Neurocomputing 590 (2024) 127759

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep learning and artificial intelligence used in cybersecurity

Marek Pawlicki [a,b], Aleksandra Pawlicka [a,c], Rafał Kozik [a,b], Michał Choraś [a,b,*]

[a] ITTI, Poznań, Poland
[b] Bydgoszcz University of Science and Technology, Bydgoszcz, Poland
[c] University of Warsaw, Warsaw, Poland

ARTICLE INFO

Keywords:
Cybersecurity
Explainability
Explainable AI
Future
Research directions
xAI

ABSTRACT

This paper engages in a comprehensive investigation concerning the application of Explainable Artificial Intelligence (xAI) within the context of deep learning and Artificial Intelligence, with a specific focus on its implications for cybersecurity. Firstly, the paper gives an overview of xAI techniques and their significance and benefits when applied in cybersecurity. Subsequently, the authors methodically delineate their systematic mapping study, which serves as an investigative tool for discerning the potential trajectory of the field. This strategic methodological framework lets one identify the future research directions and opportunities that underlie the integration of xAI within the realm of Deep Learning, Artificial Intelligence, and cybersecurity, which are described in-depth. Then, the paper brings together all the gathered insights from this extensive investigation and closes with final conclusions.

1. Introduction

Today, the society's reliance on the Internet and its associated services spans across all the sectors, transforming it into a critical infrastructure that supports the whole society. The stakes are exceptionally high particularly in the fields dealing with sensitive data or those with potential serious consequences, such as healthcare, law enforcement, finance, and national security. In these domains, a well-designed cyberattack could not only result in the loss of sensitive data but also damage public trust, have profound economic implications or even endanger lives.

This evolving scenario underscores the crucial importance of cybersecurity as a discipline that transcends and encompasses all the others. In an era where digital technologies are integrated into every aspect of daily life and business operations, cybersecurity is no longer an optional add-on but a fundamental layer of defense. It acts as the basis upon which the security of all the other sectors is built, ensuring the integrity, availability, and confidentiality of data [1].

In the highly interconnected digital landscape of today, where citizens face a multitude of cyber-threats on a daily basis, network intrusion detection systems (NIDS) have emerged as indispensable guardians of network security. These systems offer real-time monitoring capabilities, enabling the identification of any suspicious activities. Nevertheless, as the networks grow increasingly intricate and cyberattacks become more sophisticated, the task of accurately detecting and categorizing intrusions has grown progressively more challenging [2].

To address the challenge posed by NIDSs, machine learning (ML)/ deep learning (DL) algorithms [3,4] have emerged as a promising solution, showing significant potential in enhancing detection accuracy. And yet, these sophisticated algorithms often operate in an opaque or black-box manner, e.g. their inner workings not being understandable to human operators. Consequently, operators find it challenging to comprehend the reasoning behind specific decisions made by these algorithms [5,6]. When achieving transparency in the decision-making process proves elusive, concerns arise regarding the reliability and trustworthiness of these models. This concern becomes particularly critical in the high-stakes domains such as healthcare, law enforcement, autonomous vehicles, finance, and, notably, in the context of this research, cybersecurity [7,8].

The absence of transparency in AI systems has led to the proposition of making AI decision understandable, giving rise to the concept of explainable AI (xAI). Explainability entails the capacity of humans to

[*] Corresponding author at: Bydgoszcz University of Science and Technology, Bydgoszcz, Poland.
E-mail address: chorasm@pbs.edu.pl (M. Choraś).

https://doi.org/10.1016/j.neucom.2024.127759
Received 7 February 2024; Received in revised form 26 March 2024; Accepted 22 April 2024
Available online 25 April 2024

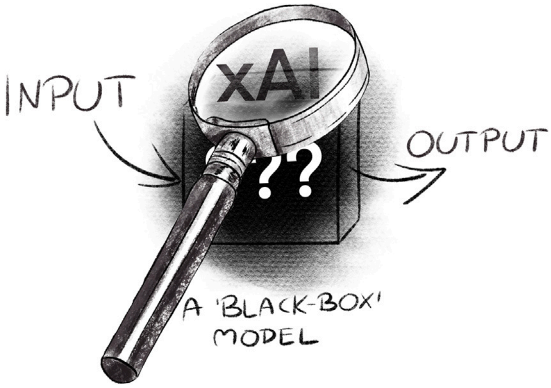


Fig. 1. The main concept behind explainable AI (xAI).

comprehend and interpret the decision-making process of ML algorithms. Within the context of network intrusion detection, the introduction of explainability can help recognize the decisive data attributes contributing to specific decisions, thereby facilitating human analysts in grasping the nature of threats and devising effective countermeasures. Furthermore, explainability can serve as a means to uncover biases and errors within models, leading to enhancements in their performance [9]. Fig. 1 shows the graphical representation of the main concept behind explainable AI, serving the role of a magnifying glass used to peek inside the "black-box" model.

This transparency, or rather, lack thereof, especially in the context of cybersecurity and network intrusion detection, may have real-world consequences and significance. To name just a few examples, explainable AI methods can aid in improved identifying and understanding the behaviour patterns indicative of threats and consequently in the identification and analysis of anomalous network traffic, which could indicate a cybersecurity threat. By understanding the reasons behind the classification of traffic as anomalous, security professionals can more effectively and accurately respond to potential threats [10]. Similarly, the use of xAI in malware detection allows for a deeper understanding of the features that lead to the classification of a file or process as malicious. This can improve trust in malware detection systems and facilitate the development of more effective countermeasures [11]. In the broader cybersecurity context, explainability is also of exceptional value. For example, it can significantly enhance phishing detection systems by providing insights into why certain emails or websites are flagged as malicious. This can help security analysts understand the characteristics of phishing attempts and improve the accuracy of detection models over time [12].

Meske et al. have remarked that explainability should not be perceived as a byproduct "problem arising through AI". Rather, it is "as old as the topic of AI itself" [13]. And yet, as pointed out by Mathew, explainability "is still in its early stages", which leaves plenty of room for future research [14]. In turn, it is the innovative research on xAI which has the potential to positively influence the application of AI in general [15].

It has been projected that the hype cycle for xAI will last longer than 10 years [15]. In other words, the expectations and excitement about Explainable AI are expected to persist at high levels for an extended period before reaching a more stable and realistic phase of adoption. This can have various implications, such as extended investments, advancements in the field, and potential challenges in managing heightened expectations and delivering on the promises of Explainable AI. Primarily, it signifies that the field will undergo intensive research efforts. Bearing these considerations in mind, the authors of this paper aimed to identify research opportunities for xAI, and to analyse and categorize potential research directions. This work presents the results of their study.

This paper is structured as follows: in Section 2, the background for the study has been laid out. Section 3 discusses the materials and methods applied in the study presented. In Section 4, the detailed results of the study have been shown, with final conclusions coming thereafter.

In Fig. 2, the structure of this paper has been illustrated in the form of a roadmap.

2. Background

In addition to its importance in the context of AI and machine learning in general, xAI is increasingly used due to the need for transparency and accountability. The xAI techniques can greatly enhance the interpretability of machine learning models, making them more accessible for stakeholders and building trust in the models [16]. In the context of cybersecurity, explainable Artificial Intelligence plays a fundamental role by enhancing the transparency and interpretability of machine learning models, allowing security professionals to effectively scrutinize and detect potential vulnerabilities, threats, or adversarial attacks, ultimately contributing to more robust and resilient cyber-defense mechanisms. In this section, the most prominent xAI techniques applicable in cybersecurity/network intrusion detection have been presented.

2.1. Overview of xAI techniques applicable in network intrusion detection

Several explainable AI techniques can be harnessed for network intrusion detection; some of the most notable ones have been explored in this section.

One of the ways of organizing xAI methods proposed in the subject literature has been dividing them into local and global ones, and the ones which could be classified into both categories.

2.2. Local and global explanations

Local and Global Explanations are two distinct xAI methodologies. The emphasis of Local explanations is on working out the rationale behind specific decisions made by the model in question. Their aim is to answer why a model made a specific decision with regard to a specific instance. These methods give insights into how each attribute influences a specific decision. Conversely, Global explanations aim to offer a comprehensive overview of the model's decision-making process over a broad range of instances, providing a macroscopic perspective on the model's operational dynamics.

2.3. Local xAI methods

2.3.1. Scoped rules (ANCHORS)

ANCHORS is a model-agnostic perturbation-based xAI approach within explainable AI that generates "if-then" rules, establishing clear demarcations for its explanations. It concentrates on the relationships between inputs and outputs. The method navigates the challenges of computational complexity in high-dimensional spaces by probabilistically defining rules, selecting candidates until their accuracy reaches statistical significance. Once the precision threshold is surpassed, the generated explanations gain relevance across a wider array of instances [17].

Fig. 3 presents a sample screenshot showing an explanation using the ANCHORS method, in the context of NIDS, as applied by the authors of the paper.

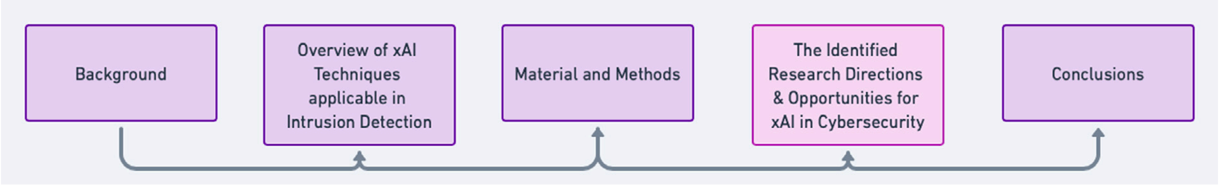
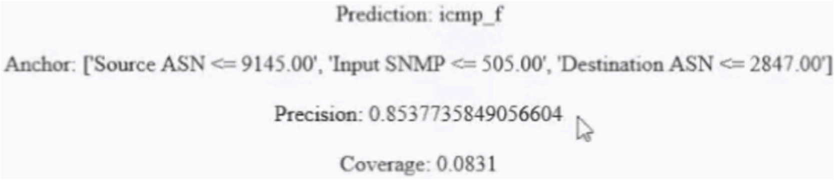

Background
Overview of xAI Techniques applicable in Intrusion Detection
Material and Methods
The Identified Research Directions & Opportunities for xAI in Cybersecurity
Conclusions

Fig. 2. The roadmap of this paper.



Prediction: icmp_f

Anchor: ['Source ASN <= 9145.00', 'Input SNMP <= 505.00', 'Destination ASN <= 2847.00']

Precision: 0.8537735849056604

Coverage: 0.0831

Fig. 3. An explanation using the ANCHORS method, in the context of NIDS; a snapshot from the authors' framework.

2.3.2. LIME (Local Interpretable Model-Agnostic Explanations)

LIME, an algorithm introduced in 2016, serves the purpose of providing explanations for the predictions made by intricate machine learning models. It accomplishes this by constructing a "local" model that approximates the behaviour of the original model within the vicinity of a particular input instance. LIME's model-agnostic nature renders it compatible with diverse machine learning models, including deep neural networks, decision trees, and support vector machines.

The LIME algorithm unfolds in several steps. Initially, it selects the instance requiring explanation. Subsequently, it generates perturbed versions of the instance by introducing random alterations or noise to the input features. The quantity of perturbed instances depends on factors such as the complexity of the original model and the desired level of accuracy. The next step involves the computation of weights assigned to interpretable features. This is achieved through training a linear model on the perturbed instances, with interpretable features as inputs and the output representing the predicted probability of the original model. The weights derived from the linear model are then employed to gauge the significance of each feature in the prediction. The final stage entails the creation of the local model. This is executed by selecting a subset of interpretable features based on their importance weights and training a straightforward interpretable model. The local model subsequently explains predictions by displaying the contribution of each feature to the output [18,19].

Fig. 4 presents a sample screenshot showing an explanation using the LIME method, in the context of NIDS, as applied by the authors of the paper.

2.4. Diverse Counterfactual Explanations (DiCE)

Counterfactual Examples (CE) are datapoints created to answer what-if scenarios on what would have to happen to a particular sample to flip its label, while perturbing it minimally. In DiCE, the adjustment of features contains constraints for proximity to the perturbed instance, and diversity, so the constructed CEs are different from one another [20].

Fig. 5 presents a sample screenshot showing an explanation using the DiCE method, in the context of NIDS, as applied by the authors of the paper.

2.4.1. Individual Conditional Expectation (ICE)

ICE plots are a visualization tool used in data analysis to explore the relationship between a feature and the target outcome on a per-instance basis. They enable a detailed examination of how changes in a feature's value affect the prediction for each individual sample, providing a comprehensive understanding of the feature's impact across the dataset. This approach offers a more personalized insight into the data, highlighting the variability of the feature's effect among different instances, without assuming the features to be independent.

2.5. Global xAI methods

2.5.1. Decision trees

The decision tree technique, which constructs a tree-like representation of decisions and their potential outcomes, serves as a prominent illustration of a rule-based approach. Each node within the tree signifies a decision based on a specific data feature or attribute, while the edges represent the diverse consequences associated with that decision [21].

Roth et al. (2021) devised a reinforcement learning algorithm aimed at determining collision-free routes for robots. To mitigate errors, they transformed the algorithm into a decision tree, naming their approach XAI-N [22]. A distinct approach was introduced by Schaaf et al. (2019), where L1-orthogonal regularization was incorporated during network training to enhance the alignment of a decision tree with deep neural networks [23]. In the realm of cybersecurity, decision tree models have found application in enhancing the trust management of machine learning algorithms [24]. The authors posit that artificial intelligence derives conclusions by scrutinizing vast datasets to uncover potentially concealed patterns and subtle signals.

Fig. 6 presents a sample screenshot showing an explanation using the Decision Trees method, in the context of NIDS, as applied by the authors of the paper [25].

2.5.2. Rule lists

The production rule system, often referred to as a rule-based expert system, stands as another prevalent rule-based methodology [27]. This system comprises an ensemble of production rules that delineate the relationships between input and output variables. Production rules are commonly expressed in "if-then" statements, where the antecedent constitutes the conditions or prerequisites to trigger the rule, and the consequent represents the action executed upon rule activation [28]. Expert systems frequently find application in diagnostic and decision-making contexts [29].

The principal advantage of rule lists lies in their interpretability. They also tend to outperform decision trees, as they only necessitate a single pass through the input features for categorization [30]. Bahani et al. (2021) implemented a fuzzy algorithm for the classification of heart disease diagnoses [31].

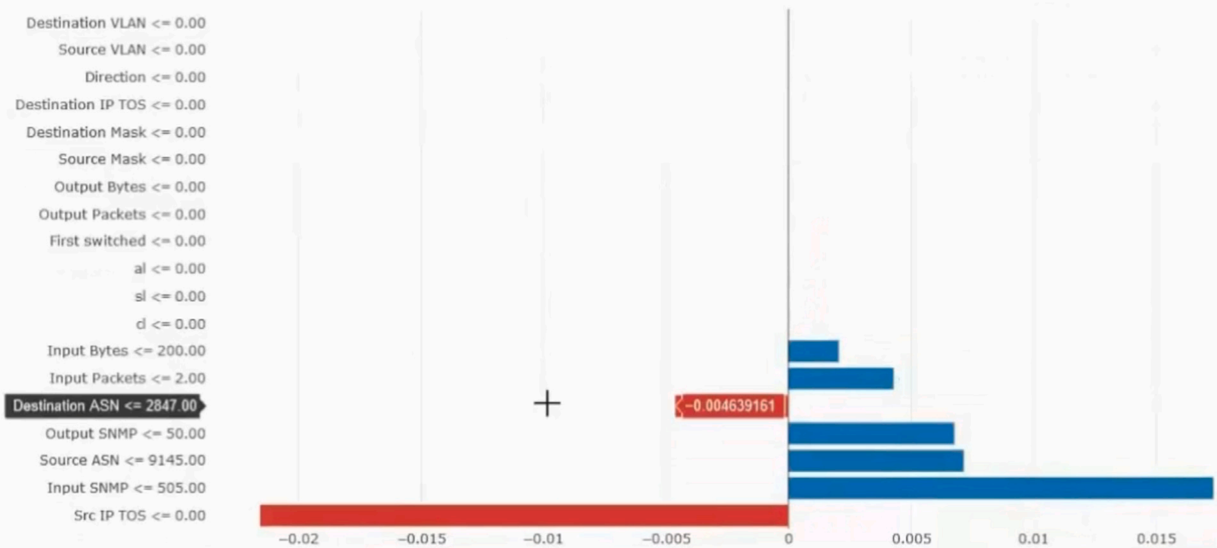
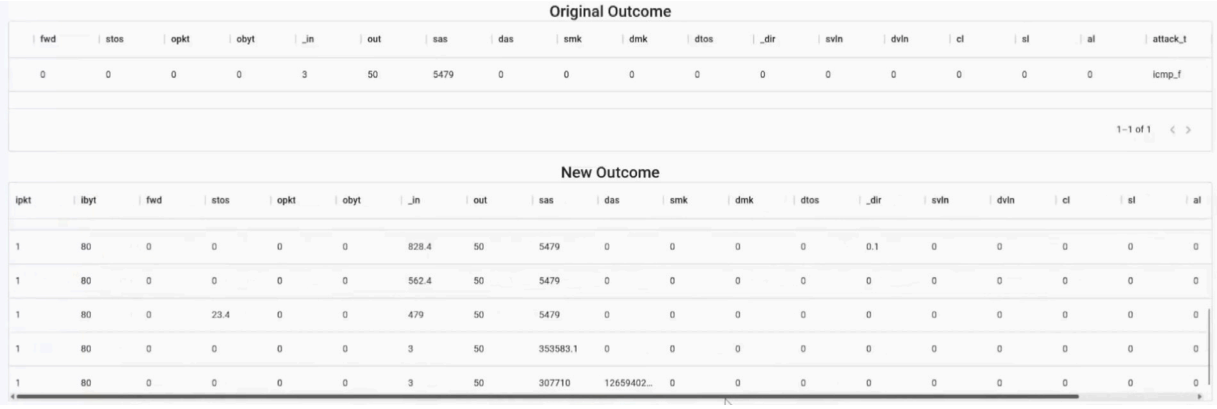


Fig. 4. An explanation using the LIME method, in the context of NIDS; a visualization coming from the LIME library [18] integrated into the authors' framework.


Original Outcome

fwd	stos	opkt	obyt	_in	out	sas	das	smk	dmk	dtos	_dir	svln	dvln	cl	sl	al	attack_t
0	0	0	0	3	50	5479	0	0	0	0	0	0	0	0	0	0	icmp_f

1–1 of 1

New Outcome

ipkt	ibyt	fwd	stos	opkt	obyt	_in	out	sas	das	smk	dmk	dtos	_dir	svln	dvln	cl	sl	al
1	80	0	0	0	0	828.4	50	5479	0	0	0	0	0.1	0	0	0	0	0
1	80	0	0	0	0	562.4	50	5479	0	0	0	0	0	0	0	0	0	0
1	80	0	23.4	0	0	479	50	5479	0	0	0	0	0	0	0	0	0	0
1	80	0	0	0	0	3	50	353583.1	0	0	0	0	0	0	0	0	0	0
1	80	0	0	0	0	3	50	307710	12659402...	0	0	0	0	0	0	0	0	0

Fig. 5. An explanation using the DiCE method, in the context of NIDS; a snapshot from the authors' framework.

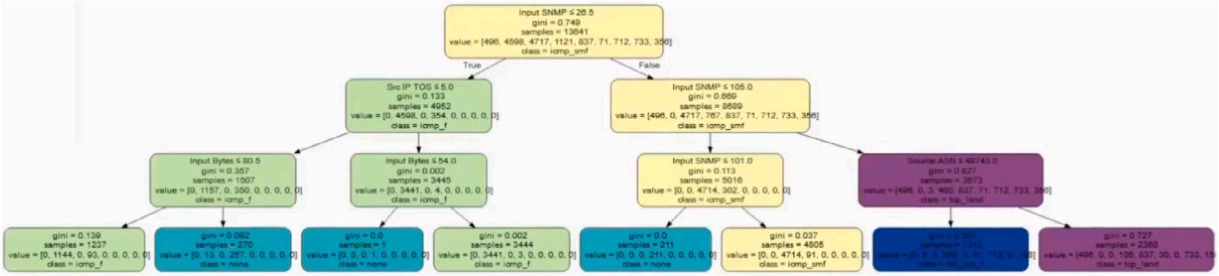


Fig. 6. An explanation using the Decision Trees method, in the context of NIDS; a visualization made with the Scikit-learn library [26] integrated into the authors' framework.

2.5.3. RuleFit

RuleFit represents a machine learning methodology that fuses decision trees with linear models to create a hybrid model proficient in capturing both linear and non-linear data associations. To detect non-linear relationships within the data, the RuleFit algorithm initially constructs a decision tree ensemble, frequently employing a random forest. Subsequently, through a process called rule extraction, the decision tree algorithm is transformed into a set of rules. These extracted rules are then amalgamated with linear models like linear regression or logistic regression to craft the hybrid model. While the linear models capture the linear data patterns, the non-linear correlations are encapsulated by the rules sourced from the decision tree ensemble. During training, the model learns the weights associated with each component [19,32].

Luo, Chao, et al. applied this methodology in the context of cancer prediction. Their results demonstrated the RuleFit-based nomogram's accuracy in predicting survival among individuals with nasopharyngeal carcinoma. In terms of discrimination and calibration, the nomogram surpassed previous models that omitted inflammatory markers [33].

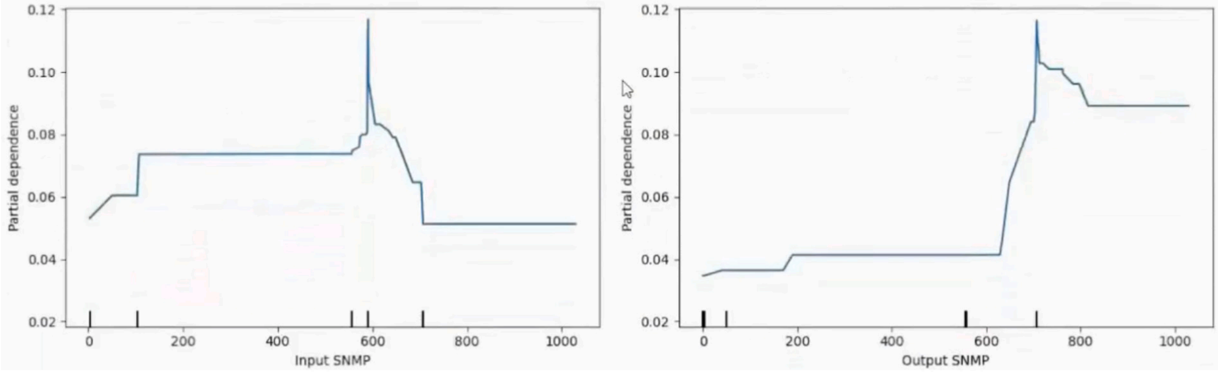


Fig. 7. An explanation using the PDP method, in the context of NIDS; a visualization coming from the PDP library integrated into the authors' framework.

2.5.4. Linear models

represent mathematical models explaining the relationship between a dependent variable and one or more independent variables. Frequently deployed in regression tasks, they predict the dependent variable's value based on the independent variables' values. Linear models offer interpretability by furnishing insights into the magnitude and direction (sign) of the coefficients associated with the independent variables [34].

2.5.5. Partial Dependence Plots (PDP)

PDPs are a tool for visualizing the connection between a subset of input features and the predicted outcome, while treating other features as independent. This makes it possible to understand the global model behaviour. Given the cognitive constraints of humans, features are analysed either individually or in small sets. For the feature, PDPs calculate the average predicted response for variations within a specific feature; thus, they facilitate an understanding of how changes in the feature affect the target variable [18].

Fig. 7 presents a sample screenshot showing an explanation using the PDP method, in the context of NIDS, as applied by the authors of the paper.

2.6. Both global and local xAI methods

2.6.1. Rule-based methods

Rule-based techniques represent a form of explainable AI that operates by constructing a set of explicit rules to explain the decision-making process of the model. These rules are intelligible to humans, and their logic is readily understandable [35]. The rules can either be established manually by domain experts or acquired through a rule-learning algorithm [19,21].

There are several advantages associated with the adoption of rule-based strategies in contrast to other machine learning models. They possess an inherent readability, and their decision-making process is transparent, facilitating the detection and rectification of model errors. Furthermore, they exhibit high scalability and adaptability in handling both continuous and discrete data [36]. Nevertheless, rule-based methods also carry certain drawbacks. They demand substantial prior domain knowledge and might fall short in capturing intricate interactions among variables. Additionally, they prove inadequate when handling noisy or missing data, and they are susceptible to alterations in data distribution [37].

2.6.2. Certifai

Certifai stands as a versatile tool applicable to any black-box model and various input data types, offering the CERScore, a measure of black-box model robustness that surpasses the performance of methods with access to internal model details [38]. By making specific selections, the algorithm confines the values of sampled points, enabling the generated counterfactuals to reflect a user's concept of how much they can alter their features. Building on this work, a framework named "Cortex Certifai" was introduced [39].

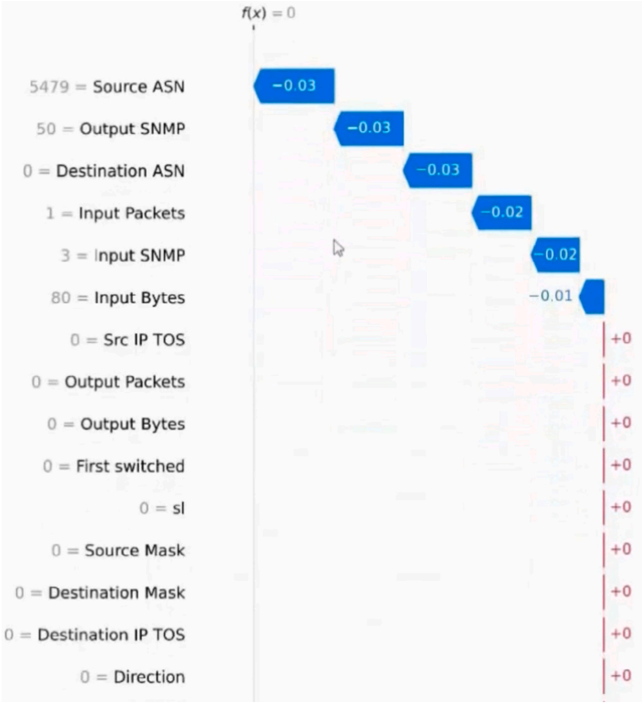


Fig. 8. An explanation using the SHAP method, in the context of NIDS; a visualization coming from the SHAP library integrated into the authors' framework.

2.6.3. SHAP (SHapley Additive exPlanations)

SHAP, an approach within machine learning, assigns scores to individual input features to explain predictions. The SHAP score represents the disparity between the expected prediction when a feature is present and when it is absent. This calculation is performed across all possible feature combinations and averaged. The first step in the SHAP algorithm involves establishing baseline predictions, which reflect the model's average prediction across the entire dataset.

The computation of SHAP values for each input feature employs the Shapley value, a concept rooted in cooperative game theory, to allocate contribution values to each feature. Considering interactions with other features, the Shapley value indicates a feature's marginal contribution to the prediction. The final step entails amalgamating these values to attain a comprehensive interpretation of the prediction. This is often achieved by displaying the values associated with each feature through bar plots or summary plots [19,40].

Fig. 8 presents a sample screenshot showing an explanation using the SHAP method, in the context of NIDS, as applied by the authors of the paper.

2.6.4. ProtoDash

ProtoDash is a technique designed to identify "Prototypical Samples" in a dataset, which is essential in understanding the defining features of a particular subset or class within the data by pointing to the samples which represent it the best. By efficiently selecting samples that maximize similarity, ProtoDash offers a concise representation of the target class, allowing for a clearer insight into its essential characteristics [41].

3. Material and methods

In this paper, the authors have conducted a systematic targeted review of the existing literature to identify the potential research directions and opportunities for explainable AI in cybersecurity. The applied approach involves a blend of quantitative and qualitative research design methodologies.

In this section, the authors have provided an overview of the methodology employed in the study presented. The study is based on a systematic mapping study conducted following the guidelines outlined by Petersen et al. [42]. The initial step involved formulating the research question: *"What are the potential research directions/ research opportunities for the explainability techniques employed in cybersecurity, i.e., network intrusion detection?"*

The answer to this question has been provided in Section 4.

3.1. The course of the study

Following the formulation of the research question, the study proceeded through the following stages:

1. Definition of Search String: A search string was defined to identify relevant papers.
2. Papers Search: The search for papers was initiated using the defined search string.
3. Inclusion and Exclusion Criteria: Inclusion and exclusion criteria were then established, and they were subsequently applied to the papers identified in the search.
4. Categorization of Papers: The identified papers were categorized according to their relevance and content.
5. Data Extraction: The necessary data was extracted from the categorized papers.
6. Data Analysis: Finally, the collected information was subjected to analysis in order to address the research question.

3.2. Material collection methodology

This study employed the systematic literature review approach, following the principles of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [43]. It involved the analysis of high-quality papers sourced from reputable sources.

The papers were retrieved from the following databases: Science Direct, IEEE Xplore, DBLP, and arXiv. Additionally, a search on ResearchGate was performed, followed by a supplementary Google search.

The selection of databases for retrieving papers was a critical step in the research process, aimed at encompassing a comprehensive and diverse range of scientific contributions. Science Direct and IEEE Xplore were chosen for their extensive collections of peer-reviewed articles in the fields of computer science and engineering, ensuring coverage of foundational and cutting-edge research in cybersecurity and AI. DBLP was included for its broad indexing of computer science bibliography, offering access to a wide array of conference proceedings and journals. Then, arXiv was selected for its repository of preprints, allowing the authors to incorporate the most recent findings not yet published in peer-reviewed venues. Finally, ResearchGate and supplementary Google searches were conducted to possibly capture grey literature and works in progress, broadening the review to include emerging insights and trends not yet formally published.

Table 1
The preliminary search results for the Search String.

(("xai" OR "explainable AI" OR "explainability") AND ("research directions" OR "research opportunities"))

Source	Number of papers found
ScienceDirect	1636
IEEE Xplore	27
DBLP	52
ResearchGate	10 000
arXiv	48
complimentary Google search	8
Total	11 771

Table 2
The search results for the complimentary Search String.

(("xai" OR "explainable AI" OR "explainability") AND "future")

Source	Number of papers found
ScienceDirect	2034
IEEE Xplore	25
DBLP	8
ResearchGate	10 000
arXiv	106
complimentary Google search	5
Total	12 178

3.3. Building the search strings

To answer the research question in the course of the literature study, proper search strings had to be determined. In order to do so, the construction of the search strings followed the PICO (Population, Intervention, Comparison, and Outcome) technique [44]. As mentioned in [45], in the case of systematic mapping studies, it is enough to use the criteria of Population and Intervention only; therefore, in this study, the strings were constructed as follows: Population: in the context of the research questions, Population is explainable AI in Network Intrusion Detection. Intervention: in this context, it is the word "research directions" and "research opportunities". Using the identified keywords, a search string was constructed:

(("xai" OR "explainable AI" OR "explainability") AND ("intrusion detection" or "cybersecurity") AND ("research directions" OR "research opportunities"))

As this initial search string proved to be too specific and the number of results was unsatisfactory, it was then simplified as follows:

(("xai" OR "explainable AI" OR "explainability") AND ("research directions" OR "research opportunities"))

The number of primary search results for this new string has been shown in Table 1.

The initial search yielded a total of 11 771 results.

Following a discussion, the authors wished to enhance the quality of the results; in order to do so, they decided to apply one more search string, incorporating the concept of a more general "future" of xAI. Thus, the complimentary search string was used:

(("xai" OR "explainable AI" OR "explainability") AND "future")

In this case, the numbers of results were as presented in Table 2.

This search uncovered additional 12 178 potential hits. Collected together, the results underwent thorough screening by the authors. In cases where the number of results was particularly extensive, the search outcomes were scrutinized until they reached a point of diminishing relevance or inclusion of relevant materials. This rigorous selection process resulted in 985 papers being chosen for further examination.

It must be noted that in the final analysis and selection of papers for this study, the decision was ultimately guided by the extensive expertise of the authors, ensuring that despite the broad initial search, the focus remained sharply on works of substantial relevance and depth within the cybersecurity domain.

Following this, to ensure the selection of the most relevant and valuable papers, specific inclusion and exclusion criteria were defined.

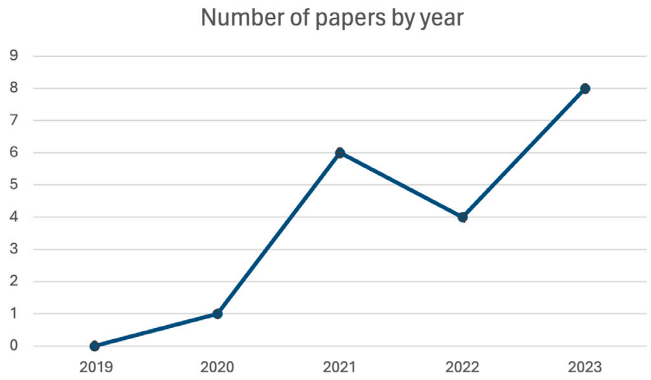


Fig. 9. The papers used for the study, by year of publication.

These criteria were developed based on both the research expertise of the authors and a comprehensive review of pertinent survey papers in related fields. The inclusion criteria for selecting papers were carefully designed to ensure the relevance, currency, and quality of the literature included in this systematic mapping study. The ultimate inclusion criteria for the papers were as follows:

- Peer-reviewed scientific papers;
  they were prioritized to guarantee the credibility and rigour of the research which was analysed, adhering to the academic standards of evidence-based inquiry.
- Written in English;
  limiting the review to works written in this language allowed for a broad yet manageable scope, considering English as the predominant language of scientific communication.
- No older than 5 years old;
  the imposed publication window of no more than five years helped focus on contemporary developments in the field, reflecting the rapid pace of innovation in cybersecurity and explainable AI.
- Available to authors;
  accessibility to them was essential to ensure that all selected studies could be thoroughly reviewed and analysed, thus avoiding the potential bias of excluding significant research not readily available.
- No duplicates;
  the exclusion of them was crucial for maintaining the clarity and efficiency of the data synthesis process, ensuring that each selected piece of literature contributed unique insights to the study.

The criteria were subsequently employed to assess the downloaded papers. Initially, they were applied to the abstract and available metadata, resulting in the retention of 68 papers.

Finally, when necessary, the inclusion and exclusion criteria were extended to encompass the full papers. The authors further refined the results; if a paper was deemed especially relevant to the study but did not fully meet the inclusion criteria, it was included following a group discussion.

As a result, a total of 19 papers were ultimately selected for inclusion in this category of the study, [46–64].

In Fig. 9, the number of papers included in the final part of the analyses have been presented according to their date of publication.

The composition of the selected papers, with 8 from 2023, 4 from 2022, 6 from 2021, and 1 from 2020, underscores the cutting-edge nature of our research corpus. The absence of papers from 2019 and the concentration of more recent publications not only adhere to the authors' criterion of studying works no older than five years but also ensure that the analysis presented is grounded in the latest developments and trends in the domain of cybersecurity and explainable AI.

Once the papers for data extraction were chosen, the authors engaged in a group discussion to determine the specific data to be extracted and included in the data extraction form. It was agreed among the authors that one author would carry out the data extraction process, while the remaining authors would verify the results of this process.

To address the research question effectively, it was essential to identify suggestions and ideas regarding the future directions of xAI research and the emerging research opportunities. The additional data extracted included the authors of the paper, the publication year, the paper's title, and the URL to the PDF (if available).

The results of the data extraction process, as well as the conclusions drawn from it, will be presented in the subsequent sections.

4. The identified research directions & opportunities for xAI in cybersecurity

The following section showcases the results of the targeted literature study — the identified key research directions for xAI, with the particular focus on its application in the cybersecurity domain. These include developing intuitive, user-centred explanations for complex AI models, establishing robust evaluation metrics, integrating xAI into various applications, and addressing ethical and regulatory concerns. A number of more obscure, less-discussed but equally interesting research opportunities have been discussed as well. These opportunities collectively drive the advancement of xAI and its potential to enhance the transparency and trustworthiness of AI systems.

4.1. Adopting more user-centred approaches

One of the research opportunities of xAI which are out the most are the aspects of making the explanations user-centred.

The fundamental objective of contemporary AI initiatives is to contribute to the creation of artificial intelligent systems that prioritize human needs and expectations. These systems are designed to engage with humans in an interpretable and explainable manner, with a primary focus on ensuring fairness, transparency, and accountability [53]. In traditional automated systems, automation itself has been at the core, expecting users to adapt to its functionalities. However, advanced automation does not necessarily lead to improved operator performance. Human-centred AI seeks to revolutionize this paradigm by placing user needs, goals, and capabilities at the forefront of automation design [53].

A user-centred approach, grounded in human factors, cognitive science, and user experience, is advocated to engineer user-friendly AI solutions for process industries [53] Yet, as mentioned by [63], interpreting and validating the reasoning behind machine learning models can be a daunting task. After all, the act of providing explanations, whether by a human or an algorithm (as in xAI), is inherently social. Effective communication of explanations necessitates their adaptation to the context of the recipients. As such, meaningless or overly complex explanations can erode user trust in the system. Furthermore, the requirement for explanations varies depending on the context, and understanding what qualifies as a meaningful explanation to the user requires a deep understanding of both the user and their context [53]. To mitigate this challenge, they advocate for the development of novel methods to identify a meaningful subset of the dataset for interpretation, subsequently facilitating the interpretation of relationships between various data samples and subsets [63]. A number of methodologies have been proposed to understand users better, such as mental model elicitation, Cognitive Task Analysis, and contextual inquiry, as tools for comprehending how expert users assimilate information and make decisions [51]. As noticed by Islam et al. Co-creation and participatory design approaches can help tailor explanations to specific domains [53].

The interaction between humans and AI systems, especially adaptive models that customize explanations based on user profiles, holds substantial importance, too. Insights from social sciences and human behavioural studies have significant potential in the domains of xAI and human-centred AI. There is a call for greater integration between the Human–Computer Interaction (HCI) community and these emerging fields [51,53]. As AI systems increasingly incorporate post-hoc explainability, it becomes crucial to carefully assess the implications and second-order effects of these approaches. The content, modality, and purpose of information communicated through xAI elements should align with rigorous analysis of use cases and requirements from the outset [51]. Research indicates that the quality of explanations can influence trust and reliance. However, there is a noticeable lack of research on the impact of xAI from the user's perspective. These areas represent opportunities for future investigation [47,54]. Another point related to this is that as AI algorithms become increasingly complex, a critical question arises: do end-users, such as cybersecurity experts or healthcare professionals, need a comprehensive understanding of AI, or should they simply trust its predictive accuracy based on past performance? [60] Lastly, a number of researchers highlight the challenges posed by increasing AI complexity and the importance of tailoring explanations to individual users' diverse and evolving needs [59].

In addition to this, Evans et al. have postulated that safe and effective xAI must strike a balance between usability and the fidelity with which it represents AI decision-making processes. Consequently, two investigative avenues are proposed, the parallel and the orthogonal approaches. In the first one, explainability elements align with components of the AI decision-making process that closely match users' decision-making processes. This approach seeks to enhance user understanding by drawing parallels between human and AI reasoning [54]. On the other hand, with the orthogonal approach, explainability elements are based on AI decision-making components with minimal resemblance to human reasoning. By deliberately emphasizing the distinct nature of AI systems, this approach aims to mitigate potential negative effects. The latter approach is said to be setting up more realistic expectations of what AI is capable of doing. However, it may require additional user training and support [54]. Evans et al. admit that while the orthogonal approach may risk alienating users, it holds potential for creating synergy in human–AI cooperation, it acknowledges the unique aspects of machine intelligence that can contribute to a more sustainable xAI strategy [47,54].

4.2. Developing domain-specific xAI evaluation metrics

Another research direction heavily referenced by the domain experts that xAI needs is developing domain-specific evaluation metrics.

There has been an ongoing debate in the scientific community regarding how to evaluate and measure explainability. So far, plenty of methods have been proposed; still, the scientific consensus on which metrics to use has been far from being reached.

Some researchers have agreed that the existing evaluation metrics could be roughly divided into two types: human-centred [50] and computer-centred (technical) ones [49,65]. Yet, right beyond this point there is no agreement as to what the categories encompass. The absence of agreed-upon evaluation criteria in xAI poses a significant challenge. On top of that, evaluation methods often lack rigour and tend to be based solely on the views of computer scientists and AI engineers. However, to harness the benefits of xAI in various domains, including construction, it is essential to cultivate evaluation frameworks that include measurable outcomes aligned with the stakeholders' interests, goals, expectations, and demands within specific contexts [61].

Scientists notice that as the existing research often lacks a balance between functional evaluation and user evaluation and over-reliance on one over the other can lead to inaccurate insights drawn from overly simplified evaluation scenarios [57]. Thus, a more harmonious blend of these evaluation methods is essential to improve the quality of xAI research [59]. In this vein, Brasse et al. suggest three research directions. First, xAI approaches should frequently undergo human-grounded evaluation to account for human risks associated with novel xAI methods. This includes robust evaluations that incorporate human users. Second, the focus should shift towards application-grounded evaluation with real users in real settings. This approach assesses the utility, quality, and efficacy of xAI methods in practical, real-life scenarios. Lastly, novel evaluation strategies should be explored, combining functional and human-grounded evaluation. This approach allows for a robust comparison of xAI approaches while considering the intricate social aspects involved [59].

Another facet of this research direction is that while some general metrics exist, there is a pressing need for the research community to focus more on domain-specific metrics. Metrics should be tailored to the application, considering the unique challenges and goals within specific domains [55]. In the pursuit of advancing the field of xAI, it becomes evident that the development of domain-specific evaluation metrics is a crucial undertaking. Similarly, in the cybersecurity domain, developing proper evaluation metrics is critical. Conventional metrics may not be suitable due to unbalanced class distributions and the high-dimensional nature of cybersecurity data. Furthermore, evaluating the consequences of different types of errors is vital, as false positives and false negatives can have vastly different implications [62].

4.3. Fostering interdisciplinary research

A number of scientists underscore the necessity of collaboration with experts from diverse fields, such as social and behavioural science, philosophy, psychology, and cognitive science, to advance our understanding and capabilities in xAI [47].

This collaborative approach is motivated by the recognition that comprehensively investigating xAI requires a holistic perspective. This entails understanding xAI as a complex sociotechnical system with far-reaching implications for AI practices in both business and society. It transcends traditional disciplinary boundaries and necessitates a broad perspective [48].

xAI, by its very nature, spans multiple domains, intertwining people, information technology, and organizational contexts. To enhance our understanding, it is imperative to adopt a multidisciplinary approach. Insights from cognitive theories provide a foundation, but exploring xAI through a social lens is equally valuable. Additionally, theories from fields such as social sciences, management, and computer science should be considered to create a holistic evaluation framework [59].

In this context, xAI includes four primary dimensions: data explainability, model explainability, post-hoc explainability, and assessment of explanations. Within these dimensions, interdisciplinary overlaps emerge. While the ultimate goal remains consistent – to produce better explanations – the specific objectives may vary based on the users and contexts. For instance, designing xAI systems for AI novices requires human-centred interfaces, while addressing the needs of AI specialists demands alternative interpretability approaches. Therefore, considering various user groups becomes an additional dimension for aligning xAI goals across different research disciplines and integrating diverse research aims [57]. This again related to the fact that xAI, to be a successful paradigm, requires contributions of fields beyond computer science.

4.4. Understanding the need for context awareness in xAI

Numerous analysed studies identify building context awareness in explainable Artificial Intelligence as another crucial research direction. Similar to the emphasis on making explanations user-centred by considering the actual needs and capabilities of users, this approach highlights the importance of exploring methods to generate explanations that account for mission contexts, user roles, and targeted goals, regardless of the type of AI system. While prior research in this area has been largely conceptual, there is an increasing need for more comprehensive and practical implementations that consider broader contextual factors [64].

As related by Yang and Wei, the development of context-aware xAI is key, with a focus on generating explanations that take into account the mission context. This context encompasses the surrounding environment, various situations, and time-series datasets that influence AI system behaviour. Furthermore, mapping user roles, including end-users, domain experts, business managers, AI developers, and others, is essential. Lastly, context-aware xAI aims to align with specific goals, such as refining models, debugging system errors, detecting bias, and comprehending the AI learning process. Despite the conceptual foundation laid by previous studies, the pursuit of more generalized context-driven xAI systems and practical implementations emerges as a vital direction for future research [64].

In the quest to build context-aware xAI, domain-specific requirements play an essential role. These requirements necessitate a comprehensive understanding of the system's purpose, efficiency, and explainability. Additionally, considerations include the level of complexity in the desired explanations and the alignment of xAI solutions with the specific needs and objectives of the domain [56].

In the context of network cybersecurity, the research goal revolves around enhancing the reliability and accuracy of AI systems in interpreting and analysing data related to cybersecurity threats. This entails the development and evaluation of novel algorithms, techniques, and approaches for interpreting diverse data sources, including network traffic, logs, and user behaviour data. A key challenge in this domain is to develop machine interpretation techniques that can accurately identify and classify cybersecurity threats while minimizing the risk of false positives. Achieving this involves the analysis of contextual and background information related to the data source, alongside the utilization of machine learning algorithms to detect patterns and trends indicative of cybersecurity threats [62].

4.5. *Enhancing explainability through interactive and hybrid approaches*

In the subject literature, there is a recurring postulate to be found, regarding the enhancement of xAI through interactive explanations and hybrid explanation systems. These approaches offer new dimensions for building more human-centred and effective AI systems.

First and foremost, the concept of interactive explanations takes centre stage. Interactive explanations encompass diverse techniques such as conversation system interfaces, games, and the use of audio, visuals, and video. These approaches hold promise in creating genuinely human-centred explanations by identifying and addressing user requirements. They facilitate better collaboration between humans and AI, allowing for an iterative process that is crucial for the success of xAI systems. By incorporating theories and frameworks that enable ongoing interaction with users, interactive explanations pave the way for more effective xAI [64].

Similarly, hybrid explanations involve the fusion of heterogeneous knowledge from various sources, addressing challenges such as time-sensitive data, inconsistency, and uncertainty. In recent years, hybrid explanations have gained traction as an interesting and increasingly explored topic. This approach necessitates the development of criteria and strategies to establish a clear structure and consensus on what constitutes success and trustworthy explanations [64].

Interactive visual tools empower AI and data specialists to enhance model performance. These tools can also benefit novices in the field. Interactive methods enable users to assess the impact of their actions and adapt their queries to improve results. This section sheds light on the importance of human interactions in the development of xAI systems, emphasizing their potential to enhance user experiences and the utility of AI models [57].

4.6. *Achieving bias-free xAI*

The research findings align with previous observations, emphasizing that social and cognitive biases profoundly impact human interactions with xAI systems [54]. This includes the common tendency to anthropomorphize (x)AI systems, highlighting the need for xAI designs that are adept at recognizing, mitigating, and judiciously utilizing these biases and predispositions. To address the pervasive confirmation bias, characterized by individuals processing information in a manner that reinforces their existing beliefs, strategies are paramount. These include limiting explanations for sensitive features primarily for system development, and offering cognitive awareness training to developers and data scientists [58].

Biases in xAI can emerge from multiple sources, encompassing biased training data or subjective users' responses. Vigilance in identifying and rectifying these biases during the validation and verification stages of AI algorithms is essential [60].

In light of these observations, scientists call for recognizing, mitigating, and occasionally harnessing biases and predispositions that impact human interactions with xAI.

4.7. *Striving for model and data transparency*

Another research direction to be identified is ensuring the transparency in xAI, of both the training data and the models themselves. It highlights the role of trust in AI systems and the critical need to understand the sources of model training data, particularly concerning the credibility and diversity of expert annotators [54].

In the realm of network cybersecurity, transparency takes centre stage as a crucial aspect. xAI methods in this domain must be designed to offer comprehensible explanations of their decision-making processes. However, achieving both effectiveness and transparency poses a significant challenge, notably concerning the trade-off between accuracy and transparency. Possible solutions include developing xAI methods that can provide varying levels of transparency based on the decision's risk level and employing transparency-enhancing techniques like sensitivity analysis. This challenge necessitates innovative approaches to ensure the wide adoption of xAI in the cybersecurity domain [62].

Similar postulates have been made concerning other domains, such as biomedical engineering and healthcare informatics; where scientists emphasize the critical role of explainable Artificial Intelligence in enhancing transparency. In their context, the objective is to make individuals aware of how AI prototypes operate. Human–computer interaction technology plays a vital role in accelerating the explainability of AI models, ushering in a new era of transparency [56].

4.8. *Towards an ethical code for practical AI applications*

The topic of ethical issues in artificial intelligence (AI) is of paramount importance and encompasses a wide range of considerations. Addressing the ethical challenges posed by AI technologies necessitates interdisciplinary collaboration among experts in AI, ethics, law, and related fields. AI applications give rise to a wide array of ethical concerns encompassing areas such as bias, fairness, privacy, and security. These issues demand various technical approaches for resolution. For instance, addressing bias in AI models may require data pre-processing, algorithmic modifications, or human oversight. Ensuring the robustness and reliability of AI systems may involve techniques like adversarial training, uncertainty quantification, and fault-tolerant design. The complexity and diversity of ethical issues in AI span different application domains and involve various stakeholders. For example, medical AI systems introduce unique ethical considerations related to patient safety, informed consent, and privacy, necessitating distinct technical and legal frameworks compared to domains like finance or transportation. In this context, it becomes evident that the ethical dimensions of AI are both critically important and extensive, underscoring the need for a comprehensive and collaborative approach to address them effectively.

Antoniadi et al. remark on the ethical dimensions of Explainable Artificial Intelligence while emphasizing the importance of considering ethics, fairness, safety implications, and the cognitive capabilities of the audience when determining the appropriate type of explanations [47]. Efforts to formalize the study of ethics in practical AI applications have led to the development of frameworks and guidelines. These include initiatives like the IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems and the EU Ethics Guidelines for Trustworthy AI [57,62].

4.9. Privacy-preserving explainability

The need to research privacy-preserving xAI is paramount in our increasingly data-driven world. As AI systems become more integrated into our lives, safeguarding sensitive information and ensuring user privacy while maintaining transparency and interpretability are essential to build trust and compliance with privacy regulations. Research in this domain is crucial to strike a balance between AI's capabilities and the protection of individuals' privacy rights, fostering responsible and ethical AI adoption [57].

4.10. Empowering cybersecurity with natural language processing

The field of Natural Language Processing (NLP) has evolved from the intersection of linguistics, computer science, and AI. It focuses on the interaction between computers and human language, encompassing tasks such as speech recognition, natural language understanding, and natural language generation. NLP's capabilities span from extracting information and insights from unstructured data to organizing and categorizing data units. In the context of network cybersecurity, NLP plays a crucial role in enhancing the interpretability and transparency of AI systems. By leveraging NLP techniques, cybersecurity experts can better understand AI-driven threat detection and response, leading to more effective security measures. Collaboration between NLP research and the Human–Computer Interaction (HCI) community can further enhance human–system interactions in cybersecurity, contributing to a safer and more secure digital environment [52].

4.11. Handling in uncertainty xAI for network cybersecurity

In the context of network cybersecurity, effectively addressing uncertainty becomes a paramount concern. The dynamic and ever-evolving nature of this domain necessitates the development of robust xAI methods adept at handling uncertainty. A multifaceted approach can be explored, encompassing the utilization of probabilistic models such as Bayesian networks, and the integration of robust optimization techniques to enhance the reliability of AI systems. Emphasizing the incorporation of uncertainty awareness into AI solutions is strongly advocated, as it holds the potential to not only bolster the robustness but also ensure the dependability of xAI deployments within this critical domain [46,62].

4.12. Ensuring reproducibility

Reproducibility is a fundamental aspect of xAI research, enhancing trust in algorithms and facilitating comparisons between different studies. To ensure reproducibility, Hulsen recommends that algorithms, including scripts and underlying data, should be made accessible for reuse whenever possible. This approach enables the replication of results, ultimately increasing confidence in the algorithm's performance [60].

In certain domains, such as xAI models based on electronic health records, the importance of research reproducibility may not receive adequate attention. To address this issue, researchers should utilize open data sources, clearly describe the methodology and infrastructure used, and share their code with the research community. Additionally, publication venues should establish reproducibility standards to be followed by authors as part of the publication process. These measures aim to enhance the reproducibility of xAI research and foster a culture of transparency and trust [63].

4.13. Balancing correlation and fusion of information

Correlating and fusing information from multiple sources can significantly enhance the interpretability and transparency of machine learning and deep learning models. This approach often leads to more accurate inferences than analysing a single dataset. However, it is essential to balance this enrichment of explainability with privacy concerns, as the fusion of data can compromise data privacy. Therefore, AI studies, especially in fields like construction and computer vision, should explore data and information fusion to enhance interpretability while safeguarding privacy [61].

Security and privacy are paramount in xAI. Popular xAI methods like SHAP may be computationally expensive when executed iteratively. Research should focus on developing energy-efficient xAI methods while also implementing interpreters that filter sensitive information to avoid privacy breaches and ensure compliance with intellectual property laws. These measures are essential in conveying explanations to stakeholders without compromising security or privacy [55].

4.14. Combining explanatory approaches

Lastly, to enrich user comprehension, it is advantageous to inquire the synergistic fusion of diverse explanatory approaches that complement one another. For instance, the seamless integration of local and global explanations can yield a holistic perspective on AI-driven decisions, enhancing the overall clarity of system outputs. This comprehensive approach to combining various explanation methods not only empowers xAI systems to deliver more robust insights but also equips users with a more informative and nuanced understanding of complex AI processes [59].

4.15. The future challenges of xAI in cybersecurity

Although the primary objective of this paper was not to explore future challenges directly, this comprehensive review of the literature and the systematic analysis of current trends and gaps have enabled the authors to identify several emergent challenges related to xAI in the cybersecurity domain. These challenges highlight the complexity and the multifaceted nature of integrating explainability into cybersecurity practices effectively. Some of the identified challenges has been briefly discussed below.

One of the foremost challenges is the lack of standardization in defining what constitutes explainability within cybersecurity applications. This ambiguity complicates the development, evaluation, and comparison of xAI systems [66]. There also exists an inherent trade-off between the complexity (and thus, performance) of AI models and their interpretability. Achieving optimal balance remains a significant challenge, as higher complexity often leads to reduced interpretability, impacting users' trust and understanding of the system's decisions [67].

In addition to this, crafting explanations that are meaningful and useful across different levels of user expertise presents another challenge. Useful explanations must be adaptable to cater to diverse users, from security experts to laypersons, ensuring that the xAI system's decisions are accessible and understandable to all [68]. Another very important challenge are the considerations related to privacy and security. They are connected with the fact that integrating explainability into cybersecurity solutions raises concerns about privacy and security, particularly in how detailed explanations might inadvertently expose sensitive information or system vulnerabilities [57,66].

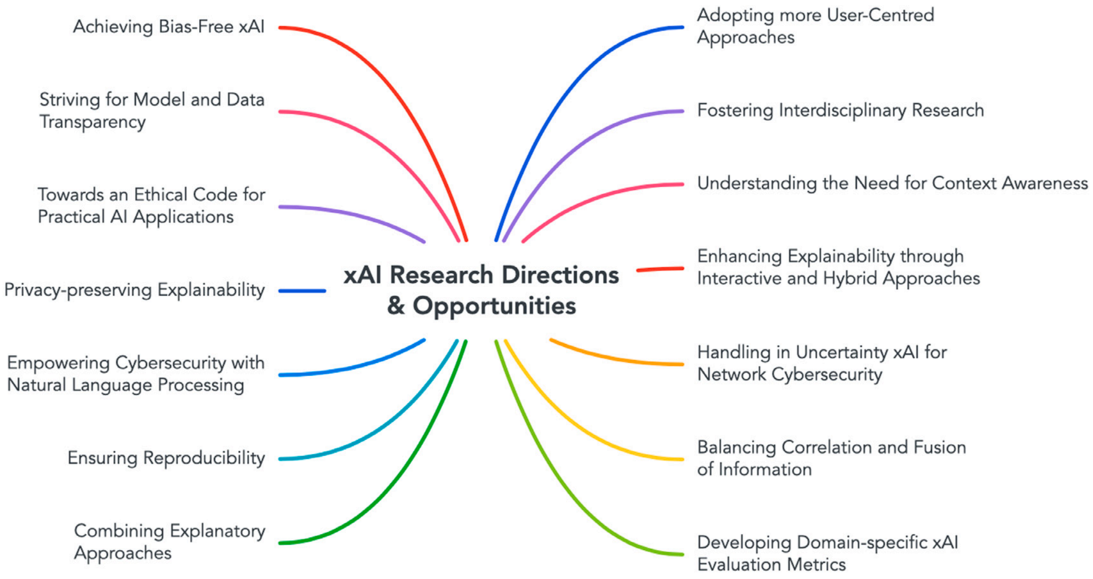

Achieving Bias-Free xAI
Striving for Model and Data Transparency
Towards an Ethical Code for Practical AI Applications
Privacy-preserving Explainability
Empowering Cybersecurity with Natural Language Processing
Ensuring Reproducibility
Combining Explanatory Approaches
xAI Research Directions & Opportunities
Adopting more User-Centred Approaches
Fostering Interdisciplinary Research
Understanding the Need for Context Awareness
Enhancing Explainability through Interactive and Hybrid Approaches
Handling in Uncertainty xAI for Network Cybersecurity
Balancing Correlation and Fusion of Information
Developing Domain-specific xAI Evaluation Metrics

Fig. 10. The research directions and opportunities for xAI identified in the study presented.

Researchers have also pointed out that the absence of robust evaluation metrics for explainability further complicates the assessment of xAI systems. There is a pressing need for metrics that can quantify the effectiveness of explanations in a manner that is both comprehensive and domain-specific [63].

Lastly, it has been noted that the use of explanation methods may present a kind of paradox. Their simplicity and universality stand as strengths, yet concerns about their robustness and potential as an attack surface cannot be overlooked. As a result, the quest for model transparency might inadvertently expose sensitive details, offering adversaries new vectors for exploitation. This duality underscores the imperative for developing robust, transparent xAI models that safeguard against such vulnerabilities without compromising on the clarity and utility of explanations [69]

It is important to note that the challenges highlighted herein represent only a subset of the many issues facing the integration of explainable AI into cybersecurity. The rapid pace of technological advancements and the evolving nature of cyberthreats ensure that this field will confront a continuously expanding set of challenges. The depth and breadth of these challenges call for ongoing, collaborative research efforts that draw on a wide range of expertise and perspectives.

5. Conclusions

The summary of this study's findings has been presented in Fig. 10.

This paper has explored a spectrum of critical research directions within the realm of explainable Artificial Intelligence, with a particular focus on its application in Network Intrusion Detection Systems. The authors have studied the imperative need for user-centred explanations, the challenges posed by increasing AI complexity, the significance of context-aware xAI, and the delicate balance between data fusion and privacy preservation. Furthermore, discussions have encompassed the vital role of ethics, transparency, and interdisciplinary collaboration in shaping the future of xAI. The less-discussed, more obscure ideas were mentioned as well.

These research directions collectively underscore the dynamic nature of xAI, reflecting its continuous evolution to meet the demands of our data-driven world. By placing human needs and ethical considerations at the forefront, the researchers pave the way for AI systems that inspire trust, enhance transparency, and contribute to a more interpretable and accountable AI landscape.

As we move forward, it is hoped that these research directions will guide scholars, practitioners, and policymakers in their pursuit of responsible and impactful xAI solutions. By addressing the challenges outlined herein and embracing the opportunities they present, a new era of AI can be ushered in—one that not only empowers individuals and organizations but also aligns seamlessly with the desired values and expectations. Ideally, good xAI solutions will provide comprehension not only for experts but also common, regular users. The journey towards xAI excellence continues, and the path is illuminated by the insights and endeavours shared within these research directions.

CRediT authorship contribution statement

Marek Pawlicki: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Aleksandra Pawlicka:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Rafał Kozik:** Writing – original draft, Visualization, Validation, Methodology, Investigation. **Michał Choraś:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marek Pawlicki reports financial support was provided by Horizon 2020 and by Horizon Europe. Aleksandra Pawlicka reports financial support was provided by Horizon 2020 and by Horizon Europe. Rafał Kozik reports financial support was provided by Horizon 2020 and by Horizon Europe. Michał Choraś reports financial support was provided by Horizon 2020 and by Horizon Europe.

Data availability

No data was used for the research described in the article.

Acknowledgements

The work described in this paper is performed in the H2020 project STARLIGHT ('Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats'). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101021797. This work is also funded under the ULTIMATE project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070162.

References

[1] L. Hernández-Álvarez, J.J. Bullón Pérez, F.K. Batista, A. Queiruga-Dios, Security threats and cryptographic protocols for medical wearables, Mathematics 10 (6) (2022) 886, http://dx.doi.org/10.3390/math10060886, URL https://www.mdpi.com/2227-7390/10/6/886.

[2] M. Pawlicki, R. Kozik, M. Choraś, A survey on neural networks for (cyber-) security and (cyber-) security of neural networks, Neurocomputing 500 (2022) 1075–1087, http://dx.doi.org/10.1016/j.neucom.2022.06.002.

[3] S. Wang, Q. Zhang, Y. He, Z. Cui, Z. Guo, K. Han, D.-S. Huang, DLoopCaller: A deep learning approach for predicting genome-wide chromatin loops by integrating accessible chromatin landscapes, in: F. Ay (Ed.), PLoS Comput. Biol. 18 (10) (2022) e1010572, http://dx.doi.org/10.1371/journal.pcbi.1010572, URL https://dx.plos.org/10.1371/journal.pcbi.1010572.

[4] Y. He, Z. Shen, Q. Zhang, S. Wang, D.-S. Huang, A survey on deep learning in DNA/RNA motif mining, Brief. Bioinform. 22 (4) (2021) http://dx.doi.org/10.1093/bib/bbaa229, URL https://academic.oup.com/bib/article/doi/10.1093/bib/bbaa229/5916939.

[5] M. Choraś, M. Pawlicki, D. Puchalski, R. Kozik, in: V.V. Krzhizhanovskaya, G. Závodszky, M.H. Lees, J.J. Dongarra, P.M.A. Sloot, S. Brissos, J.a. Teixeira (Eds.), Machine Learning – The Results Are Not the only Thing that Matters! What About Security, Explainability and Fairness? BT - Computational Science – ICCS 2020, Springer International Publishing, Cham, 2020, pp. 615–628.

[6] F. Yan, S. Wen, S. Nepal, C. Paris, Y. Xiang, Explainable machine learning in cybersecurity: A survey, Int. J. Intell. Syst. 37 (12) (2022) 12305–12334, http://dx.doi.org/10.1002/int.23088.

[7] N. Capuano, G. Fenza, V. Loia, C. Stanzione, Explainable artificial intelligence in CyberSecurity: A survey, IEEE Access 10 (2022) 93575–93600, http://dx.doi.org/10.1109/ACCESS.2022.3204171.

[8] C.I. Nwakanma, L.A.C. Ahakonye, J.N. Njoku, J.C. Odirichukwu, S.A. Okolie, C. Uzondu, C.C. Ndubuisi Nweke, D.-S. Kim, Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review, Appl. Sci. 13 (3) (2023) 1252, http://dx.doi.org/10.3390/app13031252.

[9] D.K. Sharma, J. Mishra, A. Singh, R. Govil, G. Srivastava, J.C.-W. Lin, Explainable artificial intelligence for cybersecurity, Comput. Electr. Eng. 103 (2022) 108356, http://dx.doi.org/10.1016/j.compeleceng.2022.108356.

[10] Ł. Wawrowski, M. Michalak, A. Białas, R. Kurianowicz, M. Sikora, M. Uchroński, A. Kajzer, Detecting anomalies and attacks in network traffic monitoring with classification methods and XAI-based explainability, Procedia Comput. Sci. 192 (2021) 2259–2268, http://dx.doi.org/10.1016/j.procs.2021.08.239, URL https://linkinghub.elsevier.com/retrieve/pii/S1877050921017361.

[11] S. Gulmez, A. Gorgulu Kakisim, I. Sogukpinar, XRan: Explainable deep learning-based ransomware detection using dynamic analysis, Comput. Secur. 139 (2024) 103703, http://dx.doi.org/10.1016/j.cose.2024.103703, URL https://linkinghub.elsevier.com/retrieve/pii/S016740482400004X.

[12] F. Greco, G. Desolda, A. Esposito, Explaining phishing attacks: An XAI approach to enhance user awareness and trust, in: ITASEC 2023: The Italian Conference on CyberSecurity, May 03–05, 2023, Bari, Italy, 2023, p. ..

[13] C. Meske, E. Bunde, J. Schneider, M. Gersch, Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities, Inf. Syst. Manage. 39 (1) (2022) 53–63, http://dx.doi.org/10.1080/10580530.2020.1849465, URL https://www.tandfonline.com/doi/full/10.1080/10580530.2020.1849465.

[14] A. Mathew, Explainable AI for intelligence analysis, Int. J. Eng. Res. Technol. (IJERT) 12 (02) (2023).

[15] S.S. Gill, M. Xu, C. Ottaviani, P. Patros, R. Bahsoon, A. Shaghaghi, M. Golec, V. Stankovski, H. Wu, A. Abraham, M. Singh, H. Mehta, S.K. Ghosh, T. Baker, A.K. Parlikad, H. Lutfiyya, S.S. Kanhere, R. Sakellariou, S. Dustdar, O. Rana, I. Brandic, S. Uhlig, AI for next generation computing: Emerging trends and future directions, Internet Things 19 (2022) 100514, http://dx.doi.org/10.1016/j.iot.2022.100514, URL https://linkinghub.elsevier.com/retrieve/pii/S254266052200018X.

[16] G.A. Vouros, Explainable deep reinforcement learning: state of the art and challenges, ACM Comput. Surv. 55 (5) (2022) 1–39.

[17] M. Ribeiro, S. Sing, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAAI-18), New Orleans, Louisiana, 2018.

[18] C. Molnar, Interpretable Machine Learning (Second Edition) A Guide for Making Black Box Models Explainable, Leanpub, 2022, URL https://leanpub.com/interpretable-machine-learning.

[19] W. Kurek, M. Pawlicki, A. Pawlicka, R. Kozik, M. Choraś, Explainable artificial intelligence 101: Techniques, applications and challenges, in: International Conference on Intelligent Computing, 2023, pp. 310–318.

[20] R.K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[21] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.

[22] A.M. Roth, J. Liang, D. Manocha, XAI-N: Sensor-based robot navigation using expert policies and decision trees, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2021, pp. 2053–2060.

[23] N. Schaaf, M. Huber, J. Maucher, Enhancing decision tree based interpretation of deep neural networks through l1-orthogonal regularization, in: 2019 18th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2019, pp. 42–49.

[24] B. Mahbooba, M. Timilsina, R. Sahal, M. Serrano, Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model, Complexity 2021 (2021) 1–11.

[25] M. Szczepański, M. Choraś, M. Pawlicki, R. Kozik, Achieving explainability of intrusion detection system by hybrid oracle-explainer approach, in: 2020 International Joint Conference on Neural Networks, IJCNN, 2020, pp. 1–8, http://dx.doi.org/10.1109/IJCNN48605.2020.9207199.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[27] C. Cambra Baseca, S. Sendra, J. Lloret, J. Tomas, A smart decision system for digital farming, Agronomy 9 (5) (2019) 216.

[28] B. Reddy, R. Fields, From past to present: a comprehensive technical review of rule-based expert systems from 1980–2021, in: Proceedings of the 2022 ACM Southeast Conference, 2022, pp. 167–172.

[29] A. Ambhaikar, A survey on health care and expert system, Math. Stat. Eng. Appl. 72 (1) (2023) 451–461.

[30] S. Burkhardt, J. Brugger, N. Wagner, Z. Ahmadi, K. Kersting, S. Kramer, Rule extraction from binary neural networks with convolutional rules for model validation, Front. Artif. Intell. 4 (2021) 642263.

[31] K. Bahani, M. Moujabbir, M. Ramdani, An accurate fuzzy rule-based classification systems for heart disease diagnosis, Sci. Afr. 14 (2021) e01019.

[32] J.H. Friedman, B.E. Popescu, Predictive learning via rule ensembles, Ann. Appl. Stat. (2008) 916–954.

[33] C. Luo, S. Li, Q. Zhao, Q. Ou, W. Huang, G. Ruan, S. Liang, L. Liu, Y. Zhang, H. Li, RuleFit-based nomogram using inflammatory indicators for predicting survival in nasopharyngeal carcinoma, a Bi-Center study, J. Inflamm. Res. (2022) 4803–4815.

[34] J. Grus, Data Science from Scratch: First Principles with Python, first ed., O'Reilly Media, Inc., 2015.

[35] A.G. Baydin, B.A. Pearlmutter, A.A. Radul, J.M. Siskind, Automatic differentiation in machine learning: a survey, J. Marchine Learn. Res. 18 (2018) 1–43.

[36] J. Han, M. Kamber, J. Pei, Data Mining Concepts and Techniques, third ed., University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University, 2012.

[37] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (10) (2012) 78–87.

[38] S. Sharma, J. Henderson, J. Ghosh, Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models, 2019, arXiv preprint arXiv:1905.07857.

[39] J. Henderson, S. Sharma, A. Gee, V. Alexiev, S. Draper, C. Marin, Y. Hinojosa, C. Draper, M. Perng, L. Aguirre, et al., Certifai: a toolkit for building trust in AI systems, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 5249–5251.

[40] C. Molnar, Interpretable machine learning, Lulu. com, 2020.

[41] K.S. Gurumoorthy, A. Dhurandhar, G. Cecchi, C. Aggarwal, Efficient data representation by selecting prototypes with importance weights, in: 2019 IEEE International Conference on Data Mining, ICDM, IEEE, 2019, pp. 260–269.

[42] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Inf. Softw. Technol. 64 (2015) 1–18, http://dx.doi.org/10.1016/j.infsof.2015.03.007, URL https://linkinghub.elsevier.com/retrieve/pii/S0950584915000646.

[43] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ (2021) n71, http://dx.doi.org/10.1136/bmj.n71, URL https://www.bmj.com/lookup/doi/10.1136/bmj.n71.

[44] J. Yensen, PICO search strategies, Online J. Nurs. Inform. 17 (3) (2013) doi:https://www.researchgate.net/deref/http%3A%2F%2Fojni.org%2Fissues%2F%3Fp%3D2860.

[45] M. Ashouri, P. Davidsson, R. Spalazzese, Quality attributes in edge computing for the internet of things: A systematic mapping study, Internet Things 13 (2021) 100346, http://dx.doi.org/10.1016/j.iot.2020.100346, URL https://linkinghub.elsevier.com/retrieve/pii/S2542660520301773.

[46] M. Pocevičiūtė, G. Eilertsen, C. Lundström, Survey of XAI in digital pathology, 2020, http://dx.doi.org/10.1007/978-3-030-50402-1_4, arXiv:2008.06353 URL http://arxiv.org/abs/2008.06353.

[47] A.M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B.A. Becker, C. Mooney, Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review, Appl. Sci. 11 (11) (2021) 5088, http://dx.doi.org/10.3390/app11115088, URL https://www.mdpi.com/2076-3417/11/11/5088.

[48] J. Gerlings, A. Shollo, I. Constantiou, Reviewing the need for explainable artificial intelligence (xAI), 2021, http://dx.doi.org/10.24251/HICSS.2021.156, URL http://hdl.handle.net/10125/70768.

[49] A. Hanif, X. Zhang, S. Wood, A survey on explainable artificial intelligence techniques and challenges, in: 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop, EDOCW, IEEE, 2021, pp. 81–89, http://dx.doi.org/10.1109/EDOCW52865.2021.00036, arXiv:EDOCW52865.2021.0003 URL https://ieeexplore.ieee.org/document/9626294/.

[50] Q.V. Liao, K.R. Varshney, Human-centered explainable AI (XAI): From algorithms to user experiences, 2021, arXiv:2110.10790 URL http://arxiv.org/abs/2110.10790.

[51] A. Kotriwala, B. Kloepper, M. Dix, G. Gopalakrishnan, D. Ziobro, A. Potschka, XAI for operations in the process industry – applications, theses, and research directions, in: F. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), Palo Alto, 2021.

[52] J.N. Paredes, J.C.L. Teze, G.I. Simari, M.V. Martinez, On the Importance of Domain-specific Explanations in AI-based Cybersecurity Systems, Technical Report, 2021, arXiv:2108.02006 URL http://arxiv.org/abs/2108.02006 doi:2108.02006v1.

[53] M.R. Islam, M.U. Ahmed, S. Barua, S. Begum, A systematic review of explainable artificial intelligence in terms of different application domains and tasks, Appl. Sci. 12 (3) (2022) 1353, http://dx.doi.org/10.3390/app12031353, URL https://www.mdpi.com/2076-3417/12/3/1353.

[54] T. Evans, C.O. Retzlaff, C. Geiß ler, M. Kargl, M. Plass, H. Müller, T.-R. Kiehl, N. Zerbe, A. Holzinger, The explainability paradox: Challenges for xAI in digital pathology, Future Gener. Comput. Syst. 133 (2022) 281–296, http://dx.doi.org/10.1016/j.future.2022.03.009, URL https://linkinghub.elsevier.com/retrieve/pii/S0167739X22000838.

[55] T. Senevirathna, V.H. La, S. Marchal, B. Siniarski, M. Liyanage, S. Wang, A survey on XAI for beyond 5G security: Technical aspects, use cases, challenges and research directions, 2022, arXiv:2204.12822 URL http://arxiv.org/abs/2204.12822.

[56] P.N. Srinivasu, N. Sandhya, R.H. Jhaveri, R. Raut, From blackbox to explainable AI in healthcare: Existing tools and case studies, in: S. Hakak (Ed.), Mob. Inf. Syst. 2022 (2022) 1–20, http://dx.doi.org/10.1155/2022/8167821, URL https://www.hindawi.com/journals/misy/2022/8167821/.

[57] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805, http://dx.doi.org/10.1016/j.inffus.2023.101805, URL https://linkinghub.elsevier.com/retrieve/pii/S1566253523001148.

[58] K. Bauer, M. von Zahn, O. Hinz, Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing, Inf. Syst. Res. (2023) http://dx.doi.org/10.1287/isre.2023.1199, URL http://pubsonline.informs.org/doi/10.1287/isre.2023.1199.

[59] J. Brasse, H.R. Broder, M. Förster, M. Klier, I. Sigler, Explainable artificial intelligence in information systems: A review of the status quo and future research directions, Electron. Mark. 33 (1) (2023) 26, http://dx.doi.org/10.1007/s12525-023-00644-5, URL https://link.springer.com/10.1007/s12525-023-00644-5.

[60] T. Hulsen, Explainable artificial intelligence (XAI): Concepts and challenges in healthcare, AI 4 (3) (2023) 652–666, http://dx.doi.org/10.3390/ai4030034, URL https://www.mdpi.com/2673-2688/4/3/34.

[61] P.E. Love, W. Fang, J. Matthews, S. Porter, H. Luo, L. Ding, Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction, Adv. Eng. Inform. 57 (2023) 102024, http://dx.doi.org/10.1016/j.aei.2023.102024, URL https://linkinghub.elsevier.com/retrieve/pii/S1474034623001520.

[62] G. Rjoub, J. Bentahar, O.A. Wahab, R. Mizouni, A. Song, R. Cohen, H. Otrok, A. Mourad, A survey on explainable artificial intelligence for cybersecurity, 2023, http://dx.doi.org/10.1109/TNSM.2023.3282740, arXiv:2303.12942 URL http://arxiv.org/abs/2303.12942.

[63] W. Saeed, C. Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowl.-Based Syst. 263 (2023) 110273, http://dx.doi.org/10.1016/j.knosys.2023.110273, URL https://linkinghub.elsevier.com/retrieve/pii/S0950705123000230.

[64] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M.B. Amin, B. Kang, Survey on explainable AI: From approaches, limitations and applications aspects, Hum-Cent. Intell. Syst. 3 (3) (2023) 161–188, http://dx.doi.org/10.1007/s44230-023-00038-y, URL https://link.springer.com/10.1007/s44230-023-00038-y.

[65] A. Pawlicka, M. Pawlicki, R. Kozik, W. Kurek, M. Choraś, How explainable is explainability? Towards better metrics for explainable AI, 2024, pp. 685–695, http://dx.doi.org/10.1007/978-3-031-44721-1_52, URL https://link.springer.com/10.1007/978-3-031-44721-1_52.

[66] W. Ding, M. Abdel-Basset, H. Hawash, A.M. Ali, Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey, Inform. Sci. 615 (2022) 238–292, http://dx.doi.org/10.1016/j.ins.2022.10.013, URL https://linkinghub.elsevier.com/retrieve/pii/S002002552201132X.

[67] A. Nadeem, D. Vos, C. Cao, L. Pajola, S. Dieck, R. Baumgartner, S. Verwer, Sok: Explainable machine learning for computer security applications, 2022, arXiv:2208.10605 URL http://arxiv.org/abs/2208.10605.

[68] K.P. Kalyanathaya, K.P. K., A literature review and research agenda on explainable artificial intelligence (XAI), Int. J. Appl. Eng. Manage. Lett. 6 (1) (2022) 43–59, http://dx.doi.org/10.47992/IJAEML.2581.7000.0119, URL https://srinivaspublication.com/journal/index.php/ijaeml/article/view/1576/689.

[69] R. Kozik, M. Ficco, A. Pawlicka, M. Pawlicki, F. Palmieri, M. Choraś, When explainability turns into a threat - using xAI to fool a fake news detection method, Comput. Secur. 137 (2024) 103599, http://dx.doi.org/10.1016/j.cose.2023.103599, URL https://linkinghub.elsevier.com/retrieve/pii/S0167404823005096.