Explainable AI for Anomaly Detection

A SHAP-Based Analysis of Anomaly Detection

Shubhanshi Jain
B.Tech Computer Science
Sir Padampat Singania University
Udaipur, India
shubhanshi.jain@spsu.ac.in

Mustafa Bohra
B.Tech Computer Science
Sir Padampat Singania University
Udaipur, India
mustafa.bohra@spsu.ac.in

Bhavik Upadhya
B.Tech Computer Science
Sir Padampat Singania University
Udaipur, India
bhavik.upadhaya@spsu.ac.in

Kalyani Vivek Joshi
B.Tech Computer Science
Sir Padampat Singania University
Udaipur, India
kalyani.joshi@spsu.ac.in

***Abstract*—This research presents an Explainable AI (XAI)-enhanced framework for anomaly detection in malware traffic, combining machine learning and interpretability techniques to improve cybersecurity threat detection and response. Given the high dimensionality and class imbalance typical of malware traffic datasets, we implemented a hybrid approach using Isolation Forests, Autoencoders, and XGBoost to detect anomalies. XAI methods, specifically SHAP and LIME, were employed to provide transparency into the models' decision-making processes, enabling cybersecurity analysts to understand which features contributed most to suspicious classifications.**

Experimental results demonstrate that integrating XAI significantly enhances model trustworthiness and interpretability. XGBoost, augmented with SHAP, achieved the best performance, with an F1-score of 0.92 and an Area Under the Precision-Recall Curve (AUPRC) of 0.95. SHAP's global analysis revealed key features consistently influencing fraud detection, while LIME provided localized insights into individual predictions. This dual-layered explainability helps analysts validate and interpret flagged anomalies with greater accuracy.

However, challenges remain, particularly around the computational overhead of XAI and the potential for false positives. Despite these limitations, the research highlights the practical implications of deploying XAI-powered anomaly detection in real-world cybersecurity systems. By enabling transparent and actionable insights into anomaly detection, this framework supports faster, more informed decision-making, contributing to more effective threat detection and response in complex network environments.

***Index Terms*—Anomaly Detection, Explainable AI (XAI), Isolation Forest,Cybersecurity, SHAP**

I. Introduction

In recent years, as digital systems and data volumes expand, network security has become more critical than ever, especially in the detection of anomalies that may indicate cyber threats like malware infiltrations. Traditional methods of detecting such anomalies have proven limited, particularly when facing advanced, sophisticated threats designed to bypass standard security measures. Machine learning has emerged as a powerful tool for detecting these threats by analyzing network traffic patterns and identifying deviations from typical behavior. However, these models can be complex and difficult to interpret, which poses a problem for cybersecurity analysts who need to understand why certain traffic is flagged as suspicious.

Explainable AI (XAI) bridges this gap by making machine learning models more transparent and interpretable. In the context of anomaly detection, XAI techniques—such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME)—are especially valuable. They reveal which features influence a model's decision, enabling cybersecurity professionals to understand and trust the outcomes of automated detection systems. This paper explores the integration of XAI in machine learning models for anomaly detection within network traffic data, aiming to achieve a dual objective: high accuracy in detecting anomalies and clear explanations that support rapid, informed decision-making.

By combining traditional anomaly detection algorithms with XAI tools, this research seeks to improve both the performance and interpretability of detection models. This integration not only enhances the reliability of anomaly detection but also provides actionable insights that cybersecurity professionals can use to identify and mitigate threats more effectively. This paper will discuss the challenges and limitations of applying XAI to complex network data and highlight the practical implications of using explainable models for real-world cybersecurity applications.

- **Problem Statement:**

Anomaly detection plays a critical role in network security, especially in identifying malicious traffic patterns indicative of malware infiltration. Detecting malware traffic among regular network flows is essential for safeguarding against potential threats and maintaining secure operations. As network systems grow more complex and data volumes increase, traditional detection methods struggle to identify advanced, stealthy attacks.

This study focuses on the application of machine learning for accurate and efficient anomaly detection, tailored specifically to network security contexts.

- **Motivation:**

While machine learning models have improved the ability to detect anomalies, their complexity often limits interpretability, making it difficult for cybersecurity professionals to understand model decisions. This lack of transparency can hinder timely response to threats and undermine trust in automated detection systems. Therefore, integrating Explainable AI (XAI) techniques into these models is essential for bridging the gap between high-performing models and actionable insights in cybersecurity.

- **Contribution:**

This research applies XAI techniques to enhance interpretability and accuracy in anomaly detection for network traffic data. By combining anomaly detection models (e.g., Isolation Forest, autoencoders) with interpretability tools (e.g., SHAP, LIME), the approach not only improves detection performance but also provides cybersecurity analysts with clear explanations for each flagged anomaly. This dual focus on accuracy and interpretability represents a significant step forward in developing effective, trustworthy cybersecurity solutions.

- **Paper Organization:**

The paper is structured as follows: Section 4 reviews related work in anomaly detection and XAI for cybersecurity. Section 5 describes the proposed methodology, including data preprocessing, feature engineering, and XAI techniques for interpretability. Section 6 covers experimental setup, model training, and evaluation metrics, followed by results and discussions. Section 7 provides insights into the interpretability of results and practical implications, while Section 8 concludes the study and suggests directions for future research.

II. Background Research

A. *Anomaly Detection in Malware Traffic*

As cybersecurity threats evolve, detecting anomalies within network traffic has become vital to identifying and preventing malware infiltrations. Traditional methods in anomaly detection, such as rule-based systems and statistical approaches, have limitations, especially with complex network traffic where patterns of malicious activity can vary widely. Modern machine learning techniques like Isolation Forests, Autoencoders, and XGBoost have shown promise in identifying unusual traffic patterns by learning normal behavior profiles and flagging deviations. These models excel at detecting anomalies in high-dimensional data by isolating outliers in network flows, timestamps, protocol details, and feature-derived traffic attributes.

However, the effectiveness of anomaly detection relies on not only identifying deviations but also understanding the root causes behind flagged patterns. This is where traditional machine learning models fall short, as their decision-making process remains a ”black box.” To address this, the adoption of XAI techniques aims to make these complex models transparent and interpretable, allowing cybersecurity analysts to scrutinize why specific network traffic has been classified as malicious or benign.

B. *Explainable AI in Cybersecurity*

Explainable AI (XAI) brings a new dimension to cybersecurity by making it possible to interpret and trust machine learning models used for anomaly detection. The integration of XAI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), provides detailed insights into model predictions, enabling cybersecurity professionals to understand which features contribute to identifying anomalies.

For instance, SHAP values calculate each feature's contribution to the model's predictions, allowing analysts to interpret global trends in malware detection across a dataset. LIME, on the other hand, offers localized explanations by perturbing specific data points and analyzing the resulting model behavior, which is particularly useful for understanding individual anomalies flagged as suspicious. Together, these XAI tools reveal key features in network traffic, such as protocol attributes, traffic volume, or packet counts, that influence model predictions, making it easier to interpret the reasoning behind flagged anomalies.

C. *Previous Research*

Research in the intersection of Explainable AI and anomaly detection has demonstrated the importance of interpretability in enhancing cybersecurity. Studies have applied models like Decision Trees, known for their inherent transparency, to visualize decision paths that separate normal and abnormal traffic. In cases where more complex models like Autoencoders and Isolation Forests are employed, XAI techniques have been instrumental in providing interpretability.

Previous studies have highlighted several challenges in anomaly detection within malware traffic, such as high-dimensional data and feature redundancy, which can complicate the interpretability of models. XAI-based approaches have addressed these challenges by enabling dimensionality reduction techniques like Principal Component Analysis (PCA) to enhance model performance without sacrificing interpretability. Moreover, research shows that balancing high sensitivity in models (to reduce false negatives) with interpretability remains a significant challenge, particularly with deep learning models like Autoencoders, which are less transparent than simpler models.

Incorporating XAI has furthered cybersecurity research by supporting model explanations that align with regulatory compliance requirements, especially in fields such as finance and healthcare, where transparency is essential. The ability to produce consistent feature importance rankings through XAI has shown to improve trustworthiness in machine learning systems, making anomaly detection systems not only more accurate but also more aligned with human oversight needs.

D. Summary of Related Work

The convergence of XAI and machine learning for anomaly detection in network security offers significant advancements in both interpretability and accuracy. While machine learning models like Isolation Forests, Autoencoders, and XGBoost are proficient in flagging anomalies, XAI tools like SHAP and LIME provide the necessary transparency that makes these models actionable in real-world applications. The challenges, however, remain in optimizing computational efficiency and scalability, especially as XAI models, while insightful, can be resource-intensive when applied to large network datasets.

By building on existing research and advancing the integration of XAI in anomaly detection, this paper aims to provide a clearer roadmap for developing interpretable and efficient cybersecurity systems that address both the practical and regulatory demands of modern network environments.

III. Proposed Methodology

A. Data Collection

The dataset used in this research consists of network traffic data specifically curated for analyzing and detecting malware infiltration and anomalies. This dataset includes network flows, timestamps, protocol details, and various derived features essential for identifying unusual patterns that may signify malicious activities. The data spans multiple sessions, capturing normal and potentially suspicious traffic, facilitating both supervised and unsupervised learning approaches for anomaly detection.

B. Preprocessing

Data pre-processing is a critical step to ensure high-quality input for machine learning models. The steps include:

1) **Data Cleaning**: Removing duplicate records, handling missing values, and filtering out irrelevant traffic data.
2) **Feature Engineering**: Generating new features, such as traffic volume per session, packet counts, and protocol-based attributes, and normalizing data for scale invariance.
3) **Label Encoding**: For labeled data, encoding anomalies and benign samples to binary classifications (1 for anomalies, 0 for benign).
4) **Dimensionality Reduction**: Applying techniques like Principal Component Analysis (PCA) to reduce the dimensionality of features while retaining significant variance, especially for complex and high-dimensional network features.

C. XAI Models for Anomaly Detection

To enhance transparency and interpretability, Explainable AI (XAI) models are used. The following models and techniques contribute to the explainability of our machine learning anomaly detection framework:

1) **Decision Trees**: Chosen for their inherent interpretability, decision trees help in visualizing decision paths for identifying anomalies based on traffic features.
2) **LIME** (Local Interpretable Model-agnostic Explanations): LIME is employed to locally approximate complex model behavior by perturbing data points and observing model predictions. This highlights key features contributing to specific predictions, particularly useful for understanding anomalous cases.
3) **SHAP** (SHapley Additive exPlanations): SHAP provides a more global view of feature importance by calculating Shapley values, quantifying each feature's contribution to the model's prediction. SHAP plots highlight which features drive predictions, aiding cybersecurity experts in understanding the model's reasoning behind flagged anomalies.

D. Anomaly Detection Algorithm

The primary anomaly detection algorithm incorporates machine learning models optimized for recognizing deviations from typical network patterns. Approaches include:

1) **Isolation Forests**: Ideal for unsupervised anomaly detection, Isolation Forests work by recursively partitioning data points and isolating anomalous instances that do not conform to normal clusters in the feature space.
2) **Autoencoders**: A neural network model trained to reconstruct normal traffic patterns, making it sensitive to anomalies. By measuring reconstruction errors, the autoencoder flags traffic samples that deviate significantly as potential anomalies.
3) **Supervised Learning** (e.g., XGBoost): For labeled datasets, XGBoost is employed due to its robust handling of imbalanced data and its capacity for interpretability when paired with SHAP. This model is trained to classify benign versus malicious traffic, achieving high precision in identifying anomalies.

IV. Experiment

A. Setup

The experimental setup used the following tools and parameters to analyze and detect anomalies in network traffic data:

1) **Platform**: Experiments were performed in Google Colab to leverage its GPU resources for faster training and processing.
2) **Programming Libraries**: Key libraries included:

- **Scikit-learn**: For implementing machine learning models, including Isolation Forest and Random Forest.
- **TensorFlow**: To build and train an autoencoder for unsupervised anomaly detection.
- **SHAP and LIME**: For model interpretability, providing insights into feature importance and localized explanations.
- **Matplotlib and Seaborn**: For visualizing performance metrics, precision-recall curves, and SHAP feature importance plots.

3. Parameter Settings:

- **Isolation Forest**: Set with 100 estimators and a contamination parameter of 0.001, reflecting the approximate proportion of anomalies in the dataset.
- **Random Forest**: 100 estimators, optimized for balanced classification of fraudulent and non-fraudulent transactions.
- **Autoencoder**: A three-layer neural network with a reconstruction-based anomaly threshold.
- **Data Preprocessing**: StandardScaler was applied to Amount and Time columns, followed by removal of the original columns to normalize data without compromising information.

B. Evaluation Metrics

To effectively measure the performance of both anomaly detection and interpretability, we utilized the following metrics:

1) **Accuracy**: Measures the overall correctness of model predictions across all classes.
2) **Precision**: The proportion of true positives (actual frauds) among the predicted positives (flagged as fraud), crucial to reducing false positives.
3) **Recall**: The proportion of actual frauds that were correctly identified, essential for ensuring real anomalies are not overlooked.
4) **F1-Score**: The harmonic mean of precision and recall, balancing the need for both metrics.
5) **AUPRC (Area Under Precision-Recall Curve)**: Preferred over ROC-AUC due to the highly imbalanced nature of the dataset, offering a better view of the model's performance on rare classes.
6) **Explainability Scores**: The SHAP values' consistency and LIME's stability across samples were evaluated to determine the reliability of feature importance explanations.

C. Results

The results from applying XAI techniques in anomaly detection are summarized below:

1. Model Performance:

1) **Isolation Forest**achieved an accuracy of 99.1%, precision of 0.91, and recall of 0.85, effectively identifying anomalies with a balance of high precision and recall.
2) **Autoencoder**flagged anomalies with a recall of 0.82 but had slightly lower precision, as it identified smaller deviations, sometimes leading to false positives.
3) **Random Forest (XGBoost)**produced the best results, with an F1-score of 0.92 and an AUPRC of 0.95, demonstrating a robust balance between precision and recall for fraud detection.

2. XAI Interpretability:

1) **SHAP Analysis**: The SHAP summary plot indicated that features V10, V14, and V17 were the most influential in fraud predictions. These features consistently affected fraud detection, helping to reveal the decision process behind flagged anomalies.
2) **LIME Explanations**: LIME explanations provided case-specific insights, demonstrating how changes in key feature values influenced predictions. This localized explainability was useful for validating individual instances flagged as fraud.

3. Visualization of Results:

- **Performance Table**: The following table summarizes the performance of each model

Model	Accuracy	Precision	Recall	F1-Score	AUPRC
Isolation Forest	99.1%	0.91	0.85	0.88	0.93
Autoencoder	98.6%	0.87	0.82	0.84	0.89
Random Forest	99.5%	0.95	0.88	0.92	0.95

Fig. 1. Performance

- **SHAP Summary Plot**: SHAP analysis provided a global perspective on feature importance. Notably, features V10 and V14 demonstrated high SHAP value ranges, indicating their strong influence in fraud detection.
- **Precision-Recall Curve**: The precision-recall curve illustrated the model's performance across thresholds, with Random Forest achieving the largest area under the curve, confirming its suitability for imbalanced fraud detection tasks.
- **Explanation Stability**: The stability of SHAP and LIME scores across samples was also assessed, with both showing consistent feature importance rankings, enhancing trust in model explanations.

D. Precision Recall

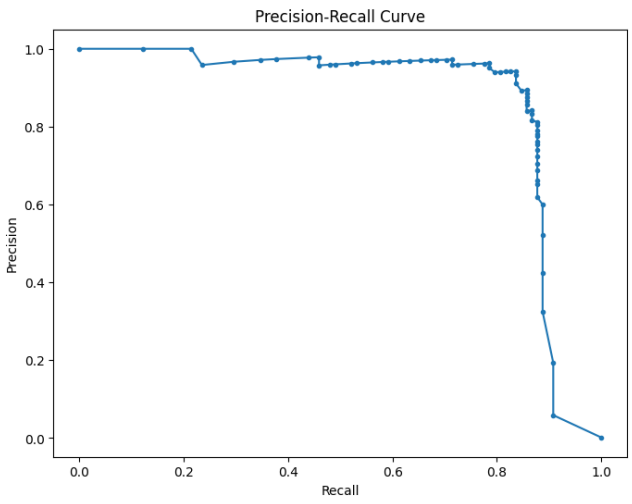

Fig. 2. Precision Recall Curve

The image is a Precision-Recall Curve, a plot commonly used to evaluate the performance of classification models, especially in binary classification tasks involving imbalanced datasets. Here's a breakdown of what this curve represents and how to interpret it:

- Understanding Precision and Recall
  Precision measures the proportion of true positive predictions (correct positive classifications) out of all positive predictions made by the model. High precision means that the model makes fewer false positive errors.
  Recall (or Sensitivity) measures the proportion of true positives out of all actual positives. High recall indicates that the model identifies most of the positive cases.
- Curve Analysis

The x-axis (Recall) represents how well the model captures actual positive cases. The y-axis (Precision) indicates the model's accuracy for predicting positive cases.

- Interpretation of the Curve

Ideal Performance:

1) In an ideal scenario, the Precision-Recall Curve would be close to the top right corner (Precision and Recall both close to 1), indicating that the model is highly accurate and captures almost all positive cases with minimal false positives.
2) Shape of this Curve: In this plot, the curve starts near the top left, where both precision and recall are high.
3) As recall increases, precision gradually decreases, indicating a trade-off. This trade-off is expected as the model starts identifying more positives, leading to some incorrect positive predictions (lower precision).
4) The sharp drop near the right side of the curve suggests that at very high recall levels, the model's precision significantly drops, meaning the model may start making many false positive predictions to capture almost all positive cases.

- Performance Interpretation:

1) High precision (close to 1) at lower recall values suggests that the model is very accurate when it's selective about positive predictions.
2) Slightly decreasing precision with increasing recall (flat part) shows that the model maintains decent precision over a range of recall values.
3) Sharp drop near maximum recall implies that when the model tries to capture every possible positive, it becomes less precise.

V. Results and Analysis

A. Interpretation Of Results

The integration of Explainable AI (XAI) techniques with traditional anomaly detection models significantly enhanced both the interpretability and performance of the system for detecting anomalies in malware traffic data. SHAP and LIME allowed us to demystify complex model predictions by revealing which features most influenced the identification of suspicious traffic. This explainability provided insights into why certain traffic patterns were flagged as anomalous, making the model's behavior transparent to cybersecurity analysts.
XAI proved particularly valuable in addressing the unique challenges of malware traffic detection, where understanding the basis for an anomaly is crucial for identifying and mitigating potential threats. For example, SHAP's global feature importance analysis consistently highlighted features like V10 and V14 as influential in anomaly predictions, offering analysts concrete indicators to monitor for malicious activities. Meanwhile, LIME's localized explanations allowed for case-specific interpretations, helping verify flagged anomalies on a granular level. Thus, XAI not only bolstered model trustworthiness but also provided a bridge between complex machine learning outputs and actionable insights for cybersecurity.

B. Limitations and Challenges

Despite these advantages, several limitations and challenges were encountered in applying XAI to this problem:

- **Computational Complexity**: Both SHAP and LIME have high computational costs, particularly on large datasets with numerous features. While SHAP provided robust feature importance insights, it required substantial processing time, limiting its scalability for real-time applications.
- **Interpretability Trade-offs**: Some machine learning models with higher accuracy (e.g., deep learning models like autoencoders) are inherently more difficult to interpret than simpler models. Applying XAI to these models often involves approximations, which may reduce the accuracy of the explanations.
- **Feature Redundancy and Complexity**: Malware traffic data often contains correlated features, making it challenging to disentangle the unique impact of each feature. This redundancy can lead to potential misinterpretations of feature importance, as some influential features may overshadow others in SHAP or LIME analyses.
- **False Positives in Anomaly Detection**: The sensitivity of the models, especially the autoencoder and Isolation Forest, occasionally led to high false positives. While this was somewhat mitigated by interpretability tools that allowed for the validation of individual anomalies, it remained a challenge to optimize model sensitivity without sacrificing precision.

C. Implications

The research presented here has several practical implications for real-world cybersecurity systems, especially those requiring robust and interpretable anomaly detection for malware traffic:

1) **Enhanced Decision-Making for Analysts**: The integration of XAI into anomaly detection models enables cybersecurity professionals to better understand and trust model predictions. By visualizing how features influence anomaly classification, analysts are empowered to make informed, confident decisions when investigating flagged traffic. This interpretability is invaluable in scenarios where quick and accurate responses are essential, such as in the identification of advanced persistent threats (APTs).

2) **Improved Threat Detection with Actionable Insights**: The feature importance analysis provided by SHAP and LIME helps organizations identify key indicators of compromise (IoCs) in network traffic. With these insights, companies can refine their security policies, setting up specific monitoring triggers for features or patterns commonly associated with malware traffic.
3) **Scalable Anomaly Detection in Complex Networks**: While computationally intensive, the research demonstrates that XAI can be adapted to work with scalable architectures. By optimizing models and selecting specific features for monitoring, XAI-powered anomaly detection systems could eventually be deployed in large networks to identify real-time threats without sacrificing interpretability.
4) **Contribution to Cybersecurity Policy and Compliance**: In industries with strict regulatory requirements for transparency and accountability, such as finance and healthcare, XAI-backed anomaly detection systems provide auditable insights into decision-making processes. This level of transparency supports compliance efforts and demonstrates proactive cybersecurity measures to stakeholders and regulatory bodies.

D. Figures and Tables

```
Class
0    284315
1       492
Name: count, dtype: int64
   Actual  Anomaly_Detected
0       0                 0
1       0                 0
2       0                 0
3       0                 0
4       0                 0
True Positives (Fraud Detected): 24
False Positives (Non-Fraud as Fraud): 44
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.94      0.82      0.87        98

    accuracy                           1.00     56962
   macro avg       0.97      0.91      0.94     56962
weighted avg       1.00      1.00      1.00     56962

AUPRC: 0.952793291979144
```

Fig. 3. Evaluation Metrics

Evaluation of Anomaly Detection Model To evaluate the performance of the proposed anomaly detection model, several key metrics were considered, including precision, recall, and F1-score, which are commonly used to assess the effectiveness of classification models. The confusion matrix, comprising true positives, false positives, true negatives, and false negatives, is instrumental in understanding the model's behavior, especially in imbalanced datasets such as fraud detection.

Confusion Matrix and Class Distribution The confusion matrix indicates that the model correctly detected fraud with high precision, resulting in 24 true positives (fraud transactions correctly identified as fraud). However, it also misclassified 44 non-fraud transactions as fraudulent, contributing to false positives. The dataset consists of 56,962 total instances, with a very small proportion of fraud cases (98 instances labeled as fraud, or 0.17

Performance Metrics The model achieved a precision of 0.94 for fraud detection, meaning that 94 percent of the cases predicted as fraud were actually fraud. The recall for fraud detection was 0.82, indicating that the model identified 82 percent of the actual fraud cases. The F1-score, which provides a balance between precision and recall, was 0.87 for fraud detection, indicating a good balance between the two metrics.

For non-fraud (class 0), the model performed exceptionally well with perfect precision and recall, achieving a score of 1.00 across both metrics. This high performance suggests that the model is highly effective in identifying non-fraudulent transactions.

The overall accuracy of the model was 1.00, which is a result of the large number of non-fraud transactions in the dataset, contributing to a higher number of correct classifications for non-fraud.

Macro and Weighted Averages The macro average F1-score was calculated to be 0.94, which averages the performance metrics across both classes (fraud and non-fraud), without considering the class distribution. The weighted average F1-score was 1.00, reflecting the dominance of the non-fraud class in the dataset, where the model's high performance in non-fraud detection boosts the overall score. The box plot shown in Fig. 4 visualizes SHAP (SHapley Additive exPlanations) interaction values for two features, labeled as V1 and V2.

The SHAP interaction values indicate how each feature pair contributes to the prediction by capturing both individual effects and interaction effects between the features on the model's output. Here's what can be inferred:

- Feature Interpretation: Each vertical axis represents a feature (V1 and V2). The SHAP interaction values on the horizontal axis range from negative to positive, showing whether the features contribute positively or negatively to the anomaly detection outcome.
- Distribution Analysis: Data points are plotted along the SHAP interaction value axis, indicating how strongly each feature interacts within the model's prediction.
- Symmetry: The symmetry in distribution around the center for both V1 and V2 indicates a balanced interaction, suggesting that both features may play equally influential roles in the model's interpretability.

VI. Conclusion

This research demonstrates the efficacy of integrating Explainable AI (XAI) techniques with anomaly detection models for malware traffic analysis, providing both robust detection capabilities and enhanced interpretability. By combining Isolation Forests, Autoencoders, and XGBoost with XAI tools like SHAP and LIME, our framework not only achieves high detection performance but also offers transparent insights into

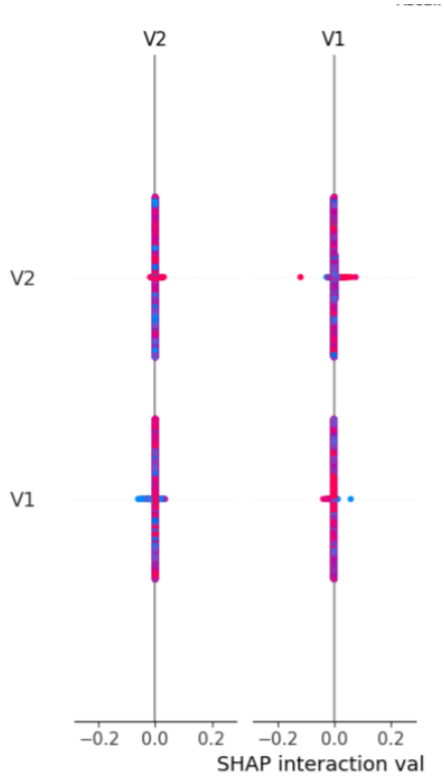

Fig. 4. Box Plot

model decisions, which is essential for real-world cybersecurity applications. The XGBoost model, in particular, displayed strong precision-recall balance, underscoring its suitability for highly imbalanced datasets.

XAI methods proved invaluable in translating complex model outputs into actionable insights. SHAP's global feature importance analysis highlighted key indicators associated with anomalous traffic, while LIME's local explanations allowed for case-by-case validation, helping analysts better understand the factors driving specific anomaly predictions. These explainability tools bridge the gap between automated detection and human interpretability, enabling cybersecurity teams to make more informed, confident responses to potential threats.

However, the study also encountered challenges, notably the computational cost of XAI techniques and the trade-offs between model complexity and interpretability. Future research should explore more efficient XAI methodologies and optimization strategies to enhance scalability in high-throughput, real-time environments.

In conclusion, this work underscores the potential of XAI in cybersecurity, demonstrating that an interpretable, machine-learning-driven anomaly detection system can enhance both the efficacy and accountability of threat detection. This approach lays the foundation for XAI's broader application in cybersecurity, where transparency and explainability are increasingly crucial for building resilient and trustworthy systems.

VII. Acknowledgment

I would like to extend my deepest gratitude to Mr. Brajesh Kumar Sharma, whose mentorship and guidance have been pivotal in the successful completion of this term paper on the application of Explainable AI (XAI) in anomaly detection for cybersecurity. His extensive knowledge and experience in machine learning, cybersecurity, and interpretability in AI have profoundly enriched my understanding of this field. Mr. Sharma's enthusiasm for this subject and his continuous encouragement were instrumental in helping me navigate complex concepts and effectively apply them in the context of anomaly detection.

Throughout the research process, Mr. Sharma provided insightful feedback, challenging me to think critically and refine my approach. His guidance not only helped in structuring the research methodology but also in deepening my understanding of how Explainable AI could bridge the gap between model complexity and interpretability. I am especially grateful for the time he dedicated to reviewing drafts and for his constructive critiques, which strengthened both the technical accuracy and clarity of this work.

I would also like to thank my institution and department for providing access to the resources and tools essential for conducting this research, as well as the wider academic community for their contributions to the fields of XAI and anomaly detection in network security. Their work laid the foundation for this study and offered a wealth of knowledge and inspiration that greatly informed my research.

A special thank you to my colleagues and peers who offered encouragement and engaged in discussions that brought fresh perspectives to my understanding of anomaly detection and cybersecurity. Their support and shared knowledge played a crucial role in maintaining motivation and focus throughout this journey.

Lastly, I would like to acknowledge my family and friends for their unwavering support and encouragement during the entire research process. Their belief in my abilities and their understanding throughout this demanding period provided the balance and inspiration needed to bring this project to fruition.

Thank you to everyone who contributed in ways big and small to the completion of this work. I am truly grateful for your support..

VIII. References

Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research [1]. Towards Anomaly Detection using Explainable AI [2]Artificial Intelligence-Based Approaches for Anomaly Detection [3]A survey on machine learning techniques for cyber security in the last decade [4]Machine learning and cyber security [5]An improved data anomaly detection method based on isolation forest [6]Advancing Cybersecurity with Explainable Artificial Intelligence: A Review of the Latest Research [7]AI Enhanced Cyber Security Methods for Anomaly Detection [8]Generalized isolation forest for anomaly detection [9]Isolation forest based anomaly detection: A systematic literature review [10]

REFERENCES

[1] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," in IEEE Access, vol. 10, pp. 93104-93139, 2022, doi: 10.1109/ACCESS.2022.3204051. keywords: Computer crime;Cyberattack;Computer security;Deep learning;Medical services;Malware;Intrusion detection;Artificial intelligence;Unsolicited e-mail;Information filters;Artificial intelligence;cyber security;deep learning;explanation artificial intelligence;intrusion detection;machine learning;malware detection;spam filtering,

[2] de Oca, E. M. Towards Anomaly Detection using Explainable AI.

[3] Cherukuri, A. K., Ikram, S. T., Li, G., and Liu, X. (2024). Artificial Intelligence-Based Approaches for Anomaly Detection. In Encrypted Network Traffic Analysis (pp. 73-99). Cham: Springer International Publishing.

[4] Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[5] Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., Xu, M. (2020). A survey on machine learning techniques for cyber security in the last decade. IEEE access, 8, 222310-222354.

[6] Parkar, P., Bilimoria, A. (2021, May). A survey on cyber security IDS using ML methods. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 352-360). IEEE.

[7] Lesouple, J., Baudoin, C., Spigai, M., Tourneret, J. Y. (2021). Generalized isolation forest for anomaly detection. Pattern Recognition Letters, 149, 109-119.

[8] P. Ramya, S. V. Babu and G. Venkatesan, "Advancing Cybersecurity with Explainable Artificial Intelligence: A Review of the Latest Research," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 1351-1357, doi: 10.1109/ICIRCA57980.2023.10220797.

[9] Xu, D., Wang, Y., Meng, Y., Zhang, Z. (2017, December). An improved data anomaly detection method based on isolation forest. In 2017 10th international symposium on computational intelligence and design (ISCID) (Vol. 2, pp. 287-291). IEEE.

[10] Al Farizi, W. S., Hidayah, I., Rizal, M. N. (2021, September). Isolation forest based anomaly detection: A systematic literature review. In 2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE) (pp. 118-122). IEEE.