6968
Impact Factor (IF): 0.7

International Journal of Applied Mathematics and Computer Science
Print ISSN: 1641-876X
Online ISSN: 2083-8492
Impact Factor (IF): 1.6

Wybrane Konferencje:

ESORICS
Pełna nazwa: European Symposium on Research in Computer Security
CORE Ranking: A

ECAI
Pełna nazwa: European Conference on Artificial Intelligence
CORE Ranking: A

IEEE WCCI 2022
Pełna nazwa: IEEE World Congress on Computational Intelligence
CORE Ranking: A

ARES
Pełna nazwa: International Conference on Availability, Reliability and Security
CORE Ranking: B

DSAA 2023
Pełna nazwa: IEEE International Conference on Data Science and Advanced Analytics
CORE Ranking: B

ICCCI
Pełna nazwa: International Conference on Computational Collective Intelligence
CORE Ranking: B

IJCNN
Pełna nazwa: International Joint Conference on Neural Networks
CORE Ranking: B

ICIC
Pełna nazwa: International Conference on Intelligent Computing
CORE Ranking: -

MobiSec

Pełna nazwa: International Conference on Mobile Internet Security
CORE Ranking: -

ICAI
Pełna nazwa: International Conference on Applied Intelligence
CORE Ranking: -

IP&C
Pełna nazwa: International Conference on Image Processing and Communications
CORE Ranking: -

IEEE CSR
Pełna nazwa: IEEE International Conference on Cyber Security and Resilience
CORE Ranking: -

EICC
Pełna nazwa: European Interdisciplinary Cybersecurity Conference
CORE Ranking: -

ISD
Pełna nazwa: International Conference on Information Systems Development
CORE Ranking: -

6.8. Prace Marka Pawlickiego spoza cyklu

- Kozik, R., Pawlicka, A., Pawlicki, M., Choraś, M., Mazurczyk, W., & Cabaj, K. (2024). A meta-analysis of state-of-the-art automated fake news detection methods. IEEE Transactions on Computational Social Systems, 5219–5229. https://doi.org/10.1109/TCSS.2023.3296627

- Uccello, F., Pawlicki, M., D'Antonio, S., Kozik, R., & Choraś, M. (2024). A novel approach to the use of explainability to mine network intrusion detection rules. W N. T. Nguyen i in. (red.), Intelligent Information and Database Systems (ss. 70–81). Springer. https://doi.org/10.1007/978-981-97-4982-9_6

- Pawlicki, M., Pawlicka, A., Kozik, R., & Choraś, M. (2024). Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep

learning and artificial intelligence used in cybersecurity. Neurocomputing, 1–13. https://doi.org/10.1016/j.neucom.2024.127759

- Pawlicka, A., Pawlicki, M., Kozik, R., Andrychowicz-Trojanowska, A., & Choraś, M. (2024). AI vs linguistic-based human judgement: Bridging the gap in pursuit of truth for fake news detection. Information Sciences, 1–29. https://doi.org/10.1016/j.ins.2024.121097

- Uccello, F., Pawlicki, M., D'Antonio, S., Kozik, R., & Choraś, M. (2024). An innovative approach to real-time concept drift detection in network security. W L. Barolli (red.), Advances in Internet, Data & Web Technologies. EIDWT 2024 (ss. 130–139). Springer. https://doi.org/10.1007/978-3-031-53555-0_13

- Uccello, F., Pawlicki, M., D'Antonio, S., Kozik, R., & Choraś, M. (2024). Comparative analysis of AI-based methods for enhancing cybersecurity monitoring systems. W O. Gervasi i in. (red.), Computational Science and Its Applications – ICCSA 2024 Workshops; Proceedings, Part II (ss. 100–112). Springer. https://doi.org/10.1007/978-3-031-65223-3_7

- Uccello, F., Pawlicki, M., D'Antonio, S., Kozik, R., & Choraś, M. (2024). Effective rules for a rule-based SIEM system in detecting DoS attacks: An association rule mining approach. W D. S. Huang, P. Premaratne & C. Yuan (red.), Applied Intelligence. ICAI 2023 (ss. 236–246). Springer. https://doi.org/10.1007/978-981-97-0827-7_21

- Pawlicka, A., Pawlicki, M., Jaroszewska-Choraś, D., Kozik, R., & Choraś, M. (2024). Enhancing clinical trust: The role of AI explainability in transforming healthcare. W Y. He, W. Hamidouche, I. Razzak i in. (red.), Proceedings 24th IEEE International Conference on Data Mining; Workshops (ss. 543–548). IEEE. https://doi.org/10.1109/ICDMW65004.2024.00075

- Komisarek, M., Pawlicki, M., D'Antonio, S., Kozik, R., Pawlicka, A., & Choraś, M. (2024). Enhancing network security through granular computing: A clustering-by-time approach to NetFlow traffic analysis. W Proceedings of the 19th International Conference on Availability, Reliability and Security (ss. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3664476.3670882

– Pawlicki, M., Pawlicka, A., Uccello, F., Szelest, S., D'Antonio, S., Kozik, R., & Choraś, M. (2024). Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. Neurocomputing, 1–19. https://doi.org/10.1016/j.neucom.2024.128282

– Kozik, R., Pawlicka, A., Pawlicki, M., & Choraś, M. (2024). From detection through display to understanding: Bridging AI and UI in disinformation and fake news analysis. W N. T. Nguyen i in. (red.), Advances in Computational Collective Intelligence 2024 (ss. 347–357). Springer. https://doi.org/10.1007/978-3-031-70248-8_27

– Pawlicka, A., Pawlicki, M., Kozik, R., Kurek, W., & Choraś, M. (2024). How explainable is explainability? Towards better metrics for explainable AI. W A. Visvizi, O. Troisi & V. Corvello (red.), Research and Innovation Forum 2023 (ss. 685–695). Springer. https://doi.org/10.1007/978-3-031-44721-1_52

– Pawlicki, M., Puchalski, D., Szelest, S., Pawlicka, A., Kozik, R., & Choraś, M. (2024). Introducing a multi-perspective xAI tool for better model explainability. W Proceedings of the 19th International Conference on Availability, Reliability and Security (ss. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3664476.3670905

– Kątek, G., Gackowska, M., Komorniczak, J., Ksieniewicz, P., Kozik, R., Pawlicki, M., & Choraś, M. (2024). Involving society to protect society from fake news and disinformation: Crowdsourced datasets and text reliability assessment. W N. T. Nguyen i in. (red.), Intelligent Information and Database Systems (ss. 384–395). Springer. https://doi.org/10.1007/978-981-97-4985-0_30

– Choraś, M., Pawlicka, A., Jaroszewska-Choraś, D., & Pawlicki, M. (2024). Not only security and privacy: The evolving ethical and legal challenges of e-commerce. W S. Katsikas i in. (red.), Computer Security. ESORICS 2023 International Workshops (ss. 167–181). Springer. https://doi.org/10.1007/978-3-031-54204-6_9

– Kozik, R., Buś, S., Gawdzik, G., Pawlicki, M., Pawlicka, A., & Choraś, M. (2024). Real-world application facilitating trustworthy human-robot collaboration with innovative explainable neuro-symbolic reasoning. W Y. He, W. Hamidouche, I. Razzak i in. (red.), Proceedings 24th

IEEE International Conference on Data Mining; Workshops (ss. 364–371). IEEE. https://doi.org/10.1109/ICDMW65004.2024.00053

- Pawlicki, M. (2023). Strengths And Weaknesses of Deep, Convolutional and Recurrent Neural Networks in Network Intrusion Detection Deployments. W A. R. da Silva, M. M. da Silva, J. Estima, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), Information Systems Development, Organizational Aspects and Societal Trends (ISD2023 Proceedings). Lisbon, Portugal: Instituto Superior Técnico. ISBN: 978-989-33-5509-1. https://doi.org/10.62036/ISD.2023.54

- Pawlicki, M., Pawlicka, A., Choraś, R., Kozik, R., & Choraś, M. (2024). Semi-supervised learning trumps active learning in cold-starting netflow-based AI network intrusion detection systems on scarce and difficult data. W Y. He, W. Hamidouche, I. Razzak I in. (red.), Proceedings 24th IEEE International Conference on Data Mining; Workshops (ss. 221–228). IEEE. https://doi.org/10.1109/ICDMW65004.2024.00035

- Choraś, M., Pawlicki, M., & Kozik, R., & Pawlicka, A. (2024). Tech developers must respect equitable AI access. Nature, 1. https://doi.org/10.1038/d41586-024-00185-7

- Pawlicki, M., Kozik, R., & Choraś, M. (2024). The impact of data scaling approaches on deep learning, random forest and nearest neighbour-based network intrusion detection systems for DoS detection in IoT networks. W I. You & M. Choraś (red.), Mobile Internet Security (ss. 197–208). Springer. https://doi.org/10.1007/978-981-97-4465-7_14

- Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, M. (2024). The rise of AI-powered writing: How ChatGPT is revolutionizing scientific communication for better or for worse. W D. S. Huang, P. Premaratne & C. Yuan (red.), Applied Intelligence. ICAI 2023 (ss. 317–327). Springer. https://doi.org/10.1007/978-981-97-0903-8_30

- Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, M. (2024). To teach to hack or not: The grey zone of cybersecurity education. W B. J. Reithel (red.), Proceedings of the fortieth INFORMATION SYSTEMS EDUCATION CONFERENCE: ISECON 2024 (ss. 69–82).

- Kozik, R., Kątek, G., Gackowska, M., Kula, S., Komorniczak, J., Ksieniewicz, P., Pawlicka, A., Pawlicki, M., & Choraś, M. (2024). Towards explainable fake news detection and automated content credibility assessment: Polish internet and digital media use-case. Neurocomputing. https://doi.org/10.1016/j.neucom.2024.128450

- Uccello, F., Pawlicki, M., D'Antonio, S., Kozik, R., & Choraś, M. (2024). Towards hybrid NIDS: Combining rule-based SIEM with AI-based intrusion detectors. W K. Daimi & A. Al Sadoon (red.), Proceedings of the Second International Conference on Advances in Computing Research (ACR'24) (ss. 244–255). Springer. https://doi.org/10.1007/978-3-031-56950-0_21

- Karampasi, A., Radoglou, P., Pawlicki, M., Choraś, R., Martin, R., Sarigiannidis, P., Puchalski, D., Pawlicka, A., Kozik, R., & Choraś, M. (2024). Towards transparent AI-powered cybersecurity in financial systems: The deployment of federated learning and explainable AI in the CaixaBank pilot. W Y. He, W. Hamidouche, I. Razzak I in. (red.), Proceedings 24th IEEE International Conference on Data Mining; Workshops (ss. 270–277). IEEE. https://doi.org/10.1109/ICDMW65004.2024.00041

- Puchalski, D., Pawlicki, M., Kozik, R., Renk, R., & Choraś, M. (2024). Trustworthy AI-based cyber-attack detector for network cyber crime forensics. W Proceedings of the 19th International Conference on Availability, Reliability and Security (ss. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3664476.3670880

- Kozik, R., Puchalski, D., Pawlicka, A., Buś, S., Główka, J., Chandramouli, K., Tiemann, M., Pawlicki, M., Renk, R., & Choraś, M. (2024). ULTIMATE project toolkit for robotic AI-based data analysis and visualization. W N. T. Nguyen I in. (red.), Intelligent Information and Database Systems (ss. 44–55). Springer. https://doi.org/10.1007/978-981-97-4985-0_4

- Kozik, R., Ficco, M., Pawlicka, A., Pawlicki, M., Palmieri, F., & Choraś, M. (2024). When explainability turns into a threat – using xAI to fool a fake news detection method. Computers & Security, 1–28. https://doi.org/10.1016/j.cose.2023.103599

- Pawlicki, M., Zadnik, M., Kozik, R., & Choraś, M. (2023). Analysis and detection of DDoS backscatter using NetFlow data, hyperband-optimised deep learning and explainability techniques. W L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz & J. M. Zurada (red.), Artificial Intelligence and Soft Computing. Springer. https://doi.org/10.1007/978-3-031-23492-7_8

- Kozik, R., Mazurczyk, W., Cabaj, K., Pawlicka, A., Pawlicki, M., & Choraś, M. (2023). Combating disinformation with holistic architecture, neuro-symbolic AI and NLU models. W 10th International Conference on Data Science and Advanced Analytics (DSAA) (ss. 1–9). IEEE. https://doi.org/10.1109/DSAA60987.2023.10302543

- Kozik, R., Mazurczyk, W., Cabaj, K., Pawlicka, A., Pawlicki, M., & Choraś, M. (2023). Deep Learning for Combating Misinformation in Multicategorical Text Contents. Sensors, 1–13. https://doi.org/10.3390/s23249666

- Kurek, W., Pawlicki, M., Pawlicka, A., Kozik, R., & Choraś, M. (2023). Explainable Artificial Intelligence 101: Techniques, Applications and Challenges. W D. S. Huang, P. Premaratne, B. Jin I in. (red.), Advanced Intelligent Computing Technology and Applications (ss. 310–318). Springer. https://doi.org/10.1007/978-981-99-4752-2_26

- Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, M. (2023). Fake news and threats to IoT – The crucial aspects of cyberspace in the times of cyberwar. W A. Visvizi, O. Troisi & M. Grimaldi (red.), Research and Innovation Forum 2022 (ss. 31–38). Springer. https://doi.org/10.1007/978-3-031-19560-0_3

- Pawlicka, A., Choraś, M., Kozik, R., & Pawlicki, M. (2023). First broad and systematic horizon scanning campaign and study to detect societal and ethical dilemmas and emerging issues spanning over cybersecurity solutions. Personal and Ubiquitous Computing, 193–202. https://doi.org/10.1007/s00779-020-01510-3

- Pawlicka, A., Tomaszewska, R., Krause, E., Jaroszewska-Choraś, D., Pawlicki, M., & Choraś, M. (2023). Has the pandemic made us more digitally literate? Innovative association rule mining study of the relationships between shifts in digital skills and cybersecurity awareness occurring whilst working remotely during the COVID-19 pandemic. Journal of Ambient

Intelligence and Humanized Computing, 14721–14731. https://doi.org/10.1007/s12652-022-04371-1

- Choraś, M., Pawlicka, A., Pawlicki, M., & Kozik, R. (2023). How to Navigate the Uncharted Waters of Cyberwar. W AMCIS 2023 Proceedings. 1. Association for Information Systems. https://aisel.aisnet.org/amcis2023/conf_theme/conf_theme/1

- Pawlicka, A., Puchalski, D., Pawlicki, M., Kozik, R., & Choraś, M. (2023). How to secure the IoT-based surveillance systems in an ELEGANT way. W 2023 IEEE International Conference on Cyber Security and Resilience (CSR) (ss. 636–640). IEEE. https://doi.org/10.1109/CSR57506.2023.10224938

- Pawlicki, M., Pawlicka, A., Śrutek, M., Kozik, R., & Choraś, M. (2023). Interpreting Intrusions – The Role of Explainability in AI-Based Intrusion Detection Systems. W Progress on Pattern Classification, Image Processing and Communications. CORES IP&C 2023 (ss. 45–53). Springer. https://doi.org/10.1007/978-3-031-41630-9_5

- Kozik, R., Pawlicka, A., Pawlicki, M., & Choraś, M. (2023). Model Stitching Algorithm for Fake News Detection Problem. W 10th International Conference on Data Science and Advanced Analytics (DSAA) (ss. 1–7). IEEE. https://doi.org/10.1109/DSAA60987.2023.10302574

- Komisarek, M., Pawlicki, M., Simic, T., Kavcnik, D., Kozik, R., & Choraś, M. (2023). Modern NetFlow network dataset with labeled attacks and detection methods. W Proceedings of the 18th International Conference on Availability, Reliability and Security (ss. 1–8). Association for Computing Machinery.

- Kozik, R., Samp, K., Choraś, M., & Pawlicki, M. (2023). Parameters Transfer Framework for Multi-domain Fake News Detection. W I. You, H. Kim & P. Angin (red.), Mobile Internet Security. MobiSec 2022 (ss. 85–96). Springer. https://doi.org/10.1007/978-981-99-4430-9_6

- Kozik, R., Komorniczak, J., Ksieniewicz, P., Pawlicka, A., Pawlicki, M., & Choraś, M. (2023). SWAROG Project Approach to Fake News Detection Problem. W P. García Bringas I in. (red.), 16th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2023) (ss. 79–88). Springer. https://doi.org/10.1007/978-3-031-42519-6_8

- Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2023). The Application of Deep Learning Imputation and Other Advanced Methods for Handling Missing Values in Network Intrusion Detection. Vietnam Journal of Computer Science, 1–23. https://doi.org/10.1142/S2196888822500257

- Pawlicki, M., Kozik, R., & Choraś, M. (2023). The Impact of Data Scaling Approaches on Deep Learning, Random Forest and Nearest Neighbour-Based Network Intrusion Detection Systems for DoS Detection in IoT Networks. W The 7th International Conference on Mobile Internet Security (MobiSec'23). https://doi.org/10.1007/978-981-97-4465-7_14

- Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, M. (2023). The Need for Practical Legal and Ethical Guidelines for Explainable AI-based Network Intrusion Detection Systems. W J. Wang, Y. He, T. N. Dinh, C. Grant, M. Qiu & W. Pedrycz (red.), 23rd IEEE International Conference on Data Mining Workshops (ss. 253–261). IEEE. https://doi.org/10.1109/ICDMW60847.2023.00038

- Pawlicki, M., Pawlicka, A., Kozik, R., & Choraś, M. (2023). The survey and meta-analysis of the attacks, transgressions, countermeasures and security aspects common to the Cloud, Edge and IoT. Neurocomputing, 1–18. https://doi.org/10.1016/j.neucom.2023.126533

- Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, M. (2023). What will the future of cybersecurity bring us, and will it be ethical? The hunt for the black swans of cybersecurity ethics. IEEE Access, 58796–58807. https://doi.org/10.1109/ACCESS.2023.3283791

- Komisarek, M., Pawlicki, M., Mihailescu, M.-E., Mihai, D., Carabas, M., Kozik, R., & Choraś, M. (2022). A novel, refined dataset for real-time Network Intrusion Detection. W Proceedings of the 17th International Conference on Availability, Reliability and Security (ss. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3538969.3544486

- Pawlicki, M., Kozik, R., & Choraś, M. (2022). Comparison of neural network paradigms for their usefulness in a real-world network intrusion detection deployment and a proposed 78ptimized approach. W D. Megías & R. Di Pietro (red.), Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference (ss. 53–56). Association for Computing Machinery. https://doi.org/10.1145/3528580.3532840

- Zyblewski, P., Pawlicki, M., Kozik, R., & Choraś, M. (2022). Cyber-Attack Detection from IoT Benchmark Considered as Data Streams. W M. Choraś, R. S. Choraś, M. Kurzyński, P. Trajdos, J. Pejaś & T. Hyla (red.), Progress in Image Processing, Pattern Recognition and Communication Systems (ss. 230–239). Springer. https://doi.org/10.1007/978-3-030-81523-3

- Kozik, R., Pawlicki, M., Szczepański, M., Renk, R., & Choraś, M. (2022). Efficient Post Event Analysis and Cyber Incident Response in IoT and E-commerce Through Innovative Graphs and Cyberthreat Intelligence Employment. W D. S. Huang, K. H. Jo, J. Jing, P. Premaratne, V. Bevilacqua & A. Hussain (red.), Intelligent Computing Methodologies. ICIC 2022 (ss. 257–266). Springer. https://doi.org/10.1007/978-3-031-13832-4_22

- Komisarek, M., Pawlicki, M., Soboński, P., Pawlicka, A., Kozik, R., & Choraś, M. (2022). Extending Machine Learning-Based Intrusion Detection with the Imputation Method. W M. Choraś, R. S. Choraś, M. Kurzyński, P. Trajdos, J. Pejaś & T. Hyla (red.), Progress in Image Processing, Pattern Recognition and Communication Systems (ss. 284–294). Springer. https://doi.org/10.1007/978-3-030-81523-3_28

- Kozik, R., Pawlicki, M., Kula, S., & Choraś, M. (2022). Fake News Detection Platform – Conceptual Architecture and Prototype. Logic Journal of the IGPL, 1005–1016. https://doi.org/10.1093/jigpal/jzac009

- Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2022). Fast Hybrid Oracle-Explainer Approach to Explainability Using Optimized Search of Comprehensible Decision Trees. W J. Z. Huang, Y. Pan, B. Hammer I in. (red.), 9th International Conference on Data Science and Advanced Analytics (DSAA) (ss. 1–10). IEEE. https://doi.org/10.1109/DSAA54385.2022.10032372

- Pawlicki, M., Pawlicka, A., Kozik, R., & Choraś, M. (2022). Graph Databases and E-commerce Cybersecurity – A Match Made in Heaven? The Innovative Technology to Enhance Cyberthreat Mitigation. W R. A. Buchmann I in. (red.), Information Systems Development: Artificial Intelligence for Information Systems Development and Operations (ISD2022 Proceedings). Babe-Bolyai University. https://doi.org/10.62036/ISD.2022.20

- Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, M. (2022). Human-driven and human-centred cybersecurity: Policy-making implications. Transforming Government: People, Process and Policy, 478–487. https://doi.org/10.1108/TG-05-2022-0073

- Komisarek, M., Pawlicki, M., Kowalski, M., Marzecki, A., Kozik, R., & Choraś, M. (2022). Hunting cyberattacks: Experience from the real backbone network. JoWUA, 128–146. https://doi.org/10.22667/JOWUA.2022.06.30.128

- Pawlicka, A., Choraś, M., & Pawlicki, M. (2022). Opening up a discussion on the native speaker myth in science. Accountability in Research, 477–481. https://doi.org/10.1080/08989621.2021.1960514

- Choraś, M., Kozik, R., & Pawlicki, M. (2022). Sensors and Pattern Recognition Methods for Security and Industrial Applications. Sensors, 1–3. https://doi.org/10.3390/s22165968

- Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, M. (2022). SIMARGL – Secure Intelligent Methods for Advanced RecoGnition of malware and stegomalware. W M. Grana (red.), Proceedings of the Basque Conference on Cyber-Physical Systems and Artificial Intelligence. CybSPEED. https://doi.org/10.5281/zenodo.6558473

- Pawlicka, A., Pawlicki, M., Renk, R., Kozik, R., & Choraś, M. (2022). The cybersecurity-related ethical issues of cloud technology and how to avoid them. W Proceedings of the 17th International Conference on Availability, Reliability and Security (ss. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3538969.3544456

- Pawlicki, M., Pawlicka, A., Komisarek, M., Kozik, R., & Choraś, M. (2022). The IoT Threat Landscape vs. Machine Learning, a.k.a. Who Attacks IoT, Why do They Do It, and How to Prevent It? W R. A. Buchmann I in. (red.), Information Systems Development: Artificial Intelligence for Information Systems Development and Operations (ISD2022 Proceedings). Babe-Bolyai University. https://doi.org/10.62036/ISD.2022.47

- Komisarek, M., Kozik, R., Pawlicki, M., & Choraś, M. (2022). Towards Zero-Shot Flow-Based Cyber-Security Anomaly Detection Framework. Applied Sciences, 1–12. https://doi.org/10.3390/app12199636

- Dutta, V., Pawlicki, M., Kozik, R., & Choraś, M. (2022). Unsupervised network traffic anomaly detection with deep autoencoders. Logic Journal of the IGPL, 912–925. https://doi.org/10.1093/jigpal/jzac002

- Kuchczyński, M., Pawlicka, A., Pawlicki, M., & Choraś, M. (2022). Using Machine Learning to Detect the Signs of Radicalization and Hate Speech on Twitter. W M. Choraś, R. S. Choraś, M. Kurzyński, P. Trajdos, J. Pejaś & T. Hyla (red.), Progress in Image Processing, Pattern Recognition and Communication Systems (ss. 210–218). Springer. https://doi.org/10.1007/978-3-030-81523-3_21

- Pawlicka, A., Choraś, M., & Pawlicki, M., Kozik, R. (2021). A $10 million question and other cybersecurity-related ethical dilemmas amid the COVID-19 pandemic. Business Horizons, 729–734. https://doi.org/10.1016/j.bushor.2021.07.010

- Kozik, R., Pawlicki, M., & Choraś, M. (2021). A new method of hybrid time window embedding with transformer-based traffic data classification in IoT-networked environment. Pattern Analysis and Applications, 1441–1449. https://doi.org/10.1007/s10044-021-00980-2

- Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, R. (2021). A Systematic Review of Recommender Systems and Their Applications in Cybersecurity. Sensors, 1–25. https://doi.org/10.3390/s21155248

- Kozik, R., Choraś, M., Kula, S., & Pawlicki, M. (2021). Distributed Architecture for Fake News Detection. W Á. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián & E. Corchado (red.), Proceedings of 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) (ss. 208–217). Springer. https://doi.org/10.1007/978-3-030-57805-3_20

- Komisarek, M., Pawlicki, M., Kozik, R., Hołubowicz, W., & Choraś, M. (2021). How to Effectively Collect and Process Network Data for Intrusion Detection? Entropy, 1–25. https://doi.org/10.3390/e23111532

- Dutta, V., Choraś, M., Kozik, R., & Pawlicki, M. (2021). Hybrid Model for Improving the Classification Effectiveness of Network Intrusion Detection. W Á. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián & E. Corchado (red.), Proceedings of 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) (ss. 405–414). Springer. https://doi.org/10.1007/978-3-030-57805-3_38

- Komisarek, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). Machine Learning Based Approach to Anomaly and Cyberattack Detection in Streamed Network Traffic Data. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Application, 3–19. https://doi.org/10.22667/JOWUA.2021.03.31.003

- Pawlicki, M., Choraś, M., Kozik, R., & Hołubowicz, W. (2021). Missing and Incomplete Data Handling in Cybersecurity Applications. W N. T. Nguyen, S. Chittayasothorn, D. Niyato & B. Trawiński (red.), Intelligent Information and Database Systems (ss. 413–427). Springer. https://doi.org/10.1007/978-3-030-73280-6_33

- Komisarek, M., Pawlicki, M., Kowalski, M., Marzecki, A., Kozik, R., & Choraś, M. (2021). Network Intrusion Detection in the Wild – the Orange Use Case in the SIMARGL Project. W The 16th International Conference on Availability, Reliability and Security (7 ss.). Association for Computing Machinery. https://doi.org/10.1145/3465481.3470091

- Szczepański, M., Pawlicki, M., Kozik, R., Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. Scientific Reports, 1–13. https://doi.org/10.1038/s41598-021-03100-6

- Pawlicki, M., & Choraś, R. (2021). Preprocessing Pipelines Including Block-Matching Convolutional Neural Network for Image Denoising to Robustify Deep Reidentification against Evasion Attacks. Entropy, 14. https://doi.org/10.3390/e23101304

- Kozik, R., Choraś, M., Pawlicki, M., Pawlicka, A., Warczak, W., & Mazgaj, G. (2021). Proposition of Innovative and Scalable Information System for Call Detail Records Analysis and Visualisation. W Á. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián & E. Corchado (red.), Proceedings of 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) (ss. 174–183). Springer. https://doi.org/10.1007/978-3-030-57805-3_17

- Szczepański, M., Choraś, M., Pawlicki, M., & Pawlicka, A. (2021). The Methods and Approaches of Explainable Artificial Intelligence. W M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra & P. M. A. Sloot (red.), Computational Science – ICCS 2021 part IV (ss. 3–17). Springer. https://doi.org/10.1007/978-3-030-77970-2_1

- Mihailescu, M.-E., Mihai, D., Carabas, M., Komisarek, M., Pawlicki, M., Hołubowicz, W., Kozik, R. (2021). The Proposition and Evaluation of the RoEduNet-SIMARGL2021 Network Intrusion Detection Dataset. Sensors, 1–20. https://doi.org/10.3390/s21134319

- Szczepański, M., Komisarek, M., Pawlicki, Kozik, R., M., & Choraś, M. (2021). The Proposition of Balanced and Explainable Surrogate Method for Network Intrusion Detection in Streamed Real Difficult Data. W K. Wojtkiewicz, J. Treur, E. Pimenidis & M. Maleszka (red.), Computational Collective Intelligence 13th International Conference (ss. 241–252). Springer. https://doi.org/10.1007/978-3-030-88113-9_19

- Pawlicka, A., Choraś, M., & Pawlicki, M. (2021). The stray sheep of cyberspace a.k.a. the actors who claim they break the law for the greater good. Personal and Ubiquitous Computing, 843–852. https://doi.org/10.1007/s00779-021-01568-7

- Caviglione, L., Choraś, M., Corona, I., Janicki, A., Mazurczyk, W., Pawlicki, M., & Wasielewska, K. (2021). Tight Arms Race: Overview of Current Malware Threats and Trends in Their Detection. IEEE Access, 5371–5396. https://doi.org/10.1109/ACCESS.2020.3048319

- Pawlicki, M., Kozik, R., Puchalski, D., & Choraś, M. (2021). Towards AI-Based Reaction and Mitigation for e-Commerce – the ENSURESEC Engine. W D. S. Huang, V. Bevilacqua & P. Premaratne (red.), Intelligent Computing Theories and Application (ss. 24–31). Springer. https://doi.org/10.1007/978-3-030-84532-2_3

- Dutta, V., Choraś, M., Pawlicki, M., & Kozik, R. (2020). A Deep Learning Ensemble for Network Anomaly and Cyber-Attack Detection. Sensors, 1–20. https://doi.org/10.3390/s20164583

- Szczepański, M., Choraś, M., Pawlicki, M., & Kozik, R. (2020). Achieving Explainability of Intrusion Detection System by Hybrid Oracle-Explainer Approach. W 2020 IEEE International Joint Conference on Neural Networks Proceedings (ss. 58–65). IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207199

- Kozik, R., Choraś, M., Pawlicki, M., Hołubowicz, W., Pallmer, D., Mueller, W., Behmer, E.-J., Loumiotis, I., Demestichas, K., Horincar, R., Laudy, C., & Faure, D. (2020). Common Representational Model and Ontologies for Effective Law Enforcement Solutions. Vietnam Journal of Computer Science, 1–18. https://doi.org/10.1142/S2196888820020017

- Pawlicka, A., Choraś, M., & Pawlicki, M. (2020). Cyberspace threats: not only hackers and criminals. Raising the awareness of selected unusual cyberspace actors – cybersecurity researchers' perspective. W ARES '20: Proceedings of the 15th International Conference on Availability, Reliability and Security (ss. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3407023.3409181

- Pawlicki, M., Choraś, M., & Kozik, R. (2020). Defending network intrusion detection systems against adversarial evasion attacks. Future Generation Computer Systems, 148–154. https://doi.org/10.1016/j.future.2020.04.013

- Dutta, V., Choraś, M., Pawlicki, M., & Kozik, R. (2020). Detection of Cyberattacks Traces in IoT Data. Journal of Universal Computer Science, 1422–1434.

- Pawlicki, M., Giełczyk, A., Kozik, R., & Choraś, M. (2020). Fault-Prone Software Classes Recognition via Artificial Neural Network with Granular Dataset Balancing. W R. Burduk, M. Kurzyński & M. Woźniak (red.), Progress in Computer Recognition Systems (ss. 130–140). Springer. https://doi.org/10.1007/978-3-030-19738-4_14

- Pawlicki, M., Marchewka, A., Choraś, M., & Kozik, R. (2020). Gated Recurrent Units for Intrusion Detection. W M. Choraś & R. S. Choraś (red.), Image Processing and Communications. Techniques, Algorithms and Applications (ss. 142–148). Springer. https://doi.org/10.1007/978-3-030-31254-1_18

- Pawlicka, A., Jaroszewska-Choraś, D., Choraś, M., & Pawlicki, M. (2020). Guidelines for Stego/Malware Detection Tools: Achieving GDPR Compliance. IEEE Technology and Society Magazine, 60–70. https://doi.org/10.1109/MTS.2020.3031848

- Pawlicka, A., Pawlicki, M., Tomaszewska, R., Choraś, M., & Gerlach, R. (2020). Innovative Machine Learning Approach and Evaluation Campaign for Predicting the Subjective Feeling of Work-Life Balance among Employees. PloS One, 1–18. https://doi.org/10.1371/journal.pone.0232771

- Choraś, M., Pawlicki, M., Puchalski, D., & Kozik, R. (2020). Machine Learning – The Results Are Not the Only Thing that Matters! What About Security, Explainability and Fairness? W V. V. Krzhizhanovskaya I in. (red.), ICCS 2020: 20th International Conference on Computational Science (ss. 615–628). Springer. https://doi.org/10.1007/978-3-030-50423-6_46

- Pawlicki, M., Choraś, M., Kozik, R., & Hołubowicz, W. (2020). On the Impact of Network Data Balancing in Cybersecurity Applications. W V. V. Krzhizhanovskaya I in. (red.), ICCS 2020: 20th International Conference on Computational Science (ss. 96–210). Springer. https://doi.org/10.1007/978-3-030-50423-6_15

- Komisarek, M., Choraś, M., Kozik, R., & Pawlicki, M. (2020). Real-time stream processing tool for detecting suspicious network patterns using machine learning. W ARES '20: Proceedings of the 15th International Conference on Availability, Reliability and Security (ss. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3407023.3409189

- Pawlicka, A., Pawlicki, M., & Choraś, M. (2020). Zawód roku 2020. Jak go zdobyć? W R. Tomaszewska (red.), Praca i kariera w warunkach nowego millennium. Implikacje dla edukacji (ss. 145–154). Wydawnictwo Uniwersytetu Kazimierza Wielkiego.

- Pawlicki, M., Kozik, R., & Choraś, M. (2019). Artificial Neural Network Hyperparameter Optimisation for Network Intrusion Detection. W D.-S. Huang, V. Bevilacqua & P. Premaratne (red.), Intelligent Computing Theories and Application (ss. 749–760). Springer. https://doi.org/10.1007/978-3-030-26763-6_72

- Kozik, R., Pawlicki, M., Choraś, M., & Pedrycz, W. (2019). Practical Employment of Granular Computing to Complex Application Layer Cyberattack Detection. Complexity, 1–9. https://doi.org/10.1155/2019/5826737

- Pawlicki, M., Choraś, M., & Kozik, R. (2019). Recent Granular Computing Implementations and Its Feasibility in Cybersecurity Domain. W Proceedings of the 13th International Conference on Availability, Reliability and Security ARES 2018 (ss. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3230833.3233259

- Choraś, M., Pawlicki, M., & Kozik, R. (2019). Recognizing Faults in Software Related Difficult Data. W J. M. F. Rodrigues I in. (red.), International Conference on Computational Science (ss. 263–272). Springer. https://doi.org/10.1007/978-3-030-22744-9_20

- Choraś, M., Pawlicki, M., Kozik, R., Demestichas, K., Kosmides, P., & Gupta, M. (2019). SocialTruth Project Approach to Online Disinformation (Fake News) Detection and Mitigation. W Proceedings of the 14th International Conference on Availability, Reliability and Security (ss. 1–10). Association for Computing Machinery. https://doi.org/10.1145/3339252.3341497

- Choraś, M., Kozik, R., Pawlicki, M., Hołubowicz, W., & Franch, X. (2019). Software Development Metrics Prediction Using Time Series Methods. W K. Saeed, R. Chaki & V. Janev (red.), Computer Information Systems and Industrial Management (ss. 311–323). Springer. https://doi.org/10.1007/978-3-030-28957-7_26

- Kozik, R., Pawlicki, M., & Choraś, M. (2019). Sparse Autoencoders for Unsupervised Netflow Data Classification. W M. Choraś & R. S. Choraś (red.), Image Processing and Communications Challenges 10 (ss. 192–199). Springer. https://doi.org/10.1007/978-3-030-03658-4_23

- Choraś, M., Pawlicki, M., & Kozik, R. (2019). The Feasibility of Deep Learning Use for Adversarial Model Extraction in the Cybersecurity Domain. W H. Yin, D. Camacho, P. Tino, A. J. Tallon-Ballesteros, R. Menezes & R. Allmendinger (red.), International Conference on Intelligent Data Engineering and Automated Learning (ss. 353–360). Springer. https://doi.org/10.1007/978-3-030-33617-2_36

- Kozik, R., Choraś, M., Pawlicki, M., Hołubowicz, W., Pallmer, D., Mueller, W., Behmer, E.-J., Loumiotis, I., Demestichas, K., Horincar, R., Laudy, C., & Faure, D. (2019). The Identification and Creation of Ontologies for the Use in Law Enforcement AI Solutions – MAGNETO Platform Use Case. W N. T. Nguyen, R. Chbeir, E. Exposito, P. Aniorte & B. Trawiński (red.), Computational Collective Intelligence ICIC (ss. 335–345). Springer. https://doi.org/10.1007/978-3-030-28374-2_29

- Kozik, R., Pawlicki, M., & Choraś, M. (2018). Cost-Sensitive Distributed Machine Learning for NetFlow-Based Botnet Activity Detection. Security and Communication Networks, 1–8. https://doi.org/10.1155/2018/8753870

PODPIS ZAUFANY
MAREK
PAWLICKI
01.01.2025 13:03:23 [GMT+1]
Dokument podpisany elektronicznie podpisem zaufanym

Marek Pawlicki

……

(podpis wnioskodawcy)

Wykaz osiągnięć naukowych albo artystycznych, stanowiących znaczny wkład w rozwój określonej dyscypliny

Informacje zawarte w poszczególnych punktach tego dokumentu powinny uwzględniać podział na okres przed uzyskaniem stopnia doktora oraz pomiędzy uzyskaniem stopnia doktora a uzyskaniem stopnia doktora habilitowanego.

I. INFORMACJA O OSIĄGNIĘCIACH NAUKOWYCH ALBO ARTYSTYCZNYCH, O KTÓRYCH MOWA W ART. 219 UST. 1. PKT 2 USTAWY

1. Monografia naukowa, zgodnie z art. 219 ust. 1. pkt 2a Ustawy; lub

-

2. Cykl powiązanych tematycznie artykułów naukowych, zgodnie z art. 219 ust. 1. pkt 2b Ustawy; pt.:

Zaawansowane metody uczenia maszynowego oraz wyjaśnialnej sztucznej inteligencji (xAI) dla wykrywania ataków sieciowych.

Poniżej przedstawiono, w ramach jednotematycznego cyklu naukowego, serię publikacji. Publikacje ułożone są w kolejności chronologicznej, obejmują listę dziewięciu artykułów wydanych w latach 2020–2024. Wszystkie prezentowane wartości bibliometryczne odzwierciedlają stan na dzień 31. grudnia 2024 r., zgodnie z bazami danych naukowych:

- Web of Science (WoS)
    - https://www.webofscience.com/wos/author/record/ABC-3938-2021
- Scopus (Sco)
    - https://www.scopus.com/authid/detail.uri?authorId=57204352435
- Google Scholar (GSc)

- https://scholar.google.com/citations?user=tsqBC7QAAAAJ&hl=en

	DOI	**Publikacja**	**Punkty**	**IF/CORE**
[C01]	https://doi.org/10.1016/j.neucom.2020.07.138	Choraś, M., and **Pawlicki, M**. (2021) Intrusion detection approach based on optimised artificial neural network. In Neurocomputing, 452, (pp. 705–715). **CRediT:** Konceptualizacja, Opracowanie danych, Analiza formalna, Badania, Metodologia, Oprogramowanie, Walidacja, Pisanie - oryginalny szkic, Pisanie - recenzja i edycja Cytowania WoS: 54; Sco: 73; Gsc: 112;	140	IF 5.719
[C02]	https://doi.org/10.1016/j.neucom.2022.06.002	**Pawlicki, M**., Kozik, R. and Choraś, M. (2022) A survey on neural networks for (cyber-) security and (cyber-) security of neural networks. In Neurocomputing, 500, (pp. 1075–1087). **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja Cytowania WoS: 31; Sco: 44; Gsc: 60;	140	IF 5.779
[C03]	https://doi.org/10.1007/978-3-031-42823-4_21	**Pawlicki, M**., Pawlicka, A., Kozik, R., and Choraś, M. (2023) How to Boost Machine Learning Network Intrusion Detection Performance with Encoding Schemes. In International Conference on Computer Information Systems and Industrial Management (pp. 283-297). Cham: Springer Nature Switzerland.	40	CORE C

		CRediT: konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja, nadzór, zarządzanie projektem, pozyskiwanie funduszy Cytowania WoS: 0; Sco: 0; Gsc: 0;		
[C04]	https://doi.org/10.1145/3538969.3544428	**Pawlicki, M**., Kozik, R., and Choraś, M. (2022) Towards Deployment Shift Inhibition Through Transfer Learning in Network Intrusion Detection. In Proceedings of the 17th International Conference on Availability, Reliability and Security (pp. 1-6). **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja, nadzór, zarządzanie projektem, pozyskiwanie funduszy Cytowania WoS: 0; Sco: 1; Gsc: 2;	70	CORE B
[C05]	https://doi.org/10.1109/IJCNN54540.2023.10191496	**Pawlicki, M.,** Kozik, R., and Choraś, M. (2023), Improving Siamese Neural Networks with Border Extraction Sampling for the use in Real-Time Network Intrusion Detection. In International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, (pp. 1-8) **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja, nadzór, zarządzanie projektem, pozyskiwanie funduszy Cytowania WoS: 2; Sco: 2; Gsc: 3;	70	CORE B

[C06]	https://doi.org/10.1109/DSAA60987.2023.10302535	**Pawlicki**, **M.,** (2023) Towards Quality Measures for xAI algorithms: Explanation Stability. in IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA), (pp. 1–10) Cytowania WoS: 0; Sco: 0; Gsc: 0;	140	CORE B
[C07]	https://dl.acm.org/doi/10.1145/3665451.3665527	**Pawlicki**, **M.,** Pawlicka, A., Kozik, R. and Choraś, M., (2024) Explainability versus Security: The Unintended Consequences of xAI in Cybersecurity, in Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems, (pp. 1–7) **CRediT:** konceptualizacja, metodologia, oprogramowanie, walidacja, analiza formalna, badania, zasoby, kuratela danych, pisanie - oryginalny szkic, pisanie - recenzja i edycja, wizualizacja, nadzór, zarządzanie projektem, pozyskiwanie funduszy Cytowania WoS: 0; Sco: 0; Gsc: 2;	140	CORE A
[C08]	https://doi.org/10.1109/ACCESS.2024.3454084	**Pawlicki**, **M.,** (2024) ARIA, HaRIA, and GeRIA: Novel Metrics for Pre-Model Interpretability, in IEEE Access, vol. 12, (pp. 123561-123580) Cytowania WoS: 0; Sco: 0; Gsc: 0;	100	IF 3.4
[C09]	https://doi.org/10.1007/s10462-024-10972-3	**Pawlicki, M.**, Pawlicka, A., Kozik, R., and Choraś, M., (2024) The survey on the dual nature of xAI challenges in intrusion detection and their potential for AI innovation. In Artificial Intelligence Review, 57, (art. no. 330) **CRediT:** konceptualizacja, metodologia, analiza formalna, badania, pisanie - oryginalny szkic	140	IF 10.7

		Cytowania WoS: 0; Sco: 0; Gsc: 0;		

3. Wykaz zrealizowanych oryginalnych osiągnięć projektowych, konstrukcyjnych, technologicznych lub artystycznych, zgodnie z art. 219 ust. 1. pkt 2c Ustawy.

-

II. INFORMACJA O AKTYWNOŚCI NAUKOWEJ ALBO ARTYSTYCZNEJ

1. Wykaz opublikowanych monografii naukowych (z zaznaczeniem pozycji niewymienionych w pkt I.1).

-

2. Wykaz opublikowanych rozdziałów w monografiach naukowych.

-

3. Informacja o członkostwie w redakcjach naukowych monografii.

-

4. Wykaz opublikowanych artykułów w czasopismach naukowych (z zaznaczeniem pozycji niewymienionych w pkt I.2).

4.1. Artykuły w czasopismach naukowych opublikowane po uzyskaniu doktoratu, zgłoszone w punkcie I.2:

1. **[C09] [IF = 10.7 PKT: 140]**

   Pawlicki, M., Pawlicka, A., Kozik, R., and Choraś, M. (2024) The survey on the dual nature of xAI challenges in intrusion detection and their potential for AI innovation. In Artificial Intelligence Review, (pp. 1–32). https://doi.org/10.1007/s10462-024-10972-3

2. **[C08] [IF = 3.4 PKT: 100]**

   Pawlicki, M. (2024) ARIA, HaRIA and GeRIA: novel metrics for pre-model interpretability. In IEEE Access, (pp. 1–21). https://doi.org/10.1109/ACCESS.2024.3454084

3. **[C02] [IF = 5.779 PKT: 140]**

Pawlicki, M., Kozik, R., and Choraś, M. (2022) A survey on neural networks for (cyber-) security and (cyber-) security of neural networks. In Neurocomputing, (pp. 1075–1087). https://doi.org/10.1016/j.neucom.2022.06.002

4. **[C01] [IF = 5.719 PKT: 140]**

   Choraś, M., and Pawlicki, M. (2021) Intrusion detection approach based on optimised artificial neural network. In Neurocomputing, (pp. 705–715). https://doi.org/10.1016/j.neucom.2020.07.138

4.2. Artykuły w czasopismach naukowych opublikowane po uzyskaniu doktoratu, niezgłoszone w punkcie I.2:

1. **[IF = 4.5 PKT: 100]**

   Kozik, R., Pawlicka, A., Pawlicki, M., Choraś, M., Mazurczyk, W., and Cabaj, K. (2024) A meta-analysis of state-of-the-art automated fake news detection methods. In IEEE Transactions on Computational Social Systems, (pp. 5219–5229). https://doi.org/10.1109/TCSS.2023.3296627

2. **[IF = 5.5 PKT: 140]**

   Pawlicki, M., Pawlicka, A., Kozik, R., and Choraś, M. (2024) Advanced insights through systematic analysis: mapping future research directions and opportunities for xAI in deep learning and artificial intelligence used in cybersecurity. In Neurocomputing, (pp. 1–13). https://doi.org/10.1016/j.neucom.2024.127759

3. **[IF = 5.91 PKT: 200]**

   Pawlicka, A., Pawlicki, M., Kozik, R., Andrychowicz-Trojanowska, A., and Choraś, M. (2024) AI vs linguistic-based human judgement: bridging the gap in pursuit of truth for fake news detection. In Information Sciences, (pp. 1–29). https://doi.org/10.1016/j.ins.2024.121097

4. **[IF = 5.5 PKT: 140]**

   Pawlicki, M., Pawlicka, A., Uccello, F., Szelest, S., D'Antonio, S., Kozik, R., and Choraś, M. (2024) Evaluating the necessity of the multiple metrics for assessing explainable AI: a critical examination. In Neurocomputing, (pp. 1–19). https://doi.org/10.1016/j.neucom.2024.128282

5. **[IF = 50.5 PKT: 200]**

Choraś, M., Pawlicki, M., Kozik, R., and Pawlicka, A. (2024) Tech developers must respect equitable AI access. In Nature, (p. 1). https://doi.org/10.1038/d41586-024-00185-7

6. **[IF = 5.5 PKT: 140]**

   Kozik, R., Kątek, G., Gackowska, M., Kula, S., Komorniczak, J., Ksieniewicz, P., Pawlicka, A., Pawlicki, M., and Choraś, M. (2024) Towards explainable fake news detection and automated content credibility assessment: Polish internet and digital media use-case. In Neurocomputing, 608

   https://doi.org/10.1016/j.neucom.2024.128450

7. **[IF = 4.8 PKT: 140]**

   Kozik, R., Ficco, M., Pawlicka, A., Pawlicki, M., Palmieri, F., and Choraś, M. (2024) When explainability turns into a threat – using xAI to fool a fake news detection method. In Computers & Security, (pp. 1–28) https://doi.org/10.1016/j.cose.2023.103599

8. **[IF = 0.6 PKT: 20]**

   Szczepański, M., Pawlicki, M., Kozik, R., and Choraś, M. (2023) The application of deep learning imputation and other advanced methods for handling missing values in network intrusion detection. In Vietnam Journal of Computer Science, (pp. 1–23). https://doi.org/10.1142/S2196888822500257

9. **[IF = 3.4 PKT: 100]**

   Kozik, R., Mazurczyk, W., Cabaj, K., Pawlicka, A., Pawlicki, M., and Choraś, M. (2023) Deep learning for combating misinformation in multicategorical text contents. In Sensors, (pp. 1–13). https://doi.org/10.3390/s23249666

10. **[IF =0.0 PKT: 20]**

    Pawlicka, A., Choraś, M., Kozik, R., and Pawlicki, M. (2023) First broad and systematic horizon scanning campaign and study to detect societal and ethical dilemmas and emerging issues spanning over cybersecurity solutions. In Personal and Ubiquitous Computing, (pp. 193–202). https://doi.org/10.1007/s00779-020-01510-3

11. **[IF = 5.5 PKT: 140]**

    Pawlicki, M., Pawlicka, A., Kozik, R., and Choraś, M. (2023) The survey and meta-analysis of the attacks, transgressions, countermeasures and security aspects common to the Cloud, Edge and IoT. In Neurocomputing, (pp. 1–18). https://doi.org/10.1016/j.neucom.2023.126533

12. **[IF = 3.4 PKT: 100]**

    Pawlicka, A., Pawlicki, M., Kozik, R., and Choraś, M. (2023) What will the future of cybersecurity bring us, and will it be ethical? The hunt for the black swans of cybersecurity ethics. In IEEE Access, (pp. 58796–58807). https://doi.org/10.1109/ACCESS.2023.3283791

13. **[IF = 0.0 PKT: 20]**

    Pawlicka, A., Tomaszewska, R., Krause, E., Jaroszewska-Choraś, D., Pawlicki, M., and Choraś, M. (2023) Has the pandemic made us more digitally literate? innovative association rule mining study of the relationships between shifts in digital skills and cybersecurity awareness occurring whilst working remotely during the COVID-19 pandemic. In Journal of Ambient Intelligence and Humanized Computing, (pp. 14721–14731). https://doi.org/10.1007/s12652-022-04371-1

14. **[IF = 1.0 PKT: 100]**

    Kozik, R., Pawlicki, M., Kula, S., and Choraś, M. (2022) Fake news detection platform – conceptual architecture and prototype. In Logic Journal of the IGPL, (pp. 1005–1016). https://doi.org/10.1093/jigpal/jzac009

15. **[IF = 2.4 PKT: 70]**

    Pawlicka, A., Pawlicki, M., Kozik, R., and Choraś, M. (2022) Human-driven and human-centred cybersecurity: policy-making implications. In Transforming Government: People, Process and Policy, (pp. 478–487). https://doi.org/10.1108/TG-05-2022-0073

16. **[IF = 0.0 PKT: 70]**

    Komisarek, M., Pawlicki, M., Kowalski, M., Marzecki, A., Kozik, R., and Choraś, M. (2022) Hunting cyberattacks: experience from the real backbone network. In Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, (pp. 128–146). https://doi.org/10.22667/JOWUA.2022.06.30.128

17. **[IF = 2.622 PKT: 70]**

    Pawlicka, A., Choraś, M., and Pawlicki, M. (2022) Opening up a discussion on the native speaker myth in science. In Accountability in Research, (pp. 477–481). https://doi.org/10.1080/08989621.2021.1960514

18. **[IF = 3.4 PKT: 100] (Editorial)**

Choraś, M., Kozik, R., and Pawlicki, M. (2022) Sensors and pattern recognition methods for security and industrial applications. In Sensors, (pp. 1–3). https://doi.org/10.3390/s22165968

19. **[IF = 2.7 PKT: 100]**

    Komisarek, M., Kozik, R., Pawlicki, M., and Choraś, M. (2022) Towards zero-shot flow-based cyber-security anomaly detection framework. In Applied Sciences, (pp. 1–12). https://doi.org/10.3390/app12199636

20. **[IF = 1.0 PKT: 100]**

    Dutta, V., Pawlicki, M., Kozik, R., and Choraś, M. (2022) Unsupervised network traffic anomaly detection with deep autoencoders. In Logic Journal of the IGPL, (pp. 912–925). https://doi.org/10.1093/jigpal/jzac002

21. **[IF = 10.562 PKT: 100]**

    Pawlicka, A., Choraś, M., Pawlicki, M., and Kozik, R. (2021) A $10 million question and other cybersecurity-related ethical dilemmas amid the COVID-19 pandemic. In Business Horizons, (pp. 729–734). https://doi.org/10.1016/j.bushor.2021.07.010

22. **[IF = 2.307 PKT: 70]**

    Kozik, R., Pawlicki, M., and Choraś, M. (2021) A new method of hybrid time window embedding with transformer-based traffic data classification in IoT-networked environment. In Pattern Analysis and Applications, (pp. 1441–1449). https://doi.org/10.1007/s10044-021-00980-2

23. **[IF = 3.847 PKT: 100]**

    Pawlicka, A., Pawlicki, M., Kozik, R., and Choraś, R. (2021) A systematic review of recommender systems and their applications in cybersecurity. In Sensors, (art. no. 5248), (pp. 1–25). https://doi.org/10.3390/s21155248

24. **[IF = 2.738 PKT: 100]**

    Komisarek, M., Pawlicki, M., Kozik, R., Hołubowicz, W., and Choraś, M. (2021) How to effectively collect and process network data for intrusion detection? In Entropy, (pp. 1–25). https://doi.org/10.3390/e23111532

25. **[IF = 0.0 PKT: 70]**

    Komisarek, M., Pawlicki, M., Kozik, R., and Choraś, M. (2021) Machine learning based approach to anomaly and cyberattack detection in streamed network traffic data. In

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, (pp. 3–19). https://doi.org/10.22667/JOWUA.2021.03.31.003

26. **[IF = 4.997 PKT: 140]**

Szczepański, M., Pawlicki, M., Kozik, R., and Choraś, M. (2021) New explainability method for BERT-based model in fake news detection. In Scientific Reports, (pp. 1–13). https://doi.org/10.1038/s41598-021-03100-6

27. **[IF = 2.738 PKT: 100]**

Pawlicki, M., and Choraś, R. (2021) Preprocessing pipelines including block-matching convolutional neural network for image denoising to robustify deep reidentification against evasion attacks. In Entropy, (pp. 1–14). https://doi.org/10.3390/e23101304

28. **[IF = 3.847 PKT: 100]**

Mihailescu, M.-E., Mihai, D., Carabas, M., Komisarek, M., Pawlicki, M., Hołubowicz, W., and Kozik, R. (2021) The proposition and evaluation of the RoEduNet-SIMARGL2021 network intrusion detection dataset. In Sensors, (pp. 1–20). https://doi.org/10.3390/s21134319

29. **[IF = 3.006 PKT: 100]**

Pawlicka, A., Choraś, M., and Pawlicki, M. (2021) The stray sheep of cyberspace a.k.a. the actors who claim they break the law for the greater good. In Personal and Ubiquitous Computing, (pp. 843–852). https://doi.org/10.1007/s00779-021-01568-7

30. **[IF = 3.476 PKT: 100]**

Caviglione, L., Choraś, M., Corona, I., Janicki, A., Mazurczyk, W., Pawlicki, M., and Wasielewska, K. (2021) Tight arms race: overview of current malware threats and trends in their detection. In IEEE Access, (pp. 5371–5396). https://doi.org/10.1109/ACCESS.2020.3048319

31. **[IF = 7.187 PKT: 140]**

Pawlicki, M., Choraś, M., and Kozik, R. (2020) Defending network intrusion detection systems against adversarial evasion attacks. In Future Generation Computer Systems, (pp. 148–154). https://doi.org/10.1016/j.future.2020.04.013

32. **[IF = 3.576 PKT: 100]**

Dutta, V., Choraś, M., Pawlicki, M., and Kozik, R. (2020) A deep learning ensemble for network anomaly and cyber-attack detection. In Sensors, (pp. 1–20). https://doi.org/10.3390/s20164583

33. **[IF = 1.554 PKT: 40]**

Pawlicka, A., Jaroszewska-Choraś, D., Choraś, M., and Pawlicki, M. (2020) Guidelines for stego/malware detection tools: achieving GDPR compliance. In IEEE Technology and Society Magazine, (pp. 60–70). https://doi.org/10.1109/MTS.2020.3031848

34. **[IF = 1.139 PKT: 40]**

Dutta, V., Choraś, M., Pawlicki, M., and Kozik, R. (2020) Detection of cyberattacks traces in IoT data. In Journal of Universal Computer Science, (pp. 1422–1434). https://doi.org/10.3897/jucs.2020.075

4.3. Artykuły opublikowane przed uzyskaniem doktoratu:

1. **[IF = 0.0 PKT: 5]**

Kozik, R., Choraś, M., Pawlicki, M., Hołubowicz, W., Pallmer, D., Mueller, W., Behmer, E.-J., Loumiotis, I., Demestichas, K., Horincar, R., Laudy, C., and Faure, D. (2020) Common representational model and ontologies for effective law enforcement solutions. In Vietnam Journal of Computer Science, (pp. 1–18). https://doi.org/10.1142/S2196888820020017

2. **[IF = 3.240 PKT: 100]**

Pawlicka, A., Pawlicki, M., Tomaszewska, R., Choraś, M., and Gerlach, R. (2020) Innovative machine learning approach and evaluation campaign for predicting the subjective feeling of work-life balance among employees. In PLoS One, (pp. 1–18). https://doi.org/10.1371/journal.pone.0232771

3. **[IF = 2.462 PKT: 70]**

Kozik, R., Pawlicki, M., Choraś, M. R., and Pedrycz, W. (2019) Practical employment of granular computing to complex application layer cyberattack detection. In Complexity, (pp. 1–9). https://doi.org/10.1155/2019/5826737

4. **[IF = 1.376 PKT: 20]**

Kozik, R., Pawlicki, M., and Choraś, M. R. (2018) Cost-sensitive distributed machine learning for NetFlow-based botnet activity detection. In Security and Communication Networks, (pp. 1–8). https://doi.org/10.1155/2018/8753870

5. Wykaz osiągnięć projektowych, konstrukcyjnych, technologicznych (z zaznaczeniem pozycji niewymienionych w pkt I.3).

Prace w ramach projektu H2020 InfraStress, prowadzone przez zespół, w którym byłem jednym z kluczowych członków, doprowadziły do powstania Komponentu ST.CD.2. Komponent odpowiedzialny za wykrywanie ataków sieciowych przy pomocy uczenia maszynowego. Narzędzie zostało skutecznie wdrożone w firmie z grupy Johnson & Johnson, DePuy Synthes z Cork w Irlandii, zajmującej się wytwarzaniem sprzętu medycznego, zatrudniającej ponad tysiąc ludzi. Projekt Infrastress zakończył się jesienią 2021 roku; jego całokształt otrzymał aprobatę i pozytywną ocenę recenzentów z Komisji Europejskiej, a jego praktyczne wyniki i wytworzone rozwiązania będą wdrażane dla zapewnienia wyższego bezpieczeństwa obywateli Polski i Unii Europejskiej.

Zastosowanie Praktyczne Wyników Badań Naukowych

Praktyczne zastosowanie w przemysłowej infrastrukturze krytycznej: wdrożenie komponentu ST.CD.2 w DePuy Synthes (Johnson & Johnson) w Irlandii,

Link do wideo prezentującego działanie rozwiązania:

https://www.youtube.com/watch?v=KZFgrqfz5ek

Wykorzystanie komponentu w środowisku przemysłowym infrastruktury krytycznej, pokazuje bezpośrednie praktyczne zastosowanie badań.

Publikacja i Dystrybucja Wiedzy: Wyniki pilotażu i współpracy z partnerami z Collins Aerospace (firmy obsługującej cyberbezpieczeństwo DePuy Synthes) opublikowane zostały w recenzowanej pracy naukowej: Komisarek, M., Pawlicki, M., Soboński, P., Pawlicka, A., Kozik, R., & Choraś, M. (2022). Extending machine learning-based intrusion detection with the imputation method. In Progress in Image Processing, Pattern Recognition and

Communication Systems: Proceedings of the Conference (CORES, IP&C, ACS)-June 28-30 2021 12 (pp. 284-292). Springer International Publishing."

Innowacyjność i Znaczenie Wyników Badań

Projekt H2020 InfraStress przyczynił się do rozwoju nowych metod wykrywania cyberataków, co ma istotny wpływ na pole bezpieczeństwa sieciowego i ochrony danych.

Wpływ na Bezpieczeństwo Danych: ST.CD.2, oferuje nowe perspektywy w dziedzinie bezpieczeństwa sieciowego, poprzez wykorzystanie algorytmów uczenia maszynowego do wykrywania ataków sieciowych w czasie rzeczywistym, co jest istotne dla ochrony infrastruktury krytycznej.

6. Wykaz publicznych realizacji dzieł artystycznych (z zaznaczeniem pozycji niewymienionych w pkt I.3).

-

7. Informacja o wystąpieniach na krajowych lub międzynarodowych konferencjach naukowych lub artystycznych, z wyszczególnieniem przedstawionych wykładów na zaproszenie i wykładów plenarnych.

7.1. Aktywny udział (prezentacja pracy) podczas międzynarodowych konferencji naukowych po uzyskaniu stopnia doktora:

1. **Konferencja:** 24th IEEE International Conference on Data Mining (ICDM) 2024 - Workshops

   Miejsce: Abu Zabi, Zjednoczone Emiraty Arabskie

   Referaty:

   - Pawlicki Marek; Pawlicka Aleksandra; Choraś Ryszard; Kozik Rafał; Choraś Michał. “Semi-Supervised Learning Trumps Active Learning in Cold-Starting Netflow-Based AI Network Intrusion Detection Systems on Scarce and Difficult Data”. W: Proceedings 24th IEEE International Conference on Data Mining Workshops. IEEE, 2024, s. 221-228. doi: 10.1109/ICDMW65004.2024.00035

- Karampasi Aikaterini; Radoglou Panagiotis; Pawlicki Marek; Choraś Ryszard; Martin Ramon; Sarigiannidis Panagiotis; Puchalski Damian; Pawlicka Aleksandra; Kozik Rafał; Choraś Michał. "Towards Transparent AI-Powered Cybersecurity in Financial Systems: The Deployment of Federated Learning and Explainable AI in the CaixaBank pilot". W: Proceedings 24th IEEE International Conference on Data Mining Workshops. IEEE, 2024, s. 270-277. doi: 10.1109/ICDMW65004.2024.00041

2. **Konferencja:** 29th European Symposium on Research in Computer Security (ESORICS) 2024 - workshop "Computational methods for emerging problems in disinformation analysis" (DisA)

   Miejsce: Bydgoszcz, Polska

   Referat: Marek Pawlicki, Federica Uccello, Salvatore D'Antonio, Rafał Kozik and Michał Choraś "A Novel Method of Improving Intrusion Detection Systems Robustness Against Adversarial Attacks, through Feature Omission and a Committee of Classifiers"

3. **Konferencja:** 19th International Conference on Availability, Reliability and Security (ARES) 2024

   Miejsce: Wiedeń, Austria

   Referaty:

   - Komisarek Mikołaj; Pawlicki Marek; D'Antonio Salvatore; Kozik Rafał; Pawlicka Aleksandra; Choraś Michał. "Enhancing Network Security Through Granular Computing: A Clustering-by-Time Approach to NetFlow Traffic Analysis". W: Proceedings of the 19th International Conference on Availability, Reliability and Security. New York: ACM, 2024, s. 1-8.

   - Pawlicki Marek; Puchalski D.; Szelest Sebastian; Pawlicka Aleksandra; Kozik Rafał; Choraś Michał. "Introducing a Multi-Perspective xAI Tool for Better Model Explainability". W: Proceedings of the 19th International Conference on Availability, Reliability and Security. New York: ACM, 2024, s. 1-8.

4. **Konferencja:** The 7th International Conference on Mobile Internet Security (MobiSec'23) 2023

Miejsce: Okinawa, Japonia

Referat: Pawlicki Marek; Kozik Rafał; Choraś Michał. "The Impact of Data Scaling Approaches on Deep Learning, Random Forest and Nearest Neighbour-Based Network Intrusion Detection Systems for DoS Detection in IoT Networks". W: Mobile Internet Security; eds. Ilsun You, Michał Choraś (et al.). Singapore: Springer, 2024, s. 197-208.

5. **Konferencja**: 2023 IEEE International Conference on Cyber Security and Resilience (CSR)

   Miejsce: Wenecja, Włochy

   Referat: Pawlicka Aleksandra; Puchalski Damian; Pawlicki Marek; Kozik Rafał; Choraś Michał. "How to secure the IoT-based surveillance systems in an ELEGANT way". W: 2023 IEEE International Conference on Cyber Security and Resilience (CSR). IEEE, 2023, s. 636-640.

6. **Konferencja**: 10th International Conference on Data Science and Advanced Analytics (DSAA) 2023

   Miejsce: Thessaloniki, Grecja

   Referat: Pawlicki Marek. "Towards Quality Measures for xAI algorithms: Explanation Stability". W: 10th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2023, s. 1-10. doi: 10.1109/DSAA60987.2023.10302535

7. **Konferencja:** International Joint Conference on Neural Networks (IJCNN) 2023

   Miejsce: Gold Coast, Australia

   Referat: Pawlicki Marek; Kozik Rafał; Choraś Michał. "Improving Siamese Neural Networks with Border Extraction Sampling for the use in Real-Time Network Intrusion Detection". W: International Joint Conference on Neural Networks (IJCNN) Proceedings. IEEE, 2023, s. 1-8.

8. **Konferencja**: Annual Americas Conference on Information Systems (AMCIS) 2023

Miejsce: Panama, Panama

Referat: Choraś Michał; Pawlicka Aleksandra; Pawlicki Marek; Kozik Rafał. "How to Navigate the Uncharted Waters of Cyberwar". W: AMCIS 2023 Proceedings. 1. USA: Association for Information Systems, 2023.

9. **Konferencja:** 19th International Conference on Intelligent Computing (ICIC) 2023

   Miejsce: Zhengzhou, Chiny

   Referat: Kurek Wiktor; Pawlicki Marek; Pawlicka Aleksandra; Kozik Rafał; Choraś Michał. "Explainable Artificial Intelligence 101: Techniques, Applications and Challenges". W: Advanced Intelligent Computing Technology and Applications; De-Shuang Huang et al. Singapore: Springer, 2023, s. 310-318.

10. **Konferencja:** 13th International Conference on Image Processing and Communications (IP&C) 2023

    Miejsce: konferencja online

    Referat: Pawlicki Marek; Pawlicka Aleksandra; Śrutek Mścisław; Kozik Rafał; Choraś Michał. "Interpreting Intrusions - The Role of Explainability in AI-Based Intrusion Detection Systems". W: Progress on Pattern Classification, Image Processing and Communications. CORES IP&C 2023. Springer, 2023, s. 45-53.

11. **Konferencja:** 22nd International Conference on Computer Information Systems and Industrial Management Applications (CISIM) 2023

    Miejsce: Tokio, Japonia

    Referat: Pawlicki Marek; Pawlicka Aleksandra; Kozik Rafał; Choraś Michał. "How to Boost Machine Learning Network Intrusion Detection Performance with Encoding Schemes". W: Computer Information Systems and Industrial Management: CISIM 2023; eds. Saeed, K. et al. Springer, 2023, s. 283-297. doi: 10.1007/978-3-031-42823-4_21

12. **Konferencja**: The 31st International Conference on Information Systems Development (ISD) 2023

    Miejsce: Lizbona, Portugalia

    Referat: Pawlicki Marek "Strengths And Weaknesses of Deep, Convolutional and Recurrent Neural Networks in Network Intrusion Detection Deployments". W: A. R. da Silva, M. M. da Silva, J. Estima, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), Information Systems Development, Organizational Aspects and Societal Trends (ISD2023 Proceedings). Lisbon, Portugal: Instituto Superior Técnico. ISBN: 978-989-33-5509-1. https://doi.org/10.62036/ISD.2023.54

13. **Konferencja:** The 21st International Conference on Artificial Intelligence and Soft Computing (ICAISC) 2022

    Miejsce: Zakopane, Polska

    Referat: Pawlicki Marek; Zadnik Martin; Kozik Rafał; Choraś Michał. "Analysis and Detection of DDoS Backscatter Using NetFlow Data, Hyperband-Optimised Deep Learning and Explainability Techniques". W: Artificial Intelligence and Soft Computing; eds. Leszek Rutkowski et al. Switzerland: Springer, 2023.

14. **Konferencja**: The 30th International Conference on Information Systems Development (ISD) 2022
    Miejsce: Cluj-Napoca, Rumunia
    Referat: Pawlicki Marek; Pawlicka Aleksandra; Kozik Rafał; Choraś Michał. "Graph Databases and E-commerce Cybersecurity - A Match Made in Heaven? The Innovative Technology to Enhance Cyberthreat Mitigation". W: *ISD2022 Proceedings*. Romania: Babe-Bolyai University, 2022.

15. **Konferencja:** 17th International Conference on Availability, Reliability and Security (ARES) 2022
    Miejsce: Wiedeń, Austria
    Referaty:

    - Komisarek Mikołaj; Pawlicki Marek; Mihailescu Maria-Elena; Mihai Darius; Carabas Mihai; Kozik Rafał; Choraś Michał. "A novel, refined dataset for real-time Network Intrusion Detection". W: *Proceedings of the 17th International Conference on Availability, Reliability and Security*. New York: ACM, 2022, s. 1-8. doi: 10.1145/3538969.3544486

- Pawlicki Marek; Kozik Rafał; Choraś Michał. “Towards Deployment Shift Inhibition Through Transfer Learning in Network Intrusion Detection”. W: *Proceedings of the 17th International Conference on Availability, Reliability and Security*. New York: ACM, 2022, s. 1-6. doi: 10.1145/3538969.3544428

16. **Konferencja:** 18th International Conference On Intelligent Computing (Intelligent Computing) 2022
    Miejsce: Xi'an, China
    Referat: Kozik Rafał; Pawlicki Marek; Szczepański Mateusz; Renk Rafał; Choraś Michał. “Efficient Post Event Analysis and Cyber Incident Response in IoT and E-commerce Through Innovative Graphs and Cyberthreat Intelligence Employment”. W: *Intelligent Computing Methodologies. ICIC 2022*. Switzerland: Springer, 2022, s. 257-266. doi: 10.1007/978-3-031-13832-4_22

17. **Konferencja:** International Conference on Image Processing and Communications (IP&C) 2021

    Miejsce: Konferencja Online

    Referat: Kuchczyński Marcin; Pawlicka Aleksandra; Pawlicki Marek; Choraś Michał. “Using Machine Learning to Detect the Signs of Radicalization and Hate Speech on Twitter”. W: Progress in Image Processing, Pattern Recognition and Communication Systems. Cham: Springer, 2022, s. 210-218.

18. **Konferencja:** 13th Asian Conference on Intelligent Information and Database Systems (ACIIDS) 2021

    Miejsce: Konferencja Online

    Referat: Pawlicki Marek; Choraś Michał; Kozik Rafał; Hołubowicz Witold. “Missing and Incomplete Data Handling in Cybersecurity Applications”. W: Intelligent Information and Database Systems. Switzerland: Springer, 2021, s. 413-427.

19. **Konferencja:** 17th International Conference On Intelligent Computing (ICIC) 2021
    Miejsce: Shenzhen, China
    Referat: Pawlicki Marek; Kozik Rafał; Puchalski Damian; Choraś Michał. “Towards AI-Based Reaction and Mitigation for e-Commerce - the ENSURESEC Engine”. W: *Intelligent Computing Theories and Application*. Springer, 2021, s. 24-31. doi:

10.1007/978-3-030-84532-2_3

7.2 Aktywny udział (prezentacja pracy) podczas międzynarodowych konferencji naukowych przed uzyskaniem stopnia doktora:

1. **Konferencja:** 20th International Conference on Computational Science (ICCS2020)

   Miejsce: Konferencja Online

   Referaty:

   - Choraś Michał; Pawlicki Marek; Puchalski Damian; Kozik Rafał. "Machine Learning - The Results Are Not the only Thing that Matters! What About Security, Explainability and Fairness?". W: ICCS 2020: 20th International Conference on Computational Science. Cham: Springer, 2020, s. 615-628. doi: 10.1007/978-3-030-50423-6_46
   - Pawlicki Marek; Choraś Michał; Kozik Rafał; Hołubowicz Witold. "On the Impact of Network Data Balancing in Cybersecurity Applications". W: ICCS 2020: 20th International Conference on Computational Science. Cham: Springer, 2020, s. 196-210. doi: 10.1007/978-3-030-50423-6_15

2. **Konferencja:** 15th International Conference on Availability, Reliability and Security (ARES) 2020

   Miejsce: Konferencja Online

   Referat: Komisarek Mikołaj; Choraś Michał; Kozik Rafał; Pawlicki Marek. "Real-time stream processing tool for detecting suspicious network patterns using machine learning". W: ARES '20: Proceedings of the 15th International Conference on Availability, Reliability and Security. New York: ACM, 2020, s. 1-7. doi: 10.1145/3407023.3409189

3. **Konferencja:** International Conference On Intelligent Computing (ICIC) 2019

   Miejsce: Nanchang, Chiny

Referat: Pawlicki Marek; Kozik Rafał; Choraś Michał Ryszard. “Artificial Neural Network Hyperparameter Optimisation for Network Intrusion Detection”. W: Intelligent Computing Theories and Application. Cham: Springer, 2019, s. 749-760.

4. **Konferencja:** 11th International Conference on Computational Collective Intelligence (ICCCI) 2019

   Miejsce: Hendaye, Francja

   Referat: Kozik Rafał; Choraś Michał; Pawlicki Marek; Hołubowicz Witold; Pallmer Dirk; Mueller Wilmuth; Behmer Ernst-Josef; Loumiotis Ioannis; Demestichas Konstantinos; Horincar Roxana; Laudy Claire; Faure David. “The Identification and Creation of Ontologies for the Use in Law Enforcement AI Solutions - MAGNETO Platform Use Case”. W: Computational Collective Intelligence ICIC. Cham: Springer, 2019, s. 335-345. doi: 10.1007/978-3-030-28374-2_29

5. **Konferencja:** 18th International Conference on Computer Information Systems and Industrial Management (CISIM) 2019

   Miejsce: Belgrad, Serbia

   Referat: Choraś Michał Ryszard; Pawlicki Marek; Hołubowicz Witold; Franch Xavier. “Software Development Metrics Prediction Using Time Series Methods”. W: Computer Information Systems and Industrial Management. Cham: Springer, 2019, s. 311-323. doi: 10.1007/978-3-030-28957-7_26

6. **Konferencja:** 14th International Conference on Availability, Reliability and Security (ARES) 2019

   Miejsce: Canterbury, Wielka Brytania

   Referat: Choraś Michał Ryszard; Pawlicki Marek; Kozik Rafał; Demestichas Konstantinos; Kosmides Pavlos; Gupta Manik. “SocialTruth Project Approach to Online Disinformation (Fake News) Detection and Mitigation”. W: Proceedings of the 14th International Conference on Availability, Reliability and Security. New York: ACM, 2019, s. 1-10. doi: 10.1145/3339252.3341497

7. **Konferencja:** International Conference on Computational Science (ICCS) 2019

   Miejsce: Faro, Algarve, Portugalia,

   Referat: Choraś Michał Ryszard; Pawlicki Marek; Kozik Rafał. "Recognizing Faults in Software Related Difficult Data". W: International Conference on Computational Science. Cham: Springer, 2019, s. 263-272. doi: 10.1007/978-3-030-22744-1_20

8. **Konferencja:** International Conference on Image Processing and Communications (IP&C) 2019

   Miejsce: Bydgoszcz, Polska

   Referat: Pawlicki Marek; Marchewka Adam; Choraś Michał; Kozik Rafał. "Gated Recurrent Units for Intrusion Detection". W: Image Processing and Communications. Techniques, Algorithms and Applications. Cham: Springer, 2020, s. 142-148. doi: 10.1007/978-3-030-31254-1_18

9. **Konferencja:** 11th International Conference on Computer Recognition Systems (CORES) 2019

   Miejsce: Polanica Zdrój, Polska

   Referat: Pawlicki Marek; Giełczyk Agata; Kozik Rafał; Choraś Michał Ryszard. "Fault-Prone Software Classes Recognition via Artificial Neural Network with Granular Dataset Balancing". W: Progress in Computer Recognition Systems. Cham: Springer, 2020, s. 130-140. doi: 10.1007/978-3-030-19738-4_14

7.3 Aktywny udział w konferencjach naukowych - wykłady na zaproszenie (keynote), panele

Konferencja: 19th International Conference on Availability, Reliability and Security (ARES) 2024 (CORE B) - Workshop: 7th International Workshop on Emerging Network Security (ENS)

Miejsce: Wiedeń, Austria

Keynote: "Enhancing Network Cybersecurity with Novel Trustworthy AI Solutions"

https://www.ares-conference.eu/ens

Konferencja: 29th European Symposium on Research in Computer Security (ESORICS) 2024 (CORE A) - Workshop: Computational methods for emerging problems in disinformation analysis (DiSA)

Miejsce: Bydgoszcz, Polska

Keynote: "AI Explainability Explained"

https://disa-iccs.github.io/

Konferencja: Research & Innovation Forum (RII FORUM) 2023

Miejsce: Kraków, Polska

Debata Plenarna: Byłem jednym z ekspertów w debacie plenarnej "Advances in ICT & the society: Threading the thin line between progress, development and mental health' w ramach Research & Innovation Forum https://rii-forum.org/wp-content/uploads/2023/04/Rii2023-Program-FULL-01.04.2023-v3.pdf

8. Informacja o udziale w komitetach organizacyjnych i naukowych konferencji krajowych lub międzynarodowych, z podaniem pełnionej funkcji.

8.1 Współorganizacja konferencji (Conference Chair) po uzyskaniu stopnia doktora:

- ESORICS 2024 - European Symposium on Research in Computer Security (Workshop Chair)

  https://esorics2024.org/organizing_team-en

- IP&C 2023, 2021 - International Conference on Image Processing & Communications - członek komitetu programowego, członek komitetu organizacyjnego (Program Committee, Organising Committee)

  https://ipc.pwr.edu.pl/

- ICCCI International Conference on Computational Collective Intelligence 2024

  - Special Session: Special Session on Security and reliability of information, networks and social media (SIRENE) 2024 - współorganizator sesji specjalnej

https://iccci.pwr.edu.pl/2024/session.php?acronym=SIRENE

- ARES2024, 2023, 2022 - International Conference on Availability, Reliability and Security
    - International Workshop on Cyber Use of Information Hiding (CUING) - członek komitetu programowego (Program Committee)
        - 2024: https://2024.ares-conference.eu/cuing
        - 2023: https://2023.ares-conference.eu/workshops/cuing-2023/index.html
        - 2022: https://2022.ares-conference.eu/workshops/cuing-2022/index.html
    - International Workshop on Emerging Network Security (ENS) - członek komitetu programowego (Program Committee)
        - 2024: https://2024.ares-conference.eu/ens
        - 2023: https://2023.ares-conference.eu/workshops-eu-symposium/ens-2023/index.html
        - 2022: https://2022.ares-conference.eu/workshops-eu-symposium/ens-2022/index.html
- ICIC 2022, 2021 International Conference On Intelligent Computing
    - 2022 - Special Session on Intelligent Computing for Cybersecurity: Detection, Mitigation and Response in the protection of Soft Targets and Smart Cities - współorganizator sesji specjalnej

      http://www.ic-icc.cn/2022/Special%20Session.htm
    - 2021: Special Session on Intelligent Computing for Cybersecurity: Detection, Mitigation and Response - współorganizator sesji specjalnej

      http://www.ic-icc.cn/2021/Special%20Session.htm#ss7
- EICC 2021 - European Interdisciplinary Cybersecurity Conference
    - Special Session on Advanced and Reliable Solutions to Counter Malware and Stegomalware - DETONATOR 2021 - członek komitetu programowego sesji specjalnej (Program committee)

https://www.fvv.um.si/eicc2021/detonator.html

8.2 Przewodniczyłem sesjom (Session Chair) po uzyskaniu stopnia doktora:

- ICIC2021 - International Conference on Intelligent Computing 2021
- ICIC2022 - International Conference on Intelligent Computing 2022
- ISD2022 - International Conference on Information Systems Development 2022
- IJCNN2023 - International Joint Conference on Neural Networks 2023
- DSAA2023 - International Conference on Data Science and Advanced Analytics 2023
- ESORICS2024 - European Symposium on Research in Computer Security (workshop DisA)

8.3 Przewodniczyłem sesjom (Session Chair) przed uzyskaniem stopnia doktora:

- ICIC2019 - International Conference on Intelligent Computing 2019

9. Informacja o uczestnictwie w pracach zespołów badawczych realizujących projekty finansowane w drodze konkursów krajowych lub zagranicznych, z podziałem na projekty zrealizowane i będące w toku realizacji, oraz z uwzględnieniem informacji o pełnionej funkcji w ramach prac zespołów.

9.1 Pełniłem/pełnię rolę Kierownika Projektu po stronie polskiego partnera (ITTI. sp. z o.o.), pełniąc w nich również rolę Eksperta AI, w następujących projektach (po uzyskaniu doktoratu):

1. Horizon Europe AI4CYBER od 2022-09-01 (trwa)
2. H2020 ELEGANT w okresie od 2021-01-01 do 2023-12-31
3. H2020 APPRAISE w okresie od 2021-09-01 do 2024-02-29
4. H2020 ENSURESEC w okresie od 2020-06-01 do 2022-05-31

Akronim	**HE AI4CYBER**	
Tytuł	Trustworthy Artificial Intelligence for Cybersecurity Reinforcement and System Resilience	
Identyfikator umowy o grant	GA 101070450	
Lata realizacji	2022-2025	
Typ/finansowanie	HORIZON.2.3 - Civil Security for Society HORIZON.2.3.3 - Cybersecurity	
Opis projektu	Projekt dotyczy zastosowań sztucznej inteligencji w cyberbezpieczeństwie	
Rola	● Główny wykonawca i kierownik projektu po stronie polskiego partnera ● Koordynator zadania odpowiedzialnego za stworzenie rozwiązania zapewniającego wyjaśnialność (xAI) komponentów AI użytych w projekcie	

Akronim	**H2020 ELEGANT**	
Tytuł	Secure and Seamless Edge-to-Cloud Analytics	
Identyfikator umowy o grant	GA 957286	
Lata realizacji	2021-2023	
Typ/finansowanie	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT)	

Opis projektu	Projekt dotyczył wyzwań stojących przed urządzeniami IoT i domeną Big Data
Rola	● Główny wykonawca i kierownik projektu po stronie polskiego partnera ● Koordynator pakietu prac odpowiedzialnego za stworzenie rozwiązania do analizy ruchu sieciowego w czasie rzeczywistym, dostosowanego do pracy z urządzeniami IoT i charakterystycznymi wymaganiami projektu

Akronim	**H2020 APPRAISE**	
Tytuł	fAcilitating Public & Private secuRity operAtors to mitigate terrorIsm Scenarios against soft targEts	
Identyfikator umowy o grant	GA 101021981	
Lata realizacji	2021-2024	
Typ/finansowanie	SU-FCT03-2018-2019-2020 - Information and data stream management to fight against (cyber)crime and terrorism Funded under H2020-EU.3.7. H2020-EU.3.7.1. H2020-EU.3.7.8.	
Opis projektu	Projekt dotyczy bezpieczeństwa cyber-fizycznego obiektów typu soft target	
Rola	● Główny wykonawca i kierownik projektu po stronie polskiego partnera ● Koordynacja prac mających na celu wytworzenie i wdrożenie komponentów w zakresie detekcji ataków cybernetycznych na obiekty typu soft target	

Akronim	**H2020 ENSURESEC**	

Tytuł	End-to-end Security of the Digital Single Market's E-commerce and Delivery Service Ecosystem	
Identyfikator umowy o grant	883242	
Lata realizacji	2020-2022	
Typ/finansowanie	H2020 INFRA	
Opis projektu	Projekt dotyczy bezpieczeństwa cyber-fizycznego usług z obszaru e-commerce/handlu elektronicznego.	
Rola	● Główny wykonawca i kierownik projektu po stronie polskiego partnera ● Koordynacja prac mających na celu wytworzenie komponentów w zakresie reagowania, łagodzenia skutków i odbudowy systemu narażonego na atak cyber-fizyczny ● Lider zadań mających na celu wytworzenie narzędzi wspomagających: o Zarządzanie reakcją i łagodzeniem skutków ataku z użyciem mechanizmów sztucznej inteligencji o Analizę i audyt po incydencie	

9.2 Jako wykonawca brałem udział w poniższych projektach:

a. Projekty kończące się po uzyskaniu stopnia doktora:

1. **Infrostrateg SWAROG**
    a. Źródło finansowania: NCBR
    b. Budżet: 8 657 006,25 zł

c. Okres realizacji: od 2021-12-01 (trwa)

d. Rola w projekcie: Ekspert AI

e. Nr Projektu: INFOSTRATEG-I/0019/2021

2. Horyzont 2020 STARLIGHT

a. Źródło finansowania: Horizon 2020 - Secure societies

b. Budżet: € 18 947 200,63

c. Okres realizacji: od 2021-10-01 (trwa)

d. Rola w projekcie: Ekspert AI i xAI

e. Grant Agreement: 101021797

3. Horizon Europe ULTIMATE

a. Źródło finansowania:HORIZON-CL4-2021

b. Budżet: € 3 529 112,50

c. Okres realizacji: od 2022-10-01 (trwa)

d. Rola w projekcie: Ekspert AI i xAI

e. Grant Agreement: 101070162

4. ATENA: Artificial inTElligence traiNing progrAmme (2021)

a. Źródło finansowania: Spinaker NAWA

b. Budżet: 369 490,00 PLN

c. Okres realizacji: 2021-05-01 - 2023-08-31

d. Rola w projekcie: Wykładowca, trener - Ekspert AI

e. Nr umowy o dofinansowanie: PPI/SPI/2020/1/00033/U/00001

5. **Horyzont 2020 SPARTA**
    a. Źródło finansowania: Horizon 2020 - INDUSTRIAL LEADERSHIP
    b. Budżet: € 15 999 913,00
    c. Okres realizacji: od 2019-02-01 do 2022-06-30
    d. Rola w projekcie: Ekspert AI i xAI
    e. Grant Agreement: 830892

6. **Horyzont 2020 SIMARGL**
    a. Źródło finansowania: Horizon 2020 - INDUSTRIAL LEADERSHIP
    b. Budżet: € 6 076 050,00
    c. Okres realizacji: od 2019-05-01 do 2022-04-30
    d. Rola w projekcie: Ekspert AI i cyberbezpieczeństwa
    e. Grant Agreement: 833042

7. **Horyzont 2020 PREVISION**
    a. Źródło finansowania: Horizon 2020 - Secure societies
    b. Budżet: € 9 040 230,00
    c. Okres realizacji: od 2019-09-01 do 2021-12-31
    d. Rola w projekcie: Ekspert AI i cyberbezpieczeństwa
    e. Grant Agreement: 833115

8. **Horyzont 2020 InfraStress**
    a. Źródło finansowania: Horizon 2020 - Secure societies
    b. Budżet: € 10 137 673,75
    c. Okres realizacji: od 2019-06-01 do 2021-09-30

d. Rola w projekcie: Ekspert AI i cyberbezpieczeństwa

e. Grant Agreement: 833088

9. **Horyzont 2020 MAGNETO**

a. Źródło finansowania: Horizon 2020 - Secure societies

b. Budżet: € 5 320 475,00

c. Okres realizacji: od 2018-05-01 do 2021-04-30

d. Rola w projekcie: Ekspert AI

e. Grant Agreement: 786629

10. **Horyzont 2020 SocialTruth**

a. Źródło finansowania: Horizon 2020 - INDUSTRIAL LEADERSHIP

b. Budżet: € 3 178 110,00

c. Okres realizacji: od 2018-12-01 do 2021-11-30

d. Rola w projekcie: Ekspert AI

e. Grant Agreement: 825477

b. Projekty kończące się przed uzyskaniem stopnia doktora

11. **Horyzont 2020 Q-Rapids**

a. Źródło finansowania: Horizon 2020 - INDUSTRIAL LEADERSHIP

b. Budżet: € 4 997 590,00

c. Okres realizacji: od 2018-03-01 do 2019-10-31

d. Rola w projekcie: Ekspert AI

e. Grant Agreement: 732253

10. Członkostwo w międzynarodowych lub krajowych organizacjach i towarzystwach naukowych wraz z informacją o pełnionych funkcjach.

- Członkostwo w Association for Computing Machinery (ACM)
- Członkostwo w Association for Information Systems (AIS)
- Członkostwo w International Neural Network Society (INNS)
- Jestem członkiem Rady Dyscypliny PBŚ (Informatyka Techniczna i Telekomunikacja)

11. Informacja o odbytych stażach w instytucjach naukowych lub artystycznych, w tym zagranicznych, z podaniem miejsca, terminu, czasu trwania stażu i jego charakteru.

Jednostka: University of Naples PARTHENOPE (staż naukowy)

Termin Stażu: 30 kwietnia 2024 do 1 sierpnia 2024

Miejsce: University of Naples "Parthenope" w Neapolu, we Włoszech.

Opis stażu: Zakres zainteresowań badawczych grupy naukowej z Parthenope w dużej mierze pokrywa się z moimi zainteresowaniami badawczymi. W trakcie stażu skoncentrowałem się więc na intensywnej pracy naukowej z zakresu wykorzystania sztucznej inteligencji w cyberbezpieczeństwie.

Ważnym aspektem mojego pobytu było badanie zastosowania technik granular computing w analizie ruchu sieciowego (NetFlow). Wyniki eksperymentów przeprowadzonych w tym temacie przedstawiono na konferencji ARES2024 (CORE B) [N1].

W trakcie stażu opracowałem nową metodę poprawy odporności systemów wykrywania włamań na ataki typu adversarial. Metoda ta polega na pomijaniu wybranych cech oraz zastosowaniu komitetu klasyfikatorów. Praca ta została przyjęta na konferencję ESORICS2024 (CORE A) [N4].

Również w kontekście wykorzystania ML w NIDS w trakcie współpracy z University of Naples PARTHENOPE przeprowadziłem badanie metryk sprawdzających trafność algorytmów xAI [N3].

Podczas pobytu w Neapolu nawiązałem nowe kontakty, które z czasem będą owocować dalszymi publikacjami naukowymi. Zorganizowałem oraz brałem udział w warsztatach

dotyczących nowoczesnych metod w cyberbezpieczeństwie oraz sztucznej inteligencji. Umożliwiło to wymianę wiedzy i doświadczeń z innymi badaczami oraz praktykami z branży. Zdobyta w czasie współpracy z włoskim zespołem wiedza pozwoliła mi na zastosowanie nowych metod w moich badaniach oraz podniesienie jakości mojej pracy naukowej. Przeprowadziłem prezentacje dla kadry akademickiej i studentów University of Naples "Parthenope", prezentując wyniki moich badań oraz dzieląc się wiedzą z zakresu sztucznej inteligencji, wyjaśnialności i cyberbezpieczeństwa.

Publikacje wynikające ze stażu z University of Naples PARTHENOPE:

- [N1] Komisarek, M., **Pawlicki, M.**, D'Antonio, S., Kozik, R., Pawlicka, A., & Choraś, M. (2024, July). Enhancing Network Security Through Granular Computing: A Clustering-by-Time Approach to NetFlow Traffic Analysis. In Proceedings of the 19th International Conference on Availability, Reliability and Security (pp. 1-8). **(70 pkt, CORE B)**

- [N2] Uccello, F., **Pawlicki, M.**, D'Antonio, S., Kozik, R., & Choraś, M. (2024, July). Comparative Analysis of AI-Based Methods for Enhancing Cybersecurity Monitoring Systems. In International Conference on Computational Science and Its Applications (pp. 100-112). Cham: Springer Nature Switzerland. (20 pkt)

- [N3] **Pawlicki, M.**, Pawlicka, A., Uccello, F., Szelest, S., D'Antonio, S., Kozik, R., & Choraś, M. (2024). Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. Neurocomputing, 128282. **(140 pkt, IF: 5.5)**

- [N4] **Pawlicki, M.**, Uccello, F., D'Antonio, S., Kozik, R., & Choraś, M. A Novel Method of Improving Intrusion Detection Systems Robustness Against Adversarial Attacks, through Feature Omission and a Committee of Classifiers, (przyjęte na ESORICS Workshops 2024), **(140 pkt, CORE A)**

12. Członkostwo w komitetach redakcyjnych i radach naukowych czasopism wraz z informacją o pełnionych funkcjach (np. redaktora naczelnego, przewodniczącego rady naukowej, itp.).

Guest Editor: Sensors and Pattern Recognition Methods for Security and Industrial Applications (SPR-SIA), A special issue of Sensors (ISSN 1424-8220).

https://www.mdpi.com/journal/sensors/special_issues/SPR-SIA

13. Informacja o recenzowanych pracach naukowych lub artystycznych, w szczególności publikowanych w czasopismach międzynarodowych.

Wykonałem, na zaproszenie, recenzje artykułów naukowych dla poniższych czasopism i konferencji:

13.1 Wybrane Czasopisma Naukowe:

Telecommunication Systems (Springer)

Print ISSN: 1018-4864

Electronic ISSN: 1572-9451

Impact Factor (IF): 1.7

Soft Computing (Springer)

Print ISSN: 1432-7643

Electronic ISSN: 1433-7479

Impact Factor (IF): 3.1

Bulletin of the Polish Academy of Sciences: Technical Sciences

ISSN: 2300-1917

Impact Factor (IF): 1.2

Future Generation Computer Systems (Elsevier)

Print ISSN: 0167-739X

Online ISSN: 1872-7115

Impact Factor (IF): 6.2

Artificial Intelligence Review (Springer)

Print ISSN: 0269-2821

Electronic ISSN: 1573-7462

Impact Factor (IF): 10.7

Applied Soft Computing (Elsevier)

Print ISSN: 1568-4946

Online ISSN: 1872-9681

Impact Factor (IF): 7.2

Foundations of Computing and Decision Sciences

Print ISSN: 0867-6356

Electronic ISSN: 2300-3405

Impact Factor (IF): 1.8

Applied Artificial Intelligence (Taylor & Francis)

Print ISSN: 0883-9514

Online ISSN: 1087-6545

Impact Factor (IF): 2.9

Machine Learning (Springer)

Print ISSN: 0885-6125

Electronic ISSN: 1573-0565

Impact Factor (IF): 4.3

Neurocomputing (Elsevier)

Print ISSN: 0925-2312

Online ISSN: 1872-8286

Impact Factor (IF): 5.5

Applied Sciences (MDPI)

ISSN: 2076-3417

Impact Factor (IF): 2.5

The Journal of Universal Computer Science (J.UCS)

Print ISSN: 0948-695X

Online ISSN: 0948-6968

Impact Factor (IF): 0.7

International Journal of Applied Mathematics and Computer Science

Print ISSN: 1641-876X

Online ISSN: 2083-8492

Impact Factor (IF): 1.6

13.2 Wybrane Konferencje:

ESORICS

Pełna nazwa: European Symposium on Research in Computer Security

CORE Ranking: A

ECAI

Pełna nazwa: European Conference on Artificial Intelligence

CORE Ranking: A

IEEE WCCI 2022

Pełna nazwa: IEEE World Congress on Computational Intelligence

CORE Ranking: A

ARES

Pełna nazwa: International Conference on Availability, Reliability and Security

CORE Ranking: B

DSAA 2023

Pełna nazwa: IEEE International Conference on Data Science and Advanced Analytics

CORE Ranking: B

ICCCI

Pełna nazwa: International Conference on Computational Collective Intelligence

CORE Ranking: B

IJCNN

Pełna nazwa: International Joint Conference on Neural Networks

CORE Ranking: B

ICIC

Pełna nazwa: International Conference on Intelligent Computing

CORE Ranking: -

MobiSec

Pełna nazwa: International Conference on Mobile Internet Security

CORE Ranking: -

ICAI

Pełna nazwa: International Conference on Applied Intelligence

CORE Ranking: -

IP&C

Pełna nazwa: International Conference on Image Processing and Communications

CORE Ranking: -

IEEE CSR

Pełna nazwa: IEEE International Conference on Cyber Security and Resilience

CORE Ranking: -

EICC

Pełna nazwa: European Interdisciplinary Cybersecurity Conference

CORE Ranking: -

ISD

Pełna nazwa: International Conference on Information Systems Development

CORE Ranking: -

14. Informacja o uczestnictwie w programach europejskich lub innych programach międzynarodowych.

Pełniłem/pełnię rolę Kierownika Projektu i eksperta AI po stronie polskiego partnera (ITTI. sp. z o.o.) w następujących projektach:

1. HE AI4CYBER od 2022 (trwa)
2. H2020 ELEGANT w okresie od 2021 do 2023
3. H2020 APPRAISE w okresie od 2021 do 2024
4. H2020 ENSURESEC w okresie od 2020 do 2022

Jako wykonawca/ekspert AI brałem udział w poniższych projektach:

1. H2020 STARLIGHT od 2021 (trwa)
2. HE ULTIMATE od 2022 (trwa)
3. H2020 SPARTA od 2019 do 2022
4. H2020 SIMARGL od 2019 do 2022
5. H2020 PREVISION od 2019 do 2021
6. H2020 InfraStress od 2019 do 2021
7. H2020 MAGNETO od 2018 do 2021
8. H2020 SocialTruth od 2018 do 2021
9. H2020 Q-Rapids od 2018 do 2019

Szczegółowe informacje podałem w pkt II.9

Udział w projektach europejskich	Liczba
Jako Kierownik Projektu	4
Jako Ekspert AI	13
Horyzont 2020	11
Horyzont Europa	2

15. Informacja o udziale w zespołach badawczych, realizujących projekty inne niż określone w pkt. II.9.

-

16. Informacja o uczestnictwie w zespołach oceniających wnioski o finansowanie badań, wnioski o przyznanie nagród naukowych, wnioski w innych konkursach mających charakter naukowy lub dydaktyczny.

-

III. INFORMACJA O WSPÓŁPRACY Z OTOCZENIEM SPOŁECZNYM I GOSPODARCZYM

Według bazy SCOPUS na podstawie prac do końca 2023 roku, 28.8% prac w których jestem współautorem powstała przy współpracy z naukowcami z innych państw/regionów (International collaboration), 67.5% prac zawiera współautorów z przemysłu (Academic-Corporate collaboration). Jest to doświadczenie pochodzące z pracy w projektach unijnych polegających między innymi na zastosowaniu badań we współpracy z przemysłem. Współpraca ta umożliwia dostęp do nowoczesnych technologii i infrastruktury. Pozwala również na realne zastosowanie wyników badań, walidując ich przydatność w świecie rzeczywistym.

1. Wykaz dorobku technologicznego

Dostęp do rzeczywistych danych sieciowych, zapewniany przez partnerów biznesowych, jest niezbędny dla rozwoju i walidacji nowoczesnych metod detekcji zagrożeń sieciowych. Te dane umożliwiają bezpośrednie zastosowanie technologii AI i ML w praktycznych scenariuszach identyfikacji ataków sieciowych oraz przyczyniają się do rozwijania nowych metod i technik badawczych.

- Zbiór danych wynikający ze współpracy z Telprom, zawiera zróżnicowane scenariusze ataków sieciowych, opisane, by służyć jako podstawa do dalszych badań w dziedzinie wykorzystania ML w NIDS. Opublikowany na Kaggle:
    - https://www.kaggle.com/datasets/ittibydgoszcz/appraise-h2020-real-labelled-netflow-dataset
- Zbiory danych wynikający ze współpracy z RoEduNet, zawiera zróżnicowane scenariusze ataków sieciowych, opisane, by służyć jako podstawa do dalszych badań w dziedzinie wykorzystania ML w NIDS. Opublikowane na Kaggle:
    - https://www.kaggle.com/datasets/h2020simargl/simargl2022
    - https://www.kaggle.com/datasets/dhoogla/simargl-2021

Dzięki publikacji tych zbiorów na platformie Kaggle nasze wyniki są dostępne dla szerokiej społeczności naukowej, co przyczynia się do globalnej wymiany wiedzy i współpracy w dziedzinie cyberbezpieczeństwa.

2. Informacja o współpracy z sektorem gospodarczym.

Niniejszy dział przedstawia przegląd współpracy naukowej i biznesowej w zakresie wykorzystania AI w cyberbezpieczeństwie, skupiając się na partnerstwach między sektorem akademickim a przemysłem. Za szczególnie ważne uważam współpracę z takimi firmami jak Johnson & Johnson oraz operatorami sieci, takimi jak Orange Polska, CESNET, Telprom i RoEduNet. Współpraca partnerami biznesowymi zapewnia unikalny dostęp do rzeczywistych danych sieciowych, co jest niezbędne dla rozwoju i walidacji nowoczesnych metod detekcji zagrożeń sieciowych. Takie partnerstwa pozwalają na bezpośrednie zastosowanie AI i ML do identyfikacji ataków sieciowych.

Orange Polska (operator sieci)

- Komisarek, M., **Pawlicki, M.,** Kowalski, M., Marzecki, A., Kozik, R., & Choraś, M. (2021, August). Network Intrusion Detection in the Wild-the Orange use case in the SIMARGL project. In Proceedings of the 16th International Conference on Availability, Reliability and Security (pp. 1-7).

CESNET, Praga, Czechy (operator sieci)

- **Pawlicki, M**., Zadnik, M., Kozik, R., & Choraś, M. (2022, June). Analysis and Detection of DDoS Backscatter Using NetFlow Data, Hyperband-Optimised Deep Learning and Explainability Techniques. In International Conference on Artificial Intelligence and Soft Computing (pp. 82-92). Cham: Springer International Publishing.

Telprom, informacijske tehnologije d.o.o., Słowenia (operator sieci)

- Komisarek, **M., Pawlicki**, M., Simic, T., Kavcnik, D., Kozik, R., & Choraś, M. (2023, August). Modern NetFlow network dataset with labeled attacks and detection methods. In Proceedings of the 18th International Conference on Availability, Reliability and Security (pp. 1-8).

RoEduNet, Strada Mendeleev 21-25, 010362 Bukareszt, Rumunia (operator sieci akademickiej)

- Mihailescu, M. E., Mihai, D., Carabas, M., Komisarek, **M., Pawlicki**, M., Hołubowicz, W., & Kozik, R. (2021). The proposition and evaluation of the roedunet-simargl2021 network intrusion detection dataset. Sensors, 21(13), 4319.

Collins Aerospace, Cork, Irlandia (obsługa teleinformatyczna sieci dla Johnson & Johnson)

- Komisarek, M., **Pawlicki, M**., Soboński, P., Pawlicka, A., Kozik, R., & Choraś, M. (2022). Extending machine learning-based intrusion detection with the imputation method. In Progress in Image Processing, Pattern Recognition and Communication Systems: Proceedings of the Conference (CORES, IP&C, ACS)-June 28-30 2021 12 (pp. 284-292). Springer International Publishing.

Inne przykłady współpracy z sektorem gospodarczym:

- Byłem jednym z ekspertów w międzynarodowym panelu dotyczącym bezpieczeństwa AI "CLAIRE AQUA" (pt."Cybersecurity of AI and AI for Cybersecurity" https://www.youtube.com/live/u44CiZhkbnY?si=_2bcGdT8ah8UoUj2
- Byłem gościem programu "Rozmowa Dnia" w Radiu PiK (jako ekspert od AI i cyberbezpieczeństwa) 16.09.2024 https://www.radiopik.pl/2,123336,polska-jest-najmocniej-atakowanym-krajem-na-swie
- Byłem gościem programu „Pomysł Innowacja Kariera" w Radiu PiK (jako ekspert od AI i cyberbezpieczeństwa) https://archiwum.radiopik.pl/service.go?action=details&show.Record.id=5072101&new=1
- Byłem gościem programu „Regionalny Punkt Widzenia" w Radiu PiK (jako ekspert od AI i cyberbezpieczeństwa) http://www.new.archiwum.radiopik.pl/service.go?action=details&show.Record.id=5063173&new=1
- Byłem gościem programu „Pytać Każdy Może" w Radiu PiK (jako ekspert od cyberbezpieczeństwa) https://www.radiopik.pl/index.php?idp=145&idx=2&szukaj=&s=27

3. Uzyskane prawa własności przemysłowej, w tym uzyskane patenty, krajowe lub międzynarodowe.

-

4. Informacja o wdrożonych technologiach.

W trakcie realizacji projektów badawczo-rozwojowych zakładających współpracę z partnerami przemysłowymi (m.in. Collins Aerospace, Johnson & Johnson/DePuy Synthes, operatorzy sieci tacy jak Orange, RoEduNet, Telprom) opracowane i pilotowane były takie technologie i rozwiązania:

- Systemy detekcji ataków sieciowym (NIDS) z wykorzystaniem AI/ML
- Pilotaż komponentu wykorzystującego AI w NIDS w ramach projektu H2020 InfraStress do środowiska przemysłowego (infrastruktura DePuy Synthes, Johnson

& Johnson w Cork, Irlandia). Rozwiązanie to umożliwia automatyczne wykrywanie anomalii i ataków sieciowych z wykorzystaniem uczenia maszynowego oraz zaawansowanych metod analizy danych

- Implementacja algorytmów uczenia głębokiego (Deep Learning) w środowisku Python z wykorzystaniem bibliotek TensorFlow i Keras do analizy ruchu sieciowego i detekcji

- Integracja mechanizmów xAI z systemami detekcji, pozwalająca zespołom SOC (Security Operations Center) lepiej rozumieć i zaufać wynikom generowanym przez modele. (H2020 STARLIGHT, HE AI4CYBER)

5. Informacja o wykonanych ekspertyzach lub innych opracowaniach wykonanych na zamówienie instytucji publicznych lub przedsiębiorców.

Jako jeden z członków trzyosobowego zespołu odpowiedzialnego za tworzenie Gender Equality Plan (GEP) dla Politechniki Bydgoskiej, miałem okazję bezpośrednio przyczynić się do analizy, opracowania oraz wdrożenia strategii na rzecz równości płci. Uczestniczyłem w zbieraniu danych, diagnozowaniu problemów oraz formułowaniu konkretnych rekomendacji. Współpraca nad planem równości płci na Politechnice Bydgoskiej wymagała szczegółowej analizy obecnej sytuacji w zakresie równouprawnienia i dyskryminacji na uczelni. Zebrano i przeanalizowano dane dotyczące zatrudnienia, awansów oraz uczestnictwa kobiet i mężczyzn w różnych aspektach życia akademickiego. Opracowanie to było konieczne do spełnienia wymogów projektów finansowanych z zewnętrznych źródeł, które coraz częściej wymagają od uczelni wykazania działań na rzecz równości płci. Plan zawierał szczegółowe wnioski i rekomendacje dotyczące wprowadzenia konkretnych rozwiązań mających na celu eliminację wszelkich form dyskryminacji i promowanie równości.

Dokument został przyjęty i wdrożony przez JM Rektora, co oznacza, że zalecane w planie działania stały się częścią oficjalnej polityki Uczelni.

6. Informacja o udziale w zespołach eksperckich lub konkursowych.

-

7. Informacja o projektach artystycznych realizowanych ze środowiskami pozaartystycznymi.

-

IV. INFORMACJE NAUKOMETRYCZNE

1. Informacja o punktacji Impact Factor (w dziedzinach i dyscyplinach, w których parametr ten jest powszechnie używany jako wskaźnik naukometryczny).

	Liczba prac z IF	Suma IF
Razem	35	185.131
Po uzyskaniu stopnia doktora	31	178.035
Przed uzyskaniem stopnia doktora	4	7.078

2. Informacja o liczbie cytowań publikacji wnioskodawcy, z oddzielnym uwzględnieniem autocytowań.

	Liczba wszystkich cytowań	Liczba cytowań bez autocytowań
Web of Science	680	538
Scopus	1124	853
Google Scholar	1745	dane niedostępne

3. Informacja o posiadanym indeksie Hirscha.

	h-index

Web of Science	13
Scopus	17
Google Scholar	22

4. Informacja o liczbie punktów MNiSW.

	Liczba prac z listy MNiSW	Suma punktów MNiSW
Ogółem	115	9234
Po uzyskaniu stopnia doktora	96	8100
Przed uzyskaniem stopnia doktora	19	1134

Informacje zawarte w pkt. IV powinny wskazywać również na bazę danych, na podstawie której zostały podane. Przy wyborze tej bazy należy zwracać uwagę na specyfikę dziedziny i dyscypliny naukowej, w której kandydat ubiega się o nadanie stopnia doktora habilitowanego. Rada Doskonałości Naukowej informuje, że podawanie danych naukometrycznych – w opinii Rady Doskonałości Naukowej – jest wskazane i zalecane, wynika to także ze stosowanej powszechnie praktyki przez samych kandydatów ubiegających się o awans naukowy. Należy jednak podkreślić, że podane we wnioskach o wszczęcie postępowania awansowego dane naukometryczne nie mogą stanowić kryterium oceny dorobku naukowego Kandydata dla podmiotów doktoryzujących, habilitujących oraz samej Rady Doskonałości Naukowej, organów prowadzących postępowania w sprawie nadania stopnia lub tytułu. Zadaniem tych organów jest przede wszystkim ocena ekspercka dorobku naukowego Kandydata ubiegającego się o awans naukowy, zaś decyzja o nadaniu stopnia lub tytułu nie powinna być uzależniona od podania tych danych.

PODPIS ZAUFANY
MAREK
PAWLICKI
01.01.2025 13:04:52 [GMT+1]
Dokument podpisany elektronicznie
podpisem zaufanym

Marek Pawlicki

……

(podpis wnioskodawcy)