Trustworthy Machine Learning Approaches for Cyberattack Detection: A Review

Blessing Guembe[1,3(✉)], Ambrose Azeta[2,3], Sanjay Misra[2,3], and Ravin Ahuja[3,4]

[1] Department of Computer and Information Sciences, Covenant University, Ota, Ogun, Nigeria
blessingodede@gmail.com

[2] Department of Computer Science, Namibia University of Science and Technology, Windhoek, Namibia
aazeta@nust.na

[3] Department of Computer Science and Communication, Ostfold University College, Halden, Norway

[4] Delhi Skills and Entrepreneurship University, Delhi, India

Abstract. In recent years, machine learning techniques have been utilized in sensitive areas such as health, medical diagnosis, facial recognition, cybersecurity, etc. With this exponential growth comes potential large-scale ethical, safety, and social ramifications. With this enhanced ubiquity and sensitivity, concerns about ethics, trust, transparency, and accountability inevitably arise. Given the threat of sophisticated cyberattacks, it's critical to establish cybersecurity trustworthy concepts and to develop methodologies and concepts for a wide range of explainable machine cybersecurity models that will assure reliable threat identification and detection, more research is needed. This survey examines a variety of explainable machine learning techniques that can be used to implement a reliable cybersecurity infrastructure in the cybersecurity domain. The main aim of this study is to execute an in-depth review and identification of existing explainable machine learning algorithms for cyberattack detection. This study employed the seven-step survey model to determine the research domain, implement search queries, and compile all retrieved articles from digital databases. This research looks at the literature on trustworthy machine learning algorithms for detecting cyberattacks. An extensive search of electronic databases such as ArXiv, Semantic Scholar, IEEE Xplore, Wiley Library, Scopus, Google Scholar, ACM, and Springer was carried out to find relevant literature in the subject domain. From 2016 to 2022, this study looked at white papers, conference papers, and journals. Only 25 research papers were chosen for this research paper describing trustworthy cybersecurity and explainable AI cybersecurity after we retrieved 800 articles from web databases. The study reveals that the decision tree technique outperforms other state-of-the-art machine learning models in terms of transparency and interpretability. Finally, this research suggests that incorporating explainable into machine learning cybersecurity models will help uncover the root causes of defensive failures, making it easier for cybersecurity experts to enhance both cybersecurity infrastructures and development, rather than just model results, policy, and management.

Keywords: Machine learning · Trustworthiness · Trustworthy cybersecurity

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13381, pp. 265–278, 2022.
https://doi.org/10.1007/978-3-031-10548-7_20

1 Introduction

Machine learning has seen considerable industrial use in recent years, with machine learning systems displaying superhuman ability in a variety of activities. However, increased model complexity has often resulted in increased performance, making such systems into black-box techniques, and generating ambiguity about how they function and, eventually, how they draw conclusions [1]. This inconsistency has made it difficult for machine learning techniques to be implemented in sensitive yet vital fields, such as cybersecurity and medical diagnosis, where their benefit may be enormous. Since machine learning algorithms are being more widely used in practical systems, any discrimination or unfairness propagated by those models could have a direct impact on human lives [2]. Machine learning trustworthiness and fairness is a sub-domain of machine learning interpretability that evaluates machine learning algorithms in terms of fairness, discrimination, accountability, and discrimination. It is difficult to trust systems whose outcomes are difficult to explain, especially in domains like healthcare and cybersecurity, where ethical and equitable questions have inevitably developed. The need for trustworthy, fair, robust, high-performing models for real-world applications such as cybersecurity defense models prompted the revival of explainable machine learning, a field devoted to the understanding and interpretation of machine behavior that had fallen out of favor in recent years due to most research focusing on the predictive power of algorithms rather than the understanding of machine behavior [2].

Cybersecurity experts are constantly developing sophisticated cybersecurity infrastructures that depend on a variety of components, such as technology, processes, and people to scale and operate efficiently [3]. In the cybersecurity domain, where so many decisions are critical to organizational breakdown, a lack of interpretability in machine learning models might decrease trust in those systems. The ability to Incorporate trustworthiness in cybersecurity infrastructure is completely based on the relationship between these components, consisting of people and technology, organizations and people, and groups and people. These trusting components can deploy cybersecurity infrastructures that organizations, businesses, governments, and individuals can rely on to detect, prevent, and protect cyberspace against sophisticated cyberattacks [4, 5].

The lack of explainability in internal data representations and the decision-making process of machine learning systems are one of the most challenging problems in increasing the acceptance of machine learning systems, especially in critical domains such as cybersecurity, finance, medicine, etc. For instance, people tend to blindly accept the outcomes of machine learning models in critical domains such as cybersecurity, object recognition, chatbots, and games, therefore permitting the possibility of false positives, but this method of accepting machine learning and deep learning outcomes should not be tolerated when it comes to sensitive decision-making domains such as medicine, finance, or cybersecurity, where important assets or human lives are priority [6]. Given the high demand for ethical AI, individuals are generally hesitant to accept technology that is not explicitly explainable, intractable, or trustworthy. It's natural to believe that focusing solely on model performance will make the models increasingly inscrutable. As a result, model performance and transparency are mutually exclusive. Improvements in a model's explainability, understanding, and trustworthiness, on the other hand, might lead to the correction of false positives and deficiencies [7].

When building a machine learning model, it is important to consider explainability and interpretability as a design driver to ensure impartiality in decision making (i.e., to detect anomalous behavior and automatically correct for bias in the training dataset, provide robustness by highlighting potential anomalies and adversaries that could change the prediction and decision logic, and ensure that only relevant features infer the result) [7, 8]. In response, it is important to make machine learning cybersecurity models to be trustworthy. Explainable machine learning enables the development and deployment of moral philosophical and moral AI models that are completed based on philosophical and ethical solutions [9]. Given the sophistication and unpredictable nature of cybersecurity threats, understanding the reasoning driving their outputs has become increasingly desirable. Explainability approaches in machine learning focus on giving users a comparable degree of knowledge, allowing them to make knowledgeable decisions about whether to adopt a model decision. In practice, explainability techniques usually yield a new interpretable model that basically duplicates the behavior of a black-box model, or they offer some statistical information or a graphic to show roughly how the model uses inputs. When interpretations are direct accounts of model behavior and structure, explanations are comprehensible generalizations [2].

Despite the exponential growth of trustworthy machine learning algorithms, there is still an immature and undeveloped area, hampered by a lack of formality and a lack of widely accepted terminology. As a result, even though many reliable machine learning approaches and scholarly articles have been produced, they hardly constitute a vital part of trustworthy machine learning algorithms for combating cyberattacks. Hence, the embedding of explainable machine learning in cybersecurity models will eliminate the bias attached to traditional machine learning black-box models. The purpose of this review is to act as a reference source for both researchers and industry experts by offering an in-depth review of existing methods, as well as a synopsis of the best use cases for reviewed techniques and connections to their strengths and weakness in implementing a trustworthy machine learning cyberattack detection model. The main contribution of this study is to provide an in-depth review and identification of existing explainable machine learning algorithms for cyberattack detection.

In this research, we first discuss the background of explainable and trustworthy machine learning, followed by a description of the research questions and their justifications. We also discussed the search strategy and criteria for selecting the relevant articles to archive the research objective. The responses to the study questions were thoroughly reviewed, and finally, the future research directions were also discussed.

1.1 Research Methodology

In this study, we adopted the seven steps survey model. Figure 1 represents a synopsis of designing the seven-step survey model utilized in this study. The purpose of this research was to find a collection of studies on trustworthy machine learning models in cybersecurity, and the impact of explainable machine learning in the cybersecurity domain. To archive this, the study identifies relevant articles in the explainable machine learning domain with respect to trustworthy cybersecurity. The authors formulated the research questions presented in this study and explain the concept behind the research questions.

1.2 Identification of Studies and Search Criteria

We obtained the adopted articles from the following databases: ArXiv, Semantic Scholar, IEEE Xplore, Wiley Library, Scopus, Google Scholar, ACM, and Springer. For each of these databases, we utilized the same search query to retrieve the adopted literature. The search query comprises keywords and synonyms such as ("Trustworthy machine learning algorithms" OR "explainable machine learning algorithms in cybersecurity" OR "impact of explainable machine learning in cybersecurity") OR ("explainable machine learning in cybersecurity" OR "explainable machine learning approach in cybersecurity" OR "trustworthy machine learning cybersecurity models").

1.3 Research Questions

The research presented in this study is intended to review the concept of trustworthy machine learning algorithms in cyberattack detection to ensure ethics, interpretability, and fairness in machine learning cybersecurity detection models. The formulated research questions alongside the concept behind each research question are summarized in Table 1 which also forms the basis of our search criteria.

Table 1. Research questions

Research questions	Justification of research questions
RQ1: What are the trustworthy machine learning techniques?	To identify and analyze existing trustworthy machine learning algorithms
RQ2: What is the most efficient machine learning to implement an explainable cybersecurity model?	To identify and analyze the different machine learning models in implementing trustworthy cybersecurity models
RQ3: What are the impacts of explainable machine learning in the cybersecurity domain?	To identify the impacts of explainable machine learning in cybersecurity

The selected papers were entered into a data extraction spreadsheet in Microsoft Excel 2020. The following information was extracted from the studies: author(s), year, AI technique, explainable technique, types of cybersecurity attacks, and data sources.

2 Eligibility Criteria

The inclusion, exclusive, and quality selection requirement was utilized to screen the required articles for this study. The adopted papers for this study were obtained using the selection criteria.

2.1 Exclusive Criteria

This eligibility technique involves evaluating the retrieved articles based on the following criteria:

- The same literature from a different source
- Articles with the same topic and authors
- Articles that do not address the research questions
- Articles that do not discuss trustworthy machine learning models in cybersecurity, and explainable machine learning applications in cybersecurity.

2.2 Inclusion Criteria

This inclusion screening technique involves evaluating the retrieved articles based on the following criteria:

- Articles that are relevant to trustworthy machine learning cybersecurity models
- Journals, symposiums, white papers, methodologies, and conference papers that addressed trustworthy machine learning techniques in the cybersecurity domain.
- Comparative studies of trustworthy machine learning techniques in combating cyberattacks.

2.3 Quality Criteria

The downloaded papers were screened based on the following quality criteria:

- Do the articles provide relevant answers to either of the research questions?
- Do the articles emphasize explainable machine learning algorithms in cyberattack detection?
- Do the papers discuss the positive impacts of explainable machine learning algorithms in detecting cyberattacks?
- Does the research methodology in the papers appear to be adequately disclosed?

3 Selected Articles

We utilized the search criteria to perform a search query on the databases. The search result was retrieved in the bibliography reference CSV file. A total of 760 articles were retrieved from the databases and 40 articles were retrieved from other sources. From

the 800 retrieved articles, 320 articles consist of duplicated articles. The eligibility criteria were applied in the remaining 480 articles as earlier mentioned. The exclusion and inclusion screening techniques were applied in the title, abstract, and keywords which screened out a total of 281 articles. This means that a total of 199 articles are related to our study area but do not necessarily address all the required selection criteria. The titles, abstracts, and keywords were studied to identify relevant literature that discussed the research questions. After the successful completion of this phase, 199 articles were retrieved. The skimming and scanning reading approach was used to study the returned 25 research articles which involve careful study of the methodologies, headings, sub-sections, theoretical framework, and conceptual framework. After the successful completion of this phase, 174 didn't meet the selection criteria. In this paper, we select only papers that investigated trustworthy and explainable cybersecurity machine learning models, and comparative studies of explainable machine learning in cybersecurity. Figure 2 illustrates the selected year of publication. Figure 3 depicts the distribution of the selected articles based on publication year, while Fig. 3 shows the number of retrieved articles based on publication outlets.

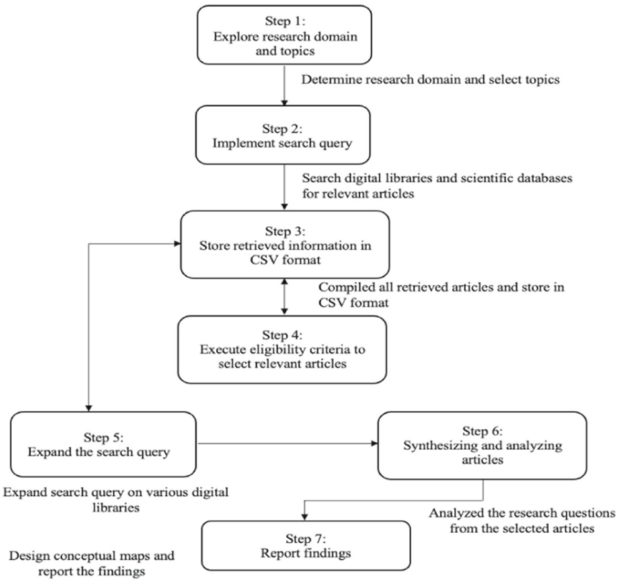

Fig. 1. Seven-step survey model

The results in Fig. 3 demonstrate that 2019 and 2020 had the most publications (7), respectively, followed by 2018 (6). While the number of selected publications (1) is lowest in 2016.

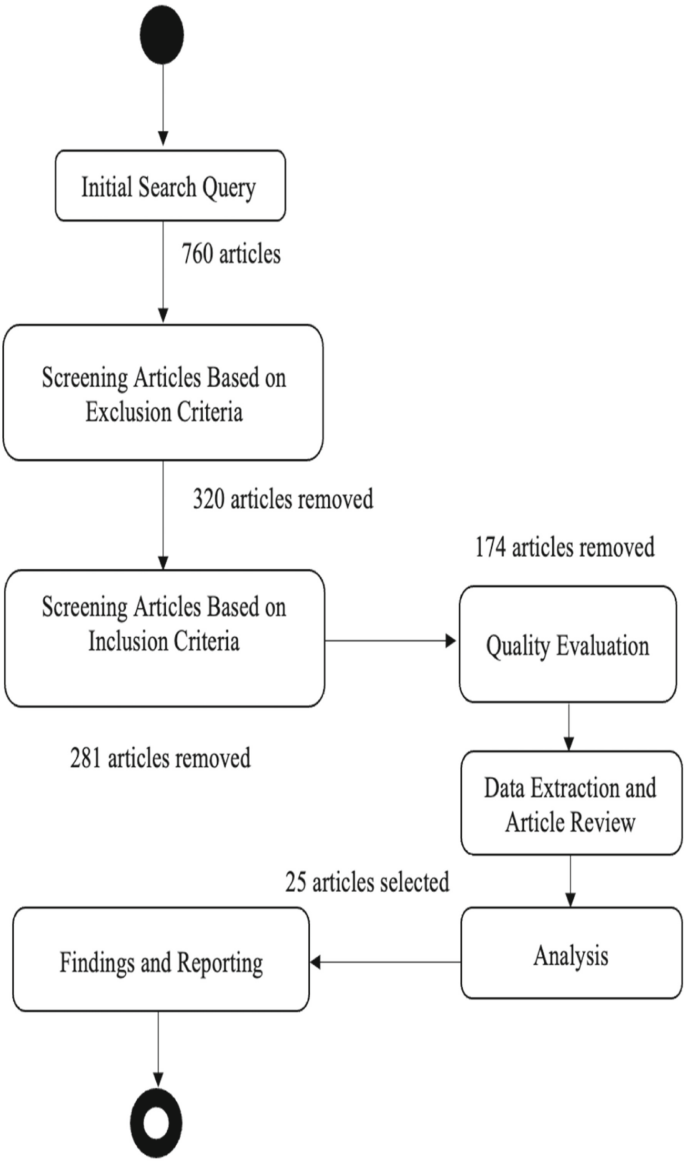

Fig. 2. Mapping of the search and selection procedure

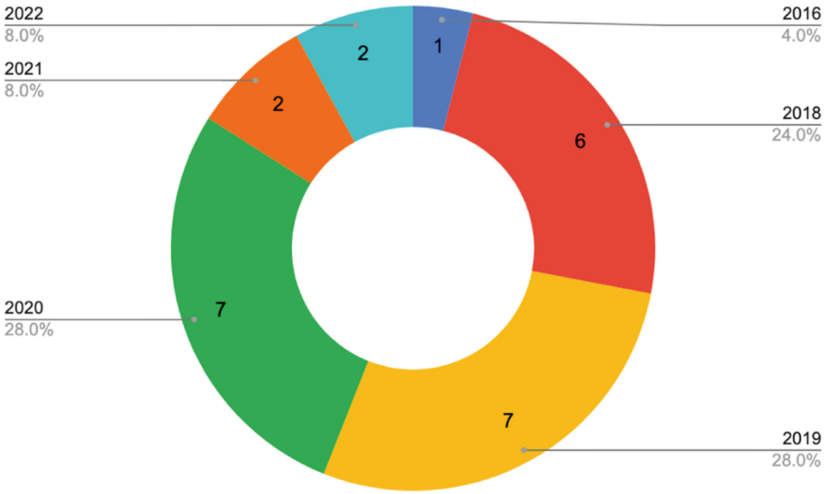

Fig. 3. Distribution of publication year

4 Research Questions

This section provides answers to the research questions in Table 1.

4.1 RQ1: What Are the Trustworthy Machine Learning Techniques?

This subsection discusses the trustworthy machine learning techniques utilized in the cybersecurity domain.

In [10] an explainable machine learning framework in intrusion Detection Systems (IDSs) to optimize the decisions of cybersecurity experts is evolved. The authors utilized SHapley Additive exPlanations (SHAP), in collaboration with global and local explanations to enhance specific inputs. The important features extracted, and the correlation between the different types of attacks and their feature values were provided by the global explanation technique. The authors also compared the interpretations between the one-vs-all classifier and the multiclass classifier. The proposed framework enhanced the transparency of any IDSs and can help cybersecurity practitioners obtain a quality understanding of IDSs judgment. 'The results obtained from the compared classifiers can play a vital role in designing reliable cybersecurity infrastructure, and design reliable structures for IDSs. [11] evolved an explainable machine learning model to enhance trust management in IDSs. The authors asserted that existing studies only demonstrated the performance accuracy of different machine learning classification techniques to enhance trust in IDSs without consideration for explainable machine learning models. Hence, the authors evolved an XAI machine learning model using a decision tree to enhance the trust management of IDSs. The authors utilized the decision tree technique because it is easy to read, and the decision tree technique is like a human expert decision-making process in terms of splitting the choice into various smaller sub-choices in IDSs. [12] evolved an adversarial approach for XAI in IDSs to trigger explanations for inaccurate

classifications produced by data-driven IDSs. The authors obtained the minimum qualification required to adequately classify a set of misclassified samples using the adversarial technique. The most relevant features that interpret the reason for the misclassification were visualized using the magnitude of the qualifications. The result shows the explanations generated by the model efficiently describe the reason behind the misclassifications with the interpretation that matches the cybersecurity practitioner's knowledge. Variable importance techniques, also known as feature importance, estimate the statistical impact of each feature on a model's performance to generate explanations [2, 13]. A machine learning cybersecurity implementation pipeline can utilize feature importance to determine whether a model is learning correctly and whether the features contributing to the predictions are trustworthy. Global explanations are provided by feature importance techniques. Mean decrease in impunity, permutation importance, and conditional feature importance are examples of feature important techniques. When a single feature is randomly shuffled, permutation importance evaluates how much black-box models predicted accuracy suffers [2]. Permutation importance works by retesting a model multiple times, each time with the values of a separate feature mixed up randomly. When each feature is shuffled, this technique measures prediction performance. If rearranging the values of a feature has no effect on predictive performance, the feature is not contributing significantly to the model's performance. Permutation importance also provides simple explanations that are generally easily understandable [2]. The heuristic technique of permutation importance seeks to correct the feature importance bias by normalizing feature importance measures. The method attempts to estimate the parameters of a feature based on the premise that its random relevance follows some probability distribution. This is accomplished by permuting the output array of predictions periodically and then assessing the importance distribution for each feature on the non-permuted output. The adjusted measure of feature importance is proxied by the computed p-value [1]. The permutation importance technique has the advantage of being computationally efficient because they do not require constant retraining of models. However, in circumstances where a model uses strongly correlated features, permutation importance can provide incorrect outcomes. In such circumstances, conditional variable importance can be used, however, it has the disadvantage of being less computationally efficient.

Surrogate techniques, such as short decision trees or logistic regression models, explain a black-box machine learning technique by generating a new interpretable model based on the black-box model's prediction. The purpose of training with the black-box model is to get the new interpretable model as close as possible to the black-box model. The interpretable model can be used to acquire a basic concept of how the black-box model operates by users looking for an explanation [14]. To explain relationships between a machine learning prediction outcome and a set of variables of interest, visualization methods for black-box interpretations can be employed. Malicious patterns can be explained using these techniques [15].

Rule-based learning refers to any model that develops rules to describe the data from which it intends to learn [4]. The simplistic conditional if-then rules or more complicated combinations of simple rules may be used to create their knowledge. Fuzzy rule-based systems, which are connected to this general family of models, are built for a larger variety of actions, enabling the formulation of vocally defined rules across imprecise

domains. Firstly, they enable more intelligible models since they function in linguistic terms. Second, in scenarios with varying degrees of uncertainty, they outperform standard rule systems. Rule-based models are transparent models that may be used to explain complicated models by establishing rules that explain their outcomes [4]. Rule-based models are great models in terms of interpretability across domains. They are good for understanding and describing other models due to their seamless and natural interaction with human behavior. A rule wrapper may be deemed to include enough information about a model to explain its behavior to a non-expert if a particular degree of coverage is achieved, without compromising the ability to utilize the derived rules as an independent prediction mode [4]. Hence, this technique is efficient for ensuring trustworthy cyberattack detection.

4.2 RQ2: What Is the Most Efficient Machine Learning to Implement an Explainable Cybersecurity Model?

Only a few explainable machine learning models have been utilized in literature to archive trustworthy cybersecurity. In this section, the study identifies and analyzes the different machine learning models utilized in literature in terms of explainability in the cybersecurity domain. [11] compared the decision tree technique with other state-of-the-art machine learning techniques to ascertain their performance. The result shows that the decision tree technique outperformed other state-of-the-art techniques in terms of marginal improvement. The performance accuracy, recall, F1score, and precision of the decision tree algorithm outperformed other state-of-the-art techniques in terms of explainability. The authors asserted that the decision tree technique is easy to evaluate, and it is computationally cheaper than other supervised learning techniques. Among the supervised learning approaches, a decision tree is nonparametric. It makes decisions using a tree-like technique and shows the probability of event occurrences. This technique includes a conditional control statement and provides straightforward and interpretable decision rules. Unlike other supervised models, it allows you to observe and interpret the data logic. The decision tree technique assesses the importance of features, generates explainable rules for threat detection, and achieves comparable performance with state-of-the-art techniques in terms of accuracy [11]. SVM is a compelling technique, and it is also computationally expensive, and non-interpretable. The missing value in the dataset does not affect the process of development and deployment in the decision tree technique which is a competitive advantage over other state-of-the-art machine learning techniques. Another advantage of the decision tree is that it does not need to be normally distributed. However, there may be a chance of overfitting if the model encounter noise in the dataset. Hence, the decision tree technique might perform inadequately in the testing set but perform better in the testing set. Another drawback of the decision tree technique is that categorical features with a different number of levels, and information gained in the technique is biased in favor of those features with more levels. This action might influence prediction performance, but the decision tree technique is considered the most efficient technique in terms of interpretability and trustworthiness [16].

4.3 R3: What Are the Impacts of Explainable Machine Learning in the Cybersecurity Domain

In this subsection, the study impact of explainable machine learning algorithms in the cybersecurity domain. The transparency, fairness, and trustworthiness of explainable machine learning cybersecurity models are becoming increasingly important. These models can be utilized to eliminate the risk of making cybersecurity decisions that are unethical, unjustifiable, and difficult to explain to users and cybersecurity experts. The misclassification of cyberattacks can cause serious damage to any organization and create vulnerability to organizations' sensitive resources [11]. Explainable machine learning plays a crucial role in ensuring trust management and understanding the impact of malicious data to detect any system infiltration. The focus on explainable cybersecurity detection machine learning models in recent times, which is very important in enhancing practical deployment of machine learning cybersecurity infrastructures can enhance their design and empower their implementation because interpretability and fairness can strengthen the robustness of cybersecurity models. This draws cybersecurity experts' attention to possible adversaries that could negatively influence or manipulate machine learning predictive models. Explainable machine learning models are also capable of ensuring the trust of cybersecurity models through offering meaningful variable inference and model reasoning causality. Explainable machine learning models can also help ensure impartiality in the decision-making process, for instance, to detect and correct biases such as imbalanced datasets in the training set [11, 17, 18]. A lot of research using black-box techniques based on the KDD original dataset has reported high-performance accuracy for malicious threat detection. Detecting malicious threats is essential but understanding and explaining how they are detected, and their decision-making process is a critical issue. Black-box machine learning techniques are very complicated and complex to understand, and it is sometimes very convoluted how they derive their conclusion. This makes them complicated I am making sensitive decisions such as malicious threat detection. It is essential that the decision-making process of machine learning in the cybersecurity domain need to be understood. Threat intelligent experts have a right to the details and decision logic of cyberattack detection models [19–23].

Explainability techniques aim to overcome the concerns of black-box machine learning models, which have long been a major roadblock to machine learning adoption in sensitive areas such as cybersecurity. Explainability can improve the robustness of cybersecurity infrastructures by highlighting possible adversaries that could alter the predictions. Explainability ensures that only contextual accurate variables deduce the outcome (i.e., in the reasoning method, there must be an underlying truthful causality).

Cyberattacks can be easily interpreted if they are explained as rules. Simple interpreted rules, such as If, then statements, are an example. Also, the explanation should be able to explain which network features are infringed, which sections of the target of the attack, and which parts of the security regulations are violated. Explainable machine learning address cybersecurity transparency requirements, and there is some indication that it can improve trustworthiness in cybersecurity systems. Hence, explainable machine learning is necessary for proving fairness, trustworthiness, and reliable interpretation of the outcome of machine learning cyberattack detection models [11, 19, 24, 25].

5 Conclusion

The contribution of this study is to execute an in-depth review and identification of existing explainable machine learning algorithms for cyberattack detection. Existing explainable machine learning techniques for defending against cyberattacks were identified using this selected article identified in this study. Machine learning takes decisions by analyzing vast volumes of data using a dynamic system to find potentially hidden patterns and weak signals. Despite the accuracy and excellent performance of machine learning cybersecurity detection models, their rising complexities are huge setbacks when users are involved, as these models cannot provide any explanation or information about the logic for their outcome. Interpretability in machine learning-based cybersecurity solutions has become critical for increasing confidence in machine learning-based cybersecurity infrastructures. Cybersecurity infrastructures are sensitive programs that can be attacked maliciously. However, understanding how a cybersecurity system makes decisions is important for establishing confidence, resolving issues of justice, and debugging a model to ensure trustworthy cybersecurity. This paper investigates the explainable machine learning techniques that can be utilized to enhance trustworthiness in cyberattack detection. In this study, extensive literature and state-of-the-art explainable machine learning techniques were reviewed. The impacts of trustworthy machine learning techniques were also reviewed to ascertain the best machine learning model in terms of interpretability in the cybersecurity domain. The study reveals that the decision tree technique outperforms other state-of-the-art machine learning models in terms of transparency and interpretability. Finally, this research suggests that incorporating explainable into machine learning cybersecurity models will help uncover the root causes of defensive failures, making it easier for cybersecurity experts to enhance both cybersecurity infrastructures and development, rather than just model results, policy, and management.

The number of research on trustworthy and explainable machine learning techniques in combating cyberattacks that have been published in recent years has highlighted the potential for improvement by considering the advantages and enhancements that these techniques can bring to existing machine learning frameworks in combating cyberattacks, but it has also exposed their flaws, weaknesses, and how much performance they lack. In any case, we anticipate that explainable machine learning still has many untapped potential areas to examine in combating cyberattacks in the future. In the future, we consider implementing a trustworthy federated learning framework for combating fraudulent credit card transactions.

References

1. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 18 (2020). https://doi.org/10.3390/e23010018
2. Petch, J., Di, S., Nelson, W.: Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can. J. Cardiol. **38**(2), 204–213 (2022). https://doi.org/10.1016/j.cjca.2021.09.004
3. Pienta, D., Tams, S., Thatcher, J.: Can trust be trusted in cybersecurity? In: Proceedings of the 53rd Hawaii International Conference on System Sciences (2020). https://doi.org/10.24251/hicss.2020.522

4. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **82**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
5. Perarasi, T., Vidhya, S., Leeban Moses, M., Ramya, P.: Malicious vehicles identifying and trust management algorithm for enhance the security in 5G-VANET. In: Second International Conference on Inventive Research in Computing Applications (ICIRCA) (2020). https://doi.org/10.1109/icirca48905.2020.9183184
6. Nassar, M., Salah, K., Rehman, M., Svetinovic, D.: Blockchain for explainable and trustworthy artificial intelligence. Wires Data Min. Knowl. Discov. **10**(1), e1340 (2019). https://doi.org/10.1002/widm.1340
7. Dosilovic, F., Brcic, M., Hlupic, N.: Explainable artificial intelligence: a survey. In: 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (2018). https://doi.org/10.23919/mipro.2018.8400040
8. Zhu, J., Liapis, A., Risi, S., Bidarra, R., Youngblood, G.: Explainable AI for designers: a human-centered perspective on mixed-initiative co-creation. In: 2018 IEEE Conference on Computational Intelligence and Games (CIG) (2018). https://doi.org/10.1109/cig.2018.8490433
9. Keeling, G.: Why trolley problems matter for the ethics of automated vehicles. Sci. Eng. Ethics **26**(1), 293–307 (2019). https://doi.org/10.1007/s11948-019-00096-1
10. Wang, M., Zheng, K., Yang, Y., Wang, X.: An explainable machine learning framework for intrusion detection systems. IEEE Access **8**, 73127–73141 (2020). https://doi.org/10.1109/access.2020.2988359
11. Mahbooba, B., Timilsina, M., Sahal, R., Serrano, M.: Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity **2021**, 1–11 (2021). https://doi.org/10.1155/2021/6634811
12. Marino, D., Wickramasinghe, C., Manic, M.: An adversarial approach for explainable AI in intrusion detection systems. In: IECON 44th Annual Conference of the IEEE Industrial Electronics Society (2018). https://doi.org/10.1109/iecon.2018.8591457
13. Daghistani, T., Elshawi, R., Sakr, S., Ahmed, A., Al-Thwayee, A., Al-Mallah, M.: Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. Int. J. Cardiol. **288**, 140–147 (2019). https://doi.org/10.1016/j.ijcard.2019.01.046
14. Jones, Y., Deligianni, F., Dalton, J.: Improving ECG classification interpretability using saliency maps. In: IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE) (2020). https://doi.org/10.1109/bibe50027.2020.00114
15. Elshawi, R., Al-Mallah, M., Sakr, S.: On the interpretability of machine learning-based model for predicting hypertension. BMC Med. Inform. Decis. Making **19**(1) (2019). doi: https://doi.org/10.1186/s12911-019-0874-0
16. Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (2018). https://doi.org/10.1109/dsaa.2018.00018
17. Carvalho, D., Pereira, E., Cardoso, J.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019). https://doi.org/10.3390/electronics8080832
18. Guo, W.: Explainable artificial intelligence for 6G: improving trust between human and machine. IEEE Commun. Mag. **58**(6), 39–45 (2020). https://doi.org/10.1109/mcom.001.2000050
19. Wu, M., Hughes, M.C., Parbhoo, S., Zazzi, M., Roth, V., Doshi-Velez, F.: Beyond sparsity: tree regularization of deep models for interpretability. In: Proceedings of AAAI Conference on Artificial Intelligence, New Orleans, LA, USA (2018)
20. Zahavy, T., Zrihem, N., Mannor, S.: Graying the black box: understanding DQNs (2022). https://doi.org/10.48550/arXiv.1602.02658. Accessed 12 May 2021

21. Luong, N., et al.: Applications of deep reinforcement learning in communications and networking: a survey. IEEE Commun. Surv. Tutor. **21**(4), 3133–3174 (2019). https://doi.org/10.1109/comst.2019.2916583
22. Huawei: AI Security White Paper (2018). https://www-file.huawei.com/-/media/corporate/pdf/trust-center/ai-securitywhitepaper.pdf. Accessed 2 Jan 2021
23. Gunning, D.: DARPA's explainable artificial intelligence (XAI) program. In: Proceedings of the 24th International Conference on Intelligent User Interfaces (2019). https://doi.org/10.1145/3301275.3308446
24. Castelvecchi, D.: Can we open the black box of AI? Nature **538**(7623), 20–23 (2016). https://doi.org/10.1038/538020a
25. Sarhan, M., Layeghy, S., Portmann, M.: Evaluating standard feature sets towards increased generalisability and explainability of ML-based network intrusion detection (2021). https://doi.org/10.48550/arXiv.2104.07183. Accesssed 9 Jan 2022