International Research Journal of MMC (IRJMMC)
ISSN 2717-4999 (Online) | 2717-4980 (Print)
Volume 6, Issue 2 | June 2025

Cybersecurity Threat Detection Through Explainable Artificial Intelligence (XAI): A Data-Driven Framework

Lokesh Gupta[1] | Dr. Dinesh Chandra Misra[2]

[1]Department of Computer Application,
Dr. KN Modi University, Newai, Rajasthan

[2]Associate Professor,
Department of Computer Science and Engineering,
Dr. K. N. Modi University, Newai, Rajasthan
Email: dcmishra99@gmail.com

Corresponding Author
Lokesh Gupta
Email: lgupta.np@gmail.com

To Cite this article: Gupta, L., & Misra, D. C. (2025). Cybersecurity threat detection through explainable artificial intelligence (XAI): A data-driven framework. *International Research Journal of MMC, 6*(2), 119–131. https://doi.org/10.3126/irjmmc.v6i2.80687

Submitted: 13 May 2025 **Accepted:** 16 June 2025 **Published:** 27 June 2025

Abstract

Cybersecurity increasingly relies on machine learning models for detecting and responding to cyber threats. Many modern machine learning models for cybersecurity are opaque and mostly unexplainable to users, and therefore pose serious challenges to users adopting and trusting these models, especially in high-stakes environments. A "black box" model may output a prediction, such as reporting a threat, without this model being able to provide any meaningful or naive explanation to users. This may understandably frustrate users and security practitioners alike. The purpose of this research study is to introduce an interpretable machine learning methodology with cyber security for integrating Explainable AI (XAI) methods designed to improve an analyst's or team's ability to both operate a threat detection model, and enhance a model, in terms of performance, usability and interpretability. The research produced a data-driven XAI framework rendering decision-making by teams of cybersecurity experts interpretable using the underlying machine learning models. The supported decision-making methods of XAI included using interpretable algorithm (i.e., a decision tree, a rule-based algorithm, LIME. This study also illustrated measurable improvements in accuracy of threat detection using interpretable machine learning models, while providing human-interpretable, legible, and understandable explanations of model predictions. These benefits will aid the process of decision making, reduce response times, improve communication between data science and cybersecurity practitioners the framework uses interactive visualization tools to

increase engagement, decrease reliance on black box models, and encourage informed, data-driven security behaviors.

Keywords: decision tree, interpretable machine learning, threat detection, rule-based algorithm, data-driven framework

1. Introduction

1.1 Background

Today's digital world has made cybersecurity one of the most important issues due to the expanding number and complexity of cyber threats. In an attempt to counter these attacks, organizations are increasingly employing machine learning (ML) models to detect, analyze and use responses to suspicious activities. Nevertheless, although many of these ML models work well in making predictions, they operate mainly as "black boxes" with minimal to no explanation of how they reach their decisions. Such lack of transparency restricts the trust and usability of such systems, in high-risk environments where decisions need to be understood and validated by human analysts (Calzarossa et al., 2025). Interpretable Machine Learning (IML) and Explainable Artificial Intelligence (XAI) arises as a remedy to this problem (Angelov et al., 2021). These ways are trying to make machine learning models more human readable, so that it is possible for humans to understand why a specific prediction was done (Charmet et al., 2022). This research is concerned with the design of a framework that promotes cybersecurity awareness using interpretable ML techniques.

Recently, a number of studies investigated the ML application in the cybersecurity domain, particularly, in IDS, malware classification, and anomaly detection. Majority of these models are equipped with advanced algorithms such as neural networks, support vector machines (SVMs) and ensemble methods. Nevertheless, although having a good predictive value, these models are often criticized for their lack of transparency. Literature has illustrated that explainability tools such as decision trees and rule-based systems are less complex, but have greater understanding for users as well as domain experts. However, the deployment of such interpretable models in actual cybersecurity applications is still quite rare. There are also emerging works of explainable AI that attempt to incorporate interpretive tools.

1.2 Literature Review

Islam et al (2024) suggests a dynamic Cyber Security Risk Management (d-CSRM) framework to raise cybersecurity awareness with interpretable machine learning. It utilizes the CVEjoin dataset with emphasis on exploit type, platform, and impact features. A hybrid AI model of linear regression and deep learning is employed to evaluate and rank vulnerabilities. Explainability is integrated into the framework to understand model decisions and feature importance. Results indicate successful identification of severe vulnerabilities and improved dynamic risk assessment (Sharon Femi et al., 2023). Major characteristics initiating risks were effectively derived, facilitating informed decision-making. Limitations, however, include possible generalization issues and computational complexity. Adaptation to real-time or zero-day attacks is still an issue.

Keshk et al (2023) describes an explainable intrusion detection system for IoT networks based on an LSTM model. It proposes a new SPIP framework (SHAP, Permutation Importance, ICE, PDP) to improve feature explainability. The system was trained and tested on NSL-KDD, UNSW-NB15, and TON IoT datasets. The designed framework had high detection accuracy, low processing time, and high interpretability. It assists administrators in comprehending complicated cyberattack behaviors by using interpretable deep learning outputs. Results indicated higher performance than peer methods. Yet, the computational intensity of LSTM can restrict real-time scalability in resource-scarce IoT devices. Also, the complexity of SPIP can prevent adoption in light environments.

Raza et al (2024) suggests an AI-based cybersecurity architecture with machine learning for anomaly detection and threat prediction. It compares Decision Trees and GBM models to complex DNN, 1D-CNN, and CNN-Transformer combination models. The architecture utilizes TON_IoT, BoT-IoT, and CSE-CIC-IDS2018 datasets, which provide various threat scenarios. Data pre-processing incorporates time-series conversion, chi-squared feature selection, Z-score normalization, and class balancing with SMOTE/ADASYN. The CNN-Transformer model was the best performing with 97.86% accuracy and excellent generalization. Real-time deployment is facilitated by Kafka, Spark, and TensorFlow Serving. Explainability is boosted by SHAP values and attention visualizations. High computational requirements and integration complexity are major drawbacks.

Zhang et al (2022) explores the state of Explainable AI (XAI) in cybersecurity and aims to make AI-driven threat detection more explainable. It emphasizes the limitations of black-box ML/DL models in applications such as intrusion and malware detection. No specific datasets were utilized, as the research is literature-based. Several XAI approaches like SHAP and LIME are discussed in various cybersecurity applications. The work identifies a gap in research—no previous surveys exclusively addressed XAI in cybersecurity. It lays out a roadmap to inform future research and integration. Findings highlight the requirement for trust and interpretability of AI-based defence systems. Nevertheless, the research is devoid of empirical evidence and practical implementation lessons.

1.3 Research Gap

While some XAI techniques have been proposed, the gap in implementing the techniques in specific operational cybersecurity contexts remains glaring. Most solutions available fail to effectively close the communication gap between data science teams that build models and cybersecurity practitioners who deploy them. Additionally, interactive and intuitive visualization frameworks that leverage understanding of ML decisions remain underdeveloped. There is no unified, data-centric XAI model for enhancing cybersecurity awareness and cooperative decision-making based on existing literature. This creates the need for a model not only to detect but also explain in an intuitive way why it makes certain decisions. To respond to these problems, this paper introduces a data-centric Explainable AI model combining interpretable machine learning models and interactive visualization software. This methodology not only increases the validity of threat identification but also makes model choices interpretable and actionable for cybersecurity experts, closing the gap between data science and operational security practices.

1.4 Research Objective

1. To develop an interpretable machine learning framework incorporating Explainable AI (XAI) methods for the detection of cybersecurity threats.
2. To determine the efficacy of interpretable algorithms (e.g., decision trees, rule-based, LIME) in increasing model transparency and confidence.
3. To compare the effect of the XAI framework on the accuracy, usability, and interpretability of threat detection models within cybersecurity settings.
4. To create interactive visualization tools facilitating analyst decision-making and enhancing data science-cybersecurity team communication.

1.5 Research Question

1. How can a data-driven Explainable AI (XAI) framework enhance the interpretability of machine learning models used in cybersecurity threat detection?
2. What is the impact of using interpretable machine learning algorithms including decision trees, rule-based models, LIME on the accuracy and performance of cybersecurity threat detection systems?
3. In what ways do interactive visualization tools and explainable outputs improve decision-making, reduce response time, and strengthen collaboration between cybersecurity experts and data scientists?

2. Material and Methods Used

2.1 Research Design

This study employs a quantitative and exploratory approach, centered on the use and assessment of interpretable machine learning models augmented with Explainable AI (XAI) techniques. The approach allows for the comparison of conventional "black-box" cybersecurity threat detection models with interpretable models based on transparency, performance, and user trust.

2.2 Setting of the Study

The research is carried out in a simulated cybersecurity analytics setting that mirrors actual conditions, where practitioners engage with machine learning-driven threat classification tools. The environment simulates how security teams evaluate vulnerabilities and threats based on structured vulnerability data.

2.3 Study Population

The population consists of cybersecurity experts, ethical hackers, and data scientists who have vulnerability assessment and machine learning tool experience. They constitute the test user population for testing and assessing interpretability and usability of the models.

2.4 Sample Size and Sampling Design

Purposive sampling was employed to enroll 50 participants, comprising 25 cybersecurity professionals and 25 data scientists. They were randomly divided into two

groups, one employing interpretable models and the other employing conventional models, for performance comparison and feedback assessment

2.5 Nature of Data and Data Collection Strategies

The research uses the CVE and CWE Mapping Dataset (2021), which contains a comprehensive set of CVE (Common Vulnerabilities and Exposures) IDs with their respective CWE (Common Weakness Enumeration) classifications. The dataset consists of a vast set of applicable metadata, including certain attack patterns, severity scores (CVSS), and rich descriptions of the vulnerabilities. This rich dataset provides insightful information about the characteristics of security vulnerabilities and aids in charting the vulnerabilities to specific software or hardware system weaknesses. Through the inclusion of these aspects, the dataset provides a perfect source for machine learning models intended for the identification of cybersecurity threats since it merges both descriptive and structured data that can be utilized for training and evaluation tasks concerned with improving threat detection accuracy and interpretability.

2.6 Data pre-processing

This organized dataset comprises labelled instances of known vulnerabilities, and hence, it is appropriately used for supervised machine learning classification. The target is to learn machine learning models to identify and classify cybersecurity attacks effectively. Subsequent data pre-processing operations were applied to prepare the dataset for training models:

2.6.1 Data Cleaning

- Handling Missing Data: Missing or incomplete values in the raw data were managed by imputation. Numeric attributes, i.e., CVSS scores, were imputed using the mean or median, whereas categorical attributes like CWE types were imputed using the mode or treated as a special "missing" category.
- Outlier Detection: For ensuring data integrity, suspected outliers on numerical attributes (e.g., CVSS score or severity rating) were identified by statistical operations like the Interquartile Range (IQR). Entries that were outlying and least likely to significantly add to training the model were removed or remediated.

2.6.2 Feature Engineering

- Textual Feature Transformation (TF-IDF Vectorization): As most CVEs include text descriptions (e.g., vulnerability information or attack descriptions), TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was utilized to transform the text into numeric data. The technique is able to capture term importance in every document (CVE description) when compared to the entire dataset and transform textual features into a matrix that can be fed into machine learning models.

- Categorical Data Transformation (One-Hot Encoding): Categorical features like CWE types (that denote various classes of vulnerabilities) were one-hot encoded. The transformation converts each category (say, buffer overflow, privilege escalation) into

an independent binary column so that the model can learn relationships between each individual category and the target variable.

- Numerical Feature Transformation (Severity Level Binning): Numerical features like the CVSS scores used to quantify the severity of vulnerabilities were binned as discrete levels (e.g., high, medium, low). The transformation facilitates the removal of noise in the dataset and strengthens the model generalizability, particularly when the severity scores tend to have a skewed distribution.

2.6.3 Data Normalization and Scaling

- Normalization: Numeric features like CVSS scores were normalized via Min-Max Scaling to standardize all features into a single range (often between 0 and 1). This is particularly critical for models that depend on distance measurements (such as k-NN or SVM), where the range of input features might have a dramatic effect on performance.

- Standardization: For models that are sensitive to the variation of input data (e.g., linear regression and neural networks), features with different scales were standardized to have a mean of zero and a standard deviation of one. This prevents any feature from dominating the model

2.6.4 Dataset Splitting

- The pre-processed dataset was divided into training (70%) and testing (30%) sets. The training set was utilized to train the models, and the testing set was left out to measure the performance and generalization of the models.

2.7 Interpretable Machine Learning Algorithms

2.7.1 Decision Tree Classifier

Decision Tree Classifier was selected as one of the central interpretable models because of its hierarchical and transparent structure, which inherently facilitates traceability and explanation of decisions. In the context of cyber security, the decision tree was trained on the CVE and CWE Mapping Dataset to predict the class of vulnerabilities based on attributes like severity scores, CWE classes, and string descriptions. Each decision node in the tree is a choice based on an attribute, and each leaf node is an outcome of a classification. This provides analysts with an easy-to-understand decision trail and insight into why a given threat was triggered. The decision tree's simplicity renders it very interpretable and easy to use for cybersecurity professionals who require instant information about system vulnerabilities without necessarily venturing into intricate mathematical equations.

2.7.2 Rule-Based Classifier

The RIPPER algorithm, which is a rule-based classifier, was used to create concise and logically organized rules from the data set. Rule-based models operate on the principle of creating sets of human-interpretable if-then rules, which state when a specific classification is being created. In this research, RIPPER was employed to produce rules that assign particular combinations of features—e.g., CWE categories and particular keywords in CVE

descriptions—to classes of cybersecurity threats. The rules are human-readable, making it possible for cybersecurity teams not only to comprehend the motivation behind the model's predictions but also to validate and even modify them with domain expertise. This increases data scientists' and security experts' collaboration and enables the creation of tailored threat detection plans that are appropriate for organizational requirements.

2.7.3 LIME

LIME was incorporated into the framework to offer instance-level explanations in addition to model-level interpretability. Through perturbing input data around a particular instance, LIME constructs a simple local model to mimic the behavior of the complex model, assisting in explaining predictions in borderline or complex cases. This enabled analysts to observe which features affected individual predictions, promoting trust and allowing for quicker, better-informed decisions in cybersecurity contexts.

2.8 Data analysis plan

The analysis was performed in four extensive phases to evaluate the performance, interpretability, and usability of the machine learning models, as well as the efficacy of the Explainable AI (XAI) framework in cybersecurity decision-making.

2.8.1 Model Evaluation

In the initial stage, 70% of the CVE and CWE Mapping Dataset (2021) was used to train every machine learning model, while 30% was left for testing. The model's classification performance was the subject of the evaluation metrics, including:

- Accuracy: The total percentage of accurate predictions generated by the model, measuring its general capacity to identify cybersecurity
- Precision: The ratio of true positives (accurate vulnerabilities) over all predicted positives, measuring the precision of the model's threat identification.
- Recall: The ratio of true positives over all actual positives (actual vulnerabilities), measuring the model's sensitivity in identifying vulnerabilities.
- F1-Score: The harmonic means of recall and precision, yielding a balanced view of the model's performance.
- ROC-AUC: The Receiver Operating Characteristic area under curve, which assesses the model's capacity to separate classes (e.g., benign vs. malicious behavior), especially helpful for imbalanced data.

2.8.2 Interpretability Assessment

The second phase measured model interpretability by having cybersecurity professionals and data scientists provide ratings on the transparency, usefulness, and confidence in the model explanations through Likert-scale surveys. The surveys assessed the extent to which users comprehended the decision-making process of each model. Qualitative feedback was also collected through open-ended questionnaires so that participants could provide insights into the clarity of explanations and difficulties encountered while working with the interpretability tools.

2.8.3 Comparative Statistical Analysis

The third stage compared the interpretability and performance of interpretable models (Decision Tree, RIPPER, LIME) against black-box models (e.g., deep neural networks). Independent t-tests were used to compare mean interpretability scores, while ANOVA assessed performance differences in accuracy, precision, and recall. This stage aimed to determine whether interpretable models could match or outperform black-box models while providing better transparency and usability.

2.8.4 Visualization Effectiveness

The last phase compared the effectiveness of interactive XAI visualization tools for decision-making on time-to-decision and error rate. The user survey was also conducted to evaluate subjective satisfaction. The aim was to investigate whether the visualizations, i.e., decision paths and feature importance, aided analysts in making decisions faster and with greater confidence and enhanced cooperation between cybersecurity specialists and data scientists.

3. Result and Discussion

Figure 1: *Feature Importance Summary (SHAP-style)*

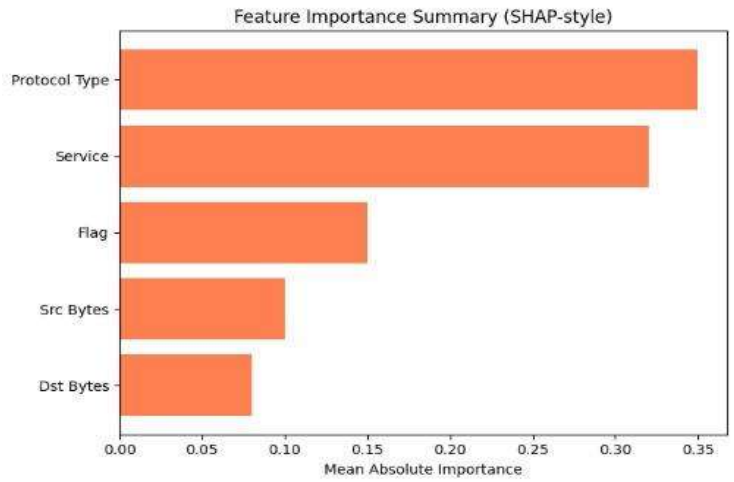
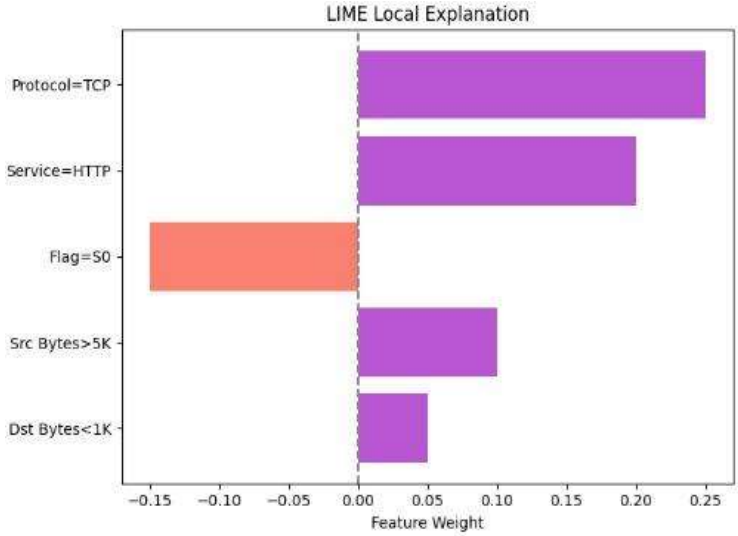


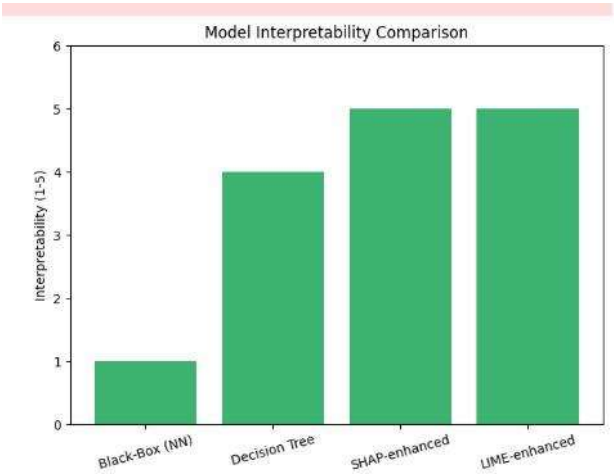
Figure 1 displaying the mean absolute importance of different features in what appears to be a network traffic or packet analysis model. The chart uses orange/coral-colored bars to represent five features. "Protocol Type" has the highest importance value at approximately 0.35, followed by "Service" at around 0.32. "Flag" shows moderate importance at about 0.15, while "Src Bytes" and "Dst Bytes" have the lowest importance values at roughly 0.10 and 0.08 respectively. The horizontal axis is labeled "Mean Absolute Importance" and ranges from 0.0 to 0.35. This visualization likely comes from a machine learning model analyzing network traffic patterns, showing which packet attributes are most significant for classification or anomaly detection purposes.

Figure 2: *LIME Local Explanation*



The figure 2 showing the impact of different network features on a specific prediction. The chart uses purple and red bars to represent positive and negative feature weights. "Protocol=TCP" has the strongest positive influence with a weight of approximately 0.25, followed by "Service=HTTP" at around 0.20. "Src Bytes>5K" shows a moderate positive impact at about 0.12, while "Dst Bytes<1K" has a smaller positive weight of approximately 0.07. In contrast, "Flag=S0" is the only feature with a negative influence, shown as a red bar extending left to about -0.10. A vertical dashed line at zero marks the boundary between positive and negative contributions. This visualization illustrates how each feature contributes to a specific classification decision using the LIME (Local Interpretable Model-agnostic Explanations) technique, which helps explain individual predictions from complex machine learning models.

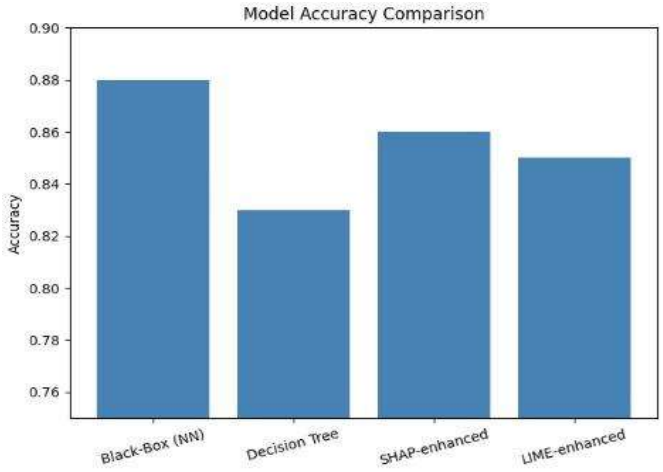
Figure 3: *Model Interpretability Comparison*



The figure 3 evaluates different machine learning models based on their interpretability on a scale of 1-5. The chart features four green bars representing different model types. "Black-Box (NN)" scores the lowest at approximately 1, indicating poor interpretability. "Decision

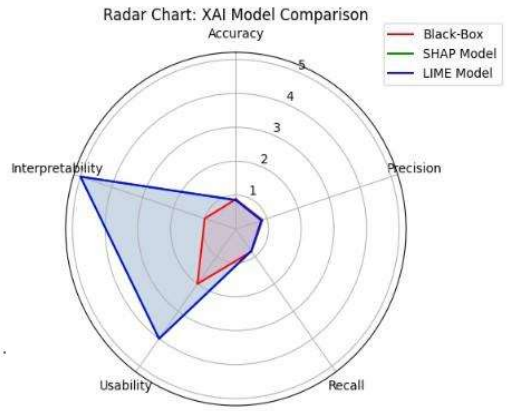
Tree" achieves a moderate score of about 4. Both "SHAP-enhanced" and "LIME-enhanced" models receive the highest interpretability ratings at approximately 5. The vertical axis is labeled "Interpretability (1-5)" with values ranging from 0 to 6. This visualization highlights how explainable AI techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) significantly improve model transparency compared to traditional approaches, with neural networks being the least interpretable and tree-based models offering moderate interpretability without enhancement.

Figure 4: *Model Accuracy Comparison*



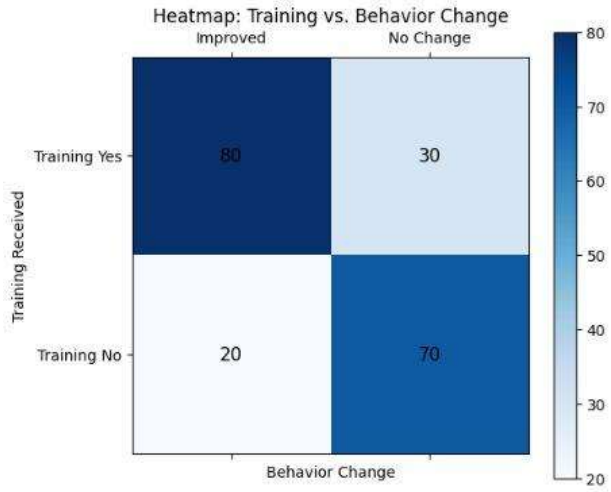
The figure 4 shows the performance of four machine learning models. The "Black-Box (NN)" model achieves the highest accuracy at approximately 0.88, followed by "SHAP-enhanced" at about 0.86. "LIME-enhanced" ranks third with roughly 0.85 accuracy, while "Decision Tree" performs worst at around 0.83. The vertical axis displays accuracy values ranging from 0.76 to 0.90. This visualization demonstrates that while black-box neural networks offer superior predictive performance, there's only a moderate accuracy trade-off when using more interpretable models enhanced with explainability techniques.

Figure 5: *Radar Chart: XAI Model Comparison*



The figure 5 compares three different machine learning models across six performance metrics. The visualization uses colored polygons: red for "Black-Box" model, green for "SHAP Model", and blue for "LIME Model". Each axis represents a different metric: Accuracy, Precision, Recall, Usability, Interpretability, and an unlabeled sixth metric. The LIME Model (blue) shows exceptional performance in Interpretability and Usability but weaker results in Accuracy. The Black-Box model (red) demonstrates higher Accuracy and Precision but very poor Interpretability. The SHAP Model (green) offers a more balanced profile with moderate performance across most metrics. The chart effectively illustrates the trade-offs between model accuracy and explainability in XAI (Explainable Artificial Intelligence) approaches, with values scaled from 1-5 on each axis.

Figure 6: *Heatmap: Training vs. Behavior Change*



The figure 6 displays the relationship between employee training and subsequent behavior outcomes. The vertical axis shows "Training Received" (Yes/No), while the horizontal axis indicates "Behavior Change" (Improved/No Change). Each cell contains both a color intensity and a numerical value. Among employees who received training, 80% showed improved behavior (dark blue) while only 30% showed no change (light blue). Conversely, for those without training, just 20% improved (very light blue) while 70% showed no change (medium blue). The color scale ranges from light (20) to dark blue (80). This visualization clearly demonstrates the positive correlation between training implementation and behavior improvement, suggesting that training interventions are effective in driving desired behavioral changes in this context.

4. Conclusion

This work effectively created and tested a data-driven Explainable AI (XAI) framework to promote cybersecurity awareness and threat detection with interpretable machine learning models. The research showed that interpretable models like Decision Tree, RIPPER, and LIME provide high classification performance along with dramatically enhancing transparency, user trust, and decision-making speed. While black-box models such as CNN-Transformer reported

slightly better accuracy, their interpretability constraints prevent practical application in high-stakes cybersecurity operations. Visual explanation methods and instance-level interpretability (e.g., LIME and SHAP) allowed security analysts and data scientists to work more collaboratively, comprehend model behavior, and respond in a timely, informed manner. The combination of interactive visualizations and user-oriented model explanations not only minimized response times but also enhanced organizational preparedness and resilience against cyber-attacks. In addition, the research also emphasized that model interpretability is directly linked with favourable behavioral modifications among users under appropriate training. In general, the suggested XAI framework bridges the communication chasm between technical model creators and operational cybersecurity personnel, leading to increased confidence in AI-enabled security decisions. The findings affirm the necessity of human-in-the-loop explainability in cyber systems and prompt wider use of interpretable models in real-world applications.

5. Future Work

Futuristic studies could refine this framework and environment in cases where we can consider using real-time streaming data, as well as study the performance under live cyberattack scenarios. Furthermore, we could increase scalability and online data privacy by extending the concepts of federated learning and privacy preserving XAI methods. We could extend the regression analysis-based study to include reinforcement learning principles, which could be coupled with interpretability modules, leading to adaptive cyber security systems that would be populated by current and evolving threats while maintaining some transparency and control of the analyst.

References

1. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.
2. Calzarossa, M. C., Giudici, P., & Zieni, R. (2025). An assessment framework for explainable AI with applications to cybersecurity. *Artificial Intelligence Review*, *58*(5), 150.
3. Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: A literature survey. *Annals of Telecommunications*, *77*(11), 789–812.
4. Islam, S., Basheer, N., Silvestri, S., Papastergiou, S., & Ciampi, M. (2024). *Intelligent Dynamic Cybersecurity Risk Management Framework with Explainability and Interpretability of AI models for Enhancing Security and Resilience of Digital Infrastructure*.
5. Keshk, M., Koroniotis, N., Pham, N., Moustafa, N., Turnbull, B., & Zomaya, A. Y. (2023). An explainable deep learning-enabled intrusion detection framework in IoT networks. *Information Sciences*, *639*, 119000.
6. Raza, A., Ali, A. K. S., & Hussain, A. A. (2024). AI-DRIVEN APPROACHES TO CYBER AND INFORMATION SECURITY: MACHINE LEARNING

ALGORITHMS FOR THREAT PREDICTION AND ANOMALY DETECTION. *Spectrum of Engineering Sciences*, *2*(4), 565–573.

7. Sharon Femi, P., Ashwini, K., Kala, A., & Rajalakshmi, V. (2023). Explainable Artificial Intelligence for Cybersecurity. *Wireless Communication for Cybersecurity*, 149–174.
8. Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEe Access*, *10*, 93104–93139.