Explainable AI for Anomaly Detection in IoT Networks Using XGBoost

Manjunath P V[1], R Rakshith Kumar[2], Sagar A[3], Thomas Sunil[4], VSaketh Nivesh[5]
[1]*Assistant Professor, Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*
[2,3,4,5] *Student, Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

***Abstract*—This paper introduces a holistic approach for identifying anomalies in Internet of Things (IoT) networks utilizing the robust XGBoost classification model and explainable artificial intelligence (XAI) methods. We leverage SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and a surrogate decision tree to supplement the model's interpretability. The performance of our method is tested on the IoT-23 dataset, which covers a variety of attack vectors as well as benign network traffic. The outcomes illustrate exceptional predictive accuracy as well as substantially improved model transparency, thereby enhancing understanding and confidence in automated systems for network security.**

***Index Terms*—Anomaly Detection, Explainable AI, SHAP, LIME, IoT Security, XGBoost, Surrogate Models, Cybersecurity, Network Intrusion Detection, Machine Learning.**

I. INTRODUCTION

With the huge growth of Internet of Things (IoT) devices, modern networks are significantly more complex and exposed. It is important that anomaly detection within such networks takes place to make them secure. Machine learning algorithms are likely to be extremely precise but uninterpretable. Such black-boxing makes it difficult for security professionals to believe in or comprehend the decision-making procedure of the system. Explainable Artificial Intelligence (XAI) fills this void by providing explanations for the behavior of the model. We introduce in this paper the use of XGBoost for anomaly detection and improve its explainability using SHAP, LIME, and surrogate decision trees. We show that the intersection of the strong prediction and explainability makes a highly useful and effective cybersecurity tool.

II. CONTEXTUAL FRAMEWORK AND RELEVANT LITERATURE

A. Anomaly Detection in IoT

IoT networks are vulnerable to numerous cyber-attacks as they have extensive attack surfaces and limited processing capabilities. Anomaly detection is the process of discovering unusual patterns that could indicate malicious behavior. Machine learning, especially supervised learning techniques, has been used to try to automate this process.

B. Constraints Related to Black-box Models

Though precise, black-box models such as random forests, XGBoost, and neural networks provide little understanding of the processes driving their predictions. This lack of transparency can be highly detrimental, particularly in sensitive domains such as cybersecurity.

C. Explainable AI (XAI)

XAI seeks to make machine learning models more explainable. Methods like SHAP and LIME interpret individual predictions by assigning importance to input features. Surrogate models, like decision trees, approximate complex models using interpretable surrogates.

D. Explainable AI (XAI)

Current research has explored XAI in the context of cybersecurity. Ribeiro et al. brought forth LIME to interpret classifier predictions. Lundberg and Lee introduced SHAP as a single, unified framework for feature attribution. In the IoT context, researchers have utilized these tools to provide insights into anomaly detection systems.

III. DATASET DESCRIPTION

A. IoT-23 Dataset

Stratosphere IPS produced the IoT-23 dataset, which is labeled network traffic of a collection of IoT devices. There is benign and malicious traffic in the dataset across many different attacks such as DDoS, C&C communications, and port scans.

B. Data Preprocessing

Data cleaning was done by removing null values and duplicate features. Categorical features were encoded with label encoders. Normalization was done where necessary. We concentrated on pertinent attack classes for balancing. The combined dataset is generated and saved as the iot23 combined.csv file.

TABLE I. COUNTS OF ATTACK TYPES FOR FILE IOT23_COMBINED.CSV

Label	Count
PartOfAHorizontalPortScan	825939
Okiru	262690
Benign	197809
DDoS	138777
Attack	3915
C&C-HeartBeat	349
C&C-FileDownload	43
C&C-Torii	30
FileDownload	13
C&C-HeartBeat-FileDownload	8
C&C-Mirai	1

C. Equations

To ensure a balance of classes, we sampled 3,500 each of the primary attack classes (for instance, Okiru, DDoS, C&C, PartOfAHorizontalPortScan). The infrequent classes were merged into a new class "Other_Attacks." This so-created dataset was shuffled and stored for model training.

TABLE II. COUNTS OF ATTACK TYPES FOR BALANCED CLASSES

Labels	Count
Attack	3500
Okiru	3500
PartOfAHorizontalPortScan	3500
DDoS	3500
Benign	3500
C&C	3500
Other_Attacks	444

IV. METHODOLOGY

A. Model Architecture

We used XGBoost, a scalable and powerful gradient boosting library. We trained the model on the balanced data with an 80-20 train-test split.

B. Ev0aluation Metrics

We employed accuracy, precision, recall, F1-score, and confusion matrix to measure model performance.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

C. Explainability Framework

a) SHAP: TreeExplainer was used to compute SHAP values for the test set. Global explanations were enabled through feature importance plots.

b) LIME: We found specific test cases and developed localized explanations with LIME. They described the impact of features on individual predictions.

c) Surrogate Decision Tree: A decision tree was trained on the predictions of XGBoost to simulate its reasoning. The tree was graphed and examined.

V. EXPERIMENTAL FRAMEWORK

Experiments were performed in Python with the following libraries: scikit-learn, XGBoost, SHAP, and LIME. The environment was executed on a machine with Intel i7 CPU, 16GB RAM, and Windows/Linux OS.

VI. FINDINGS

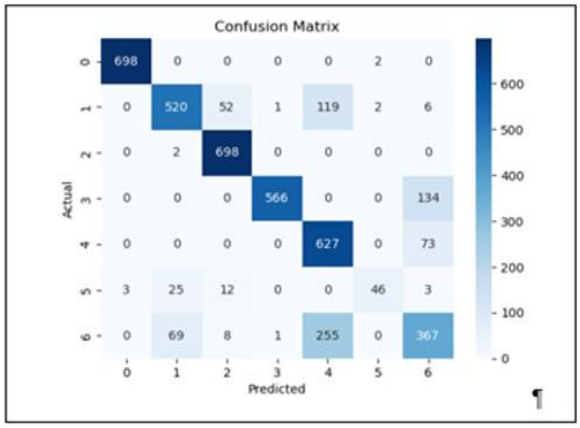
A. XGBoost Performance

The XGBoost classifier had overall accuracy of 82%, precision and recall greater than 97% for all the top classes.

Classification Report:

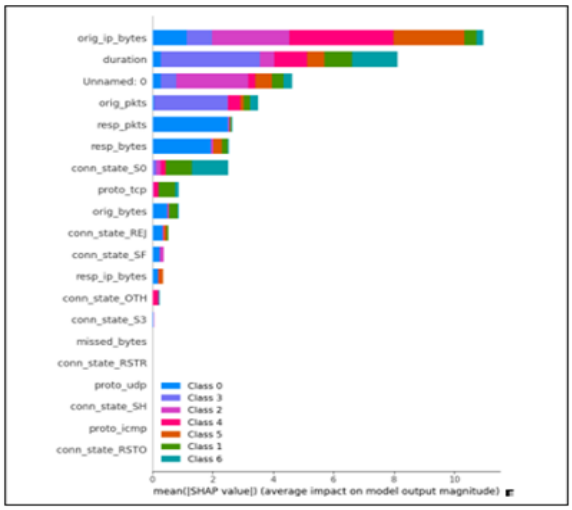
	precision	recall	f1-score	suppo
Attack	0.99	1.00	1.00	70
Benign	0.84	0.74	0.79	70
C&C	0.91	1.00	0.95	70
DDoS	0.97	0.82	0.89	70
Okiru	0.62	0.89	0.74	70
Other_Attacks	0.94	0.51	0.66	8
PartOfAHorizontalPortScan	0.63	0.50	0.56	70
accuracy			0.82	428
macro avg	0.84	0.78	0.80	428
weighted avg	0.83	0.82	0.82	428

The confusion matrix verified low misclassification.



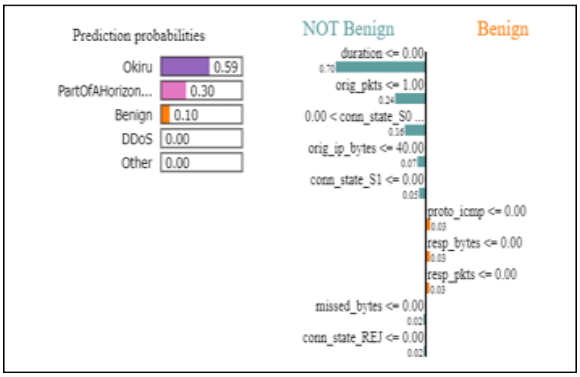
B. SHAP Analysis

Global feature importance plots revealed that packet size, duration, and destination port were the most important features. SHAP values enabled decision boundary interpretation for each class.



C. LIME Explanations

LIME provided explicit explanations at the instance level. The explanations indicated that small variations in features, like byte size, led to different predictions.



D. Surrogate Model Insights

The surrogate decision tree replicated XGBoost behavior with 95% fidelity. It was a convenient tool to track decision paths for more than one class.

VII. EXPERIMENTAL FRAMEWORK

XAI blending with XGBoost enables security analysts to audit and trust model decisions. This, as opposed to black-box systems, makes for improved human-AI cooperation. The surrogate model was also helpful in training non-technical stakeholders.

Challenges are scalability to real-time systems and explaining in terms of changing threat environments. But the modularity of XAI tools provides for future extension.

VIII. FUTURE WORK

Future research will focus on enhancing the effectiveness and practicability of the suggested approach in real-world scenarios. The primary directions will include integrating the system into real-time IDS to verify its efficiency in real-world scenarios and utilizing interpretable deep learning models to improve detection efficiency and interpretability. Additionally, efforts will be directed towards adapting the framework to support unsupervised anomaly detection to enable it to identify emerging threats in the absence of supervised information. Finally, user studies will be carried out to measure human trust and usability in XAI-based IDS systems to confirm that the system aligns to user expectations and working requirements.

IX. CONCLUSION

This paper illustrates how explainable AI methods can greatly increase the transparency and usability of machine learning models in anomaly detection within IoT networks. Our method achieves high accuracy with explainable output, which is an improvement from intelligent, reliable cybersecurity systems.

REFERENCES

[1] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," NeurIPS, 2017.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," KDD, 2016.
[3] A. Ioannou, C. Patsakis, "The IoT-23 Dataset: A Labeled Dataset for IoT Network Intrusion Detection," arXiv:2001.07621, 2020.
[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," KDD, 2016.
[5] A. Holzinger et al., "What do we need to build explainable AI systems for the medical domain?" arXiv:1712.09923, 2017.
[6] L. S. Shapley, "A Value for n-Person Games," Princeton University Press, 1953. [7] J. Kaur et al., "Explainable AI: Survey and Analysis," Information Fusion, vol. 86, 2022.