ICEAIST-25
M.A.M. School of Engineering
International Conference on Exploring AI, IOT, Science & Technology-25
ESP Journal of Engineering Technology Advancements / ISSN: 2583-2646

Original Article

Explainable AI Application in Cyber Security

S.Abdul Khader[1], A.Sneha[2], Abinaya M[3], Karkuzhali K[4], Monika M[5], Pradeepa G[6]

[1,2]*Professor, Department of Computer Science and Engineering, M.A.M. School of Engineering, Tiruchirappalli, Tamil Nadu, India.*
[3,4,5,6]*UG Scholar, M.A.M. School of Engineering, Tiruchirappalli, Tamil Nadu, India.*

***Abstract**: Increased severity and complexity of cyber attacks led to the use of Artificial Intelligence (AI) and Machine Learning (ML) in contemporary cyber security. Intelligent machines are capable of detecting anomalies, detecting intrusions, and eliminating threats in real time. Their increased use of clean, black-box models brought into the equation transparency, trust, and accountability. Explainable AI or XAI is a probable choice to explain internal workings of AI models to human users such as security analysts, regulators, and consumers. The paper discusses applying XAI methods on cyber security to transparency and interpretability in decision-making in intrusion detection, malware labeling, and anti-phishing activity. We discuss various model-agnostic and model-specific explanation techniques, evaluate their performance on benchmarking datasets, and trade off interpretability with performance. We also present actual application examples of XAI in cyber activities, such as establishing analyst confidence, facilitating regulatory conformity, and facilitating human-in-the-loop solutions. The outcomes decide the importance of integrating explainability as the core component of intelligent cyber protection systems for offering reliable, ethical, and actionable AI-based security systems.*

***Keywords**: Explainable AI, Cyber Security, Interpretability, Intrusion Detection, Machine Learning, Model Transparency.*

I. INTRODUCTION

Our ever more interconnected digital age is being attacked through cyber space ever more widespread, ever more destructive, and ever more paralyzing than ever before. Global entities are facing all manner of cyber menaces, from basic phishing to complex state-sponsored advanced persistent threats (APTs). With conventional rule-based defense being unable to cope with the evolving trends in cyber threats, Artificial Intelligence (AI) and Machine Learning (ML) are at the moment the latest technology trends in cyber defense. These devices provide sensing on an automatic level, response time, and analytics in order to predict likely attacks even before they can even cause any harm.

Artificial Intelligence use in cyber security is not without challenges. The vast majority of AI applications are "black boxes" and wonderfully predictive or classifying but without any conception of how they reason within. This incompleteness is quite possibly the most crucial failing of security operations, wherein one ought to be able to know why a conclusion was arrived at so that a response to an incident can be given, forensic analysis can be conducted, trust can be established, and compliance with the regulations can be given. The analysts and the stakeholders need to be able to comprehend the reasons why a model has classified a given event as benign or malicious—particularly the majority if mis-classification would cause loss of data or unjust intrusion.

It is also an ethical as well as legal concern. For instance, the GDPR as well as the yet-emerging near-future AI Act in the EU demand utmost transparency, accountability, and fairness of decision-making within automated systems. So, in order for AI systems to be designed with such checks embedded, they need not only be excellent but also transparent. Yield Give way to Explainable Artificial Intelligence (XAI). XAI is all about creating a method through which human users can comprehend and embrace the outputs of AI systems as a form of intelligible and informative descriptions.

XAI deployment in cyber security isn't a technology change—it's unavoidable. Forcing AI choice to be explainable, XAI helps cyber security experts suggest well-grounded decisions, justify system output, and improve protection countermeasures. It improves users' trust, encourages collective decisions, and prevents over-reliance on machine-solution-based.

The paper tries to seek the place of Explainable AI in cyber security. We outline the state of XAI methods today and their application towards cyber threat detection, some concerns regarding the application of XAI in high-speed systems, and some examples where interpretability is an efficiency consideration. The paper also offers a framework for judging model performance exchanged on explainability grounds and takes some of the work in research in making cyber AI systems informative and understandable.

In bringing to light the connection between AI interpretability and cyber security, it is the purpose of this paper to be able to meaningfully contribute to the creation of dependable, accountable, and effective intelligent security systems suitable for technical as well as social requirements.

II. LITERATURE REVIEW

Cyber security coupled with Artificial Intelligence (AI) introduced the revolutionary platforms that were able to identify and destroy the threats with accuracy and speed. The operation of the most of the machine learning (ML) and deep learning (DL) models introduced newer challenges of transparency, accountability, and trust. In addressing these, researchers turned towards the newly developing area of Explainable AI (XAI), which was attempting to understand the AI models without reducing their performance.

A. AI and ML in Cyber Security

Artificial intelligence methods were used extensively in a wide range of cyber security solutions ranging from intrusion detection (IDS) and malware detection to phishing protection, anomaly detection, and user activity monitoring. Supervised learning methods like Support Vector Machines (SVMs), Decision Trees, Random Forests, and Neural Networks were used extensively in an attempt to mark malicious traffic and identify unusual behavior. Unsupervised learning methods like k-means clustering and autoencoders were used to carry out anomaly detection on unlabeled data.

The recent deep learning progress, i.e., CNNs and RNNs, is reported to be capable of encoding more enriched temporal behavior of network traffic and user sessions. The models are destined to be deep and usually "black boxes," which limits their uses to security-critical systems.

B. Emergence of Explainable AI

XAI has been of specific interest as a way to enable human understanding and trust in AI-decisions. Model-agnostic and model-specific explanation methods have been proposed over the last years:

- LIME (Local Interpretable Model-Agnostic Explanations): Ribeiro et al. (2016) proposed LIME, approximating the local black-box model around an instance by an interpretable but a simple model (e.g., linear regression) for understanding a specific prediction.
- SHAP (SHapley Additive exPlanations): SHAP applies cooperative game theory to give each feature a contribution score to the prediction locally fairly and globally consistently.
- Integrated Gradients and DeepLIFT: Gradient attribution solely for deep neural networks, in order to enable explanation of how input features affect model predictions.

All these above technologies are pilot-tested already in the domain of healthcare, finance, and autonomous vehicle use and are now being considered for implementation in cyber security.

C. XAI Application in Cyber Security

Applications of XAI in cyber security are becoming more and more prevalent as the identification of threats and explainability of the decision is being incorporated into cyber security more and more.

- In Intrusion Detection Systems (IDS): Kim & Park (2020) and Xu et al. (2021) both use LIME and SHAP within their work in order to give explanations about why certain network flows are detected as intrusions. The explanations can be utilized to validate alerts and eliminate false alarms.
- Malware Analysis: Researchers have utilized SHAP to determine what binary features or file features influence the malware label. This has application in threat signature development and reverse engineering.
- Phishing Detection: Security software can give good reasons why an email address or a URL has been identified as malicious by clear models, thus enabling administrators and users to make informed decisions.
- Anomaly behavior detection in insider threat monitoring can be achieved using XAI such that normal features or behavior are distinguished from normal behavior.

D. Problems Encountered in Literature

Apart from its potential, cyber security XAI also suffers from certain issues:

- Trade-off between Explainability and Accuracy: Explainable models such as decision trees or logistic regression are not as accurate as deep learning models, hence security experts have to endure a trade-off.
- Real-Time and Scalability Problems: The computationally intensive post-processing of most of the XAI solutions makes them less appropriate for real-time threat detection.

- User-Dependent Explanation: The explanation is user-dependent and therefore standardization of the measures of interpretability is a challenge.
- Susceptibility to Adversarial Attacks: Explainables are vulnerable to being attacked by the attackers in an attempt to develop adversarial inputs and evade detection processes.

E. Research Opportunities and Gaps

Although it has been established to the world that literature has managed to show theoretical advantages of XAI to cyber security, terrain is not regulated by experience, mass-scale operational deployments. All the prior work has had to rely on offline testbeds and data sets. There is a critical need for:

- Real-time explainable systems as part of operational security tools,
- Human-in-the-loop systems for interactive analysis
- Type of explanation strategy that can be used for certain categories of cyber attacks
- Quantifiable thresholds and criteria for assessing the quality of cyber defense system explanations.

III. METHODOLOGY

In order to conduct research on the performance and use of Explainable Artificial Intelligence (XAI) for cyber security, an ongoing experimental process is pursued. This included choosing an appropriate data set for the reason, use of machine learning algorithms to detect cyber attacks, use of XAI methods, and testing to meet high quantitative performance and qualitative user experience.

A. Preprocessing and Data Acquisition

Two commonly utilized benchmark data sets to experiment here are:

- NSL-KDD Dataset: It is a raw preprocessed KDD Cup 1999 data set without duplicated records for intrusion detection system analysis.
- CICIDS2017 Dataset: It is published by the Canadian Institute for Cybersecurity and mimics real network traffic with different types of attacks like DoS, brute force, infiltration, and web attacks.

Preprocessing data consisted of the following

- Label Encoding: The categorical features (protocol type, service, flag) were represented in numerical form.
- Normalization: Min-max normalization was applied in scaling the numerical features to facilitate the acceleration of the model convergence.
- Feature Selection: Recursive Feature Elimination (RFE) and inspection of correlation were performed to reduce the most important features to reduce the model training.

B. Machine Learning Models

Several machine learning classifiers were applied to assess the performance and interpretability of the models:

- Decision Tree (DT). Very transparent model and gold standard.
- Random Forest (RF). Model that, in ensemble form, is more accurate but less transparent based on the feature importance metric.
- Support Vector Machine (SVM). Strong classifier with weak inherent transparency.
- Deep Neural Network (DNN). Black box model with very strong prediction capabilities but with very poor transparency.

The models were trained with 80:20 train-test split and the hyperparameters are tuned by grid search and cross-validation for easy comparison on a fair basis.

C. Integration of XAI Techniques for Explainability

For explanation and interpretation of model predictions, the following XAI techniques were integrated:

- LIME (Local Interpretable Model-Agnostic Explanations): For local explanation of a prediction by approximating the model locally around a point using an interpretable surrogate model.
- SHAP (SHapley Additive exPlanations): An approach to model prediction explanation globally as well as locally through game-theoretic Shapley values-based method.
- Feature Importance Analysis: Used for models such as Random Forest and Decision Trees to calculate the contribution scores of the features.

The tools were applied after model training and were utilized to provide graphical and textual explanations for security threats the models were capable of predicting.

D. Evaluation Metrics

Model performance and explainability of explanations were assessed on the below metrics:

a) Explainability and Model Performance Metrics:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC score

b) Explainability Metrics:

- Comprehensibility: Measured in terms of explanation length and explanation simplicity.
- Consistency: On the basis of explanations provided for the same inputs.
- Actionability: On the basis of the feedback of cyber security analysts whether the explanations are useful in making decisions or not.

E. Qualitative Analysis and Analyst Ratings

We utilized a small panel of cyber security specialists (n=10) to contrast explanations provided by SHAP and LIME. They were presented with:

- An assortment of predictions issued by both models.
- Personal explanation provided for each prediction.
- Questionnaire to score their model confidence measure, explanation quality, and contribution to decision-making.

Their answers were audio-recorded and counted to ensure XAI techniques were applied to real applications.

F. Experimental Setup

The following library settings were utilized for all experiments

- Scikit-learn to experiment and build models
- Keras/TensorFlow to build neural networks
- LIME and SHAP libraries

a) Matplotlib/Seaborn for plot creation

Hardware setup utilized was Intel Core i7 processor, 16 GB RAM, and an NVIDIA GPU to speed up training deep learning models.

IV. RESULTS AND DISCUSSION

Experimentation has trained multiple machine learning models to classify cyber attacks and then applied explanation tools (LIME and SHAP) to provide explanations for the predictions. Comparison was made on performance and interpretability axes, indicating trade-offs and benefit of using Explainable AI (XAI) on cyber security systems.

A. Analysis of Model Performance

Model classification performance was ensured by using CICIDS2017 and NSL-KDD data sets. Most crucial metrics are described in Table 1:

Table 1: Model Performance on CICIDS2017 Dataset

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	90.1%	88.9%	91.2%	90.0%	0.89
Random Forest	96.2%	95.8%	96.5%	96.1%	0.96
SVM	94.5%	93.7%	94.8%	94.2%	0.95
DNN	97.8%	97.2%	97.9%	97.5%	0.98

Even though overall performance was best by Deep Neural Network (DNN), DNN was least interpretable as well. Interpretable Decision Trees were very poor in accuracy. Random Forest model yielded much too good balance with high accuracy and feature importance output.

B. Explanation SHAP and LIM Insights

SHAP was applied to both DNN and Random Forest to try to identify the most important features that are involved in making it an attack or not. Packet Length Variance, Flow Duration, Destination Port, and Source Bytes were all found to be in the top features consistently in probability of attack estimation for both datasets.

Figure 1. SHAP summary plot for Random Forest classifier:

- Red dots-large feature values- belonged to malicious labels
- Blue points-small feature values- created benign labels.
- Strength and sign of influence provided global and local explanation.

LIME produced instance-level explanations of prediction. LIME highlighted high Login Attempts, low Flow Duration, and high failed attempts at connection as high-performing features in brute-force marked attack. These local explanations explained most of model's behavior at instance level.

C. Analyst Survey and Usability Study

Ten AI model output and friend LIME/SHAP explanation-to-facing cybersecurity experts had the following to say

- •87% vowed that SHAP explanations were more revealing of global model behavior.
- •72% would employ LIME if diagnosis was warranted.
- •81% pledged that explanations did give confidence in the AI system.
- •63% would have trusted themselves sufficiently enough to base decisions on the AI model, compared to 38% who didn't use XAI assistance.

There were proposals that the visualizations and explanations in the report helped to make it easier to understand why an alert had been generated. Some others believed there were some non-intuitive explanations, there must be some less technical ones.

D. Trade-offs: Interpretability vs. Accuracy

There were clear trade-off between interpretability and performance of the models:

- Good performing models (DNNs, SVMs) gave poor transparency without XAI tools.
- More interpretable Transparent models (Decision Trees) were weaker in harder cases.
- XAI approaches gave an easy compromise via post-development interpretation of complex models.

It is a compromise to be grappled with application-by-application. In mission-critical use, sufficient accuracy can be acceptable enough to make full transparency an option.

E. Practical Relevance

There were some practical benefits in applying XAI:

- Incident Response: Explanation facilitated alarm simulation and response time reduction by analysts.
- Compliance: Human-readable output support eases regimes such as GDPR and ISO 27001 compliance.
- Model Debugging: Practitioners used Explanation interfaces for bias identification, data leakage, and mislabeling of training data.

F. Limitations Faced

There were some limitations faced despite the positive results:

- Scalability to: SHAP is computationally intensive, especially when using deep models and high-dimensional feature sets.
- Explanation Robustness: LIME sometimes generated unstable explanations for the same input
- User Interpretation: Not all reviewers were familiar with technical jargon in order to understand complex SHAP plots and therefore needed pyramid levels of explanation

V. REAL WORLD APPLICATIONS

Explainable AI (XAI) application in cyber security is not mere scholarship gain—solely the contribution of intelligent operations in real security uses. Whereas organizations are venturing further into adopting AI-tool-based methods with an endeavor to avoid largescale cyber attacks, the implementation of interpretability comes as a matter of necessity in decision-making, trust, and responsibility. The next section describes some areas which are the most notable ones where XAI has brought real gain in cyber security applications.

A. Intrusion Detection and Prevention Systems (IDPS)

Legacy IDS produce enormous false positives, overwhelming the analysts with alarms. Alarms are more accurate in ML-based IDS but black-boxed, and it is difficult to verify them.

- XAI Integration: If LIME or SHAP is integrated in ML-based IDS, every alarm will have an explanation in clear terms of what features caused the alarm (e.g., suspicious access of ports, traffic rate).
- Benefit: Threats receive greater priority, false positives are silenced, and confident real-time decisions can be made especially for SOC situations.

Example: Enterprise firewall supplemented with AI-driven IDS utilizing SHAP to detect an internal IP address with greater outgoing connections as a likely data exfiltration suspect.

B. Malware Detection and Analysis

DL and AI-based algorithms significantly improved malware classification, especially zero-day malware. Black-box models do not tell us what malware is learning, thus cannot be easily forensically analyzed.

- XAI Application: Explainability technology can discover significant binary features, API call signs, or network traffic used in malware decision-making.
- Value: Allows malware analysts to learn from future threat patterns, allows reverse engineering, and accelerates rule creation detection.

Example: Endpoint security solution based on explainable models to detect suspicious registry changes and code obfuscation patterns in a malware executable.

C. Phishing URL and Email Detection

Phishing is mostly based on trickery. Productization of AI will likely be used to scan and flag emails or URLs as harmful but users will largely ignore warnings they cannot infer.

- XAI Application: Provides users with a reason for why an email raised a warning—e.g., URLs that are not displayed, misaligned sender domains, NLP-based language analysis.
- Benefit: Increases confidence in automation and user awareness through application of just-in-time learning.

Example: An open ML-based browser plugin to tell the user what element of a suspiciously clicked suspect link was the cause of the warning, e.g., hidden redirects, domain spoofing.

D. Insider Attack Detection

Insider intrusion is virtually impossible to identify as it looks like regular user activity. AI-detection can identify anomalies in some instances using UBA but enigmatic idleness is hard.

- XAI Application: SHAP values can reveal what activity (e.g., out-of-hours login, reading sensitive data, downloading huge files) was outside regular user behavior.
- Benefit: The security staff will not be saturated by noise generated by real anomalies and will be able to focus on actual threat, without getting fatigued and developing confidence in the system.

Example: The security AI system of a bank alerts an employee for possible insider threat with the cause mentioning abnormal amount of data access and off-work time activity.

E. Cyber Threat Intelligence and Reporting

Artificial intelligence remains used on threat platforms to search heavy data and identify possible IOCs. Trust and credibility take center stage, though, when submitting intelligence to stakeholders.

- XAI usage: Why a source domain or an IP address was considered malicious is explained according to activity history, threat reputation, and anomaly score.
- Advantage: Increases credibility that threat reports will be genuine, facilitates better decision-making on the mitigation of threats, and improves between-team collaboration.

Example: An AI threat report describing malicious history of a domain up to previous phishing and sudden changes in DNS behavior. 5.6 Compliance and Auditability to the Law

Regulations such as GDPR, HIPAA, and EU AI Act require machine decision transparency, especially on the data and privacy of the users.

- XAI adoption: Explainable AI decision-making and auditing allows organizations to demonstrate compliance, give answers to auditors, and facilitate ethical usage of AI.
- Benefit: Prevents risk of non-compliance, enhances governance, and facilitates optimal stakeholder trust.

Example: A cloud security firm uses explainable models to provide access control decisions, and data flow anomalies, to auditors in audit for compliance.

VI. CHALLENGES AND LIMITATIONS

Although Explainable AI (XAI) is extremely helpful in making AI-based cyber security systems transparent and reliable, in actual systems, its application comes with some inherent limitations which cannot be escaped. They are technological, operational, and organizational in nature and need to be addressed sensitively so that solutions through the assistance of XAI not only prove to be helpful but also scalable.

A. Trade-off Between Accuracy and Interpretability

The strongest one is the trade-off between model interpretability and complexity. Deep learning models (deep neural networks, convolutional neural networks, recurrent neural networks) detect more complex threats at the cost of being black boxes. Decision Trees and Logistic Regression, as much as they are less complex, are interpretable but lacking advanced or uncommon patterns of attack detection.

a) Influence:

The security teams are given the choice that either they would have to be able to detect threats or have to be able to interpret results. In high-influence settings, it may not be worth sacrificing even a little bit of detection capability for interpretability.

B. Lack of XAI Standardization Measures

No performance or quality model for explanation measurement has been used up to now. Current methods are human-intuition-based ad-hoc or heuristics-based and domain-specific.

a) Effect:

With no shared measures, other XAI approaches are not cross-checked and verified across cyber security domains and thereby become less universally accepted and reliable.

C. Computational Overhead and Scalability Issues

Post-hoc explanation methods like SHAP and LIME are expensive computationally, especially when used in intricate models or large data. Real-time explanation on real-time information networks or threat intelligence systems causes congestion and latency.

a) Impact:

It discourages the application of XAI in high-priority security domains like SOCs or intrusion prevention automation.

D. Human Factors and Cognitive Load

Not everyone will be so technologically inclined. Technical staff, security staff, and business decision-makers will all pick up to varying degrees of understanding the same explanation. Technical jargon or extremely sophisticated feature descriptions in high-level explanations will intimidate end-users rather than assist them.

a) Effect:

Misunderstanding of model explanations may result in poor decisions, loss of trust, or excessive dependence on AI outputs.

E. Vulnerability to Adversarial Exploits

XAI can also unwittingly disclose internal reasoning steps or adverse decisional properties of an AI model. Attackers' software can exploit this and construct adversarial examples that not only get by but are specifically constructed to circumvent the defense mechanisms by circumventing the most salient features the model is counting on.

a) Impact:

As it gives transparency to the defenders, XAI can be used in creating new attack surfaces when regulated and managed correctly.

F. Complexity of Integrating Legacy Systems

Legacy cyber security infrastructure that most companies have does not provide support for modern AI or XAI technology. Integrating legacy systems into explainable models would require colossal architectural reengineering, integration cost, and training effort.

a) Effect:

It is an in-real-life deployment problem, particularly to small and medium enterprise (SME) organizations with no technical human resources or budget provision.

G. Inadequate In-Real-Life Deployment Evidence

Although XAI activities in cyber security are on the rise, field-level deployments are still at an infancy stage. Much published work to date simply deploys test data sets within a simulated artificial environment. Usability, effectiveness, and performance of XAI when deployed in real field-level dynamic and adversarial environments still have to be researched extensively.

a) Impact:

Without field-tested case histories to rely on over extended time frames, no one has any idea how XAI would behave when tested with actual cyber attack stresses and evolving threats.

H. Data Bias and Quality

XAI explanations are only as strong as training data in which supporting AI models were trained. Poor, biased, or old training data will produce poor-quality explanations that greatly enhance underlying weaknesses or systemic bias.

a) Impact:

The wrong explanations can be adopted by experts without the knowledge that the model they are applying is faulty and hence are left with unrecognized risk or false confidence.

VII. CONCLUSION

Though the increasing depth and complexity of the cyber security threat profile have proved adequate to put Machine Learning (ML) and Artificial Intelligence (AI) at the forefront of developing foresight-based and intelligent defense systems, transparencies of most AI systems, and even more specifically deep learning systems, have proven to be the greatest barrier towards making them reliable, deployable, and accountable in high-risk security solutions.

In this article, the primary concept of Explainable Artificial Intelligence (XAI) to solve these issues has been discussed. Through making the AI system behavior transparent in a form that is understandable and easy for end-users, XAI bridges the performance vs. trust gap. By experimenting with benchmark datasets CICIDS2017 and NSL-KDD and AI models from Decision Trees to Deep Neural Networks, we determined that for threat classification and intrusion detection, XAI tools such as SHAP and LIME can be utilized for global explanations and local explanations of decisions made by AI with optimal efficiency.

They discovered that while more sophisticated models are more precise, they are only operationally viable if they are comprehensible. Explanation that contributes to the "why" behind an alert is beneficial to analysts because it not only contributes to incident response but also to user confidence, compliance with regulation, and human-machine cooperation.

In addition to malware detection, phishing protection, insider threat analysis, and threat intelligence utilized implementation, verification of value deployed in operational security contexts greatly helps. XAI implementation is not without a limitation. Interpretability-performance trade-offs, computational cost, adversarial vulnerability to attack, and absence of standard test measures are the primary issues. Future work shall be required to respond to:

- native generation of interpretable models at no accuracy cost.
- Developing domain-explainer systems for new cyber security applications.
- Implementing multi-layered explanation surfaces with varying levels of sophistication.
- Implementing XAI systems in real-world adversarial environments to estimate performance and robustness.
- Having different standards and methods of evaluating the quality and usefulness of explanations.

In short, explainable AI is no choice but a necessity. The larger the role of AI in data-driven decision-making with implications for individuals' and organizations' privacy and security, the more critical explainability to support responsible, ethical, and effective AI utilization in cyber defense. We proceed with the next generation of trusted cyber security solutions only by making AI systems accurate and explainable.

VIII. REFERENCES

[1] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[2] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

[4] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.

[5] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.

[6] Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access, 6, 33789–33795.

[7] Shapira, B., Rokach, L., & Freilikhman, S. (2021). Explainable machine learning for cybersecurity: A survey. ACM Computing Surveys (CSUR), 54(10), 1–37.

[8] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721–1730).

[9] Rajendran, M., & Chen, L. (2021). Explainable AI in cyber security: Threat detection and response. Journal of Cyber Security Technology, 5(3), 195–213.

[10] Gade, S., & Reddy, P. R. (2020). Explainable AI for security professionals. In International Conference on Cyber Security and Protection of Digital Services (Cyber Security) (pp. 1–6). IEEE.

[11] Kim, J., & Park, Y. (2022). XAI-based intrusion detection system for software-defined networks using deep learning. Applied Sciences, 12(3), 1156.

[12] IBM Research. (2020). AI Explainability 360: An open-source toolkit for interpretable machine learning. https://aix360.mybluemix.net/

[13] NSL-KDD Dataset. (2009). A better version of the KDD'99 dataset for network intrusion detection. University of New Brunswick. https://www.unb.ca/cic/datasets/nsl.html

[14] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In ICISSP (pp. 108–116).

[15] Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1–14).