

Explainable Hybrid AI Framework for Cyber Threat Detection with Anomaly Analysis

More Srushti Satish, Prof. Bangar Abhishek Popat, Prof. Bhosale. Sachin Bajirao, Prof. Karad Vandana

Dept. of Computer Engineering
Jaihind College of engineering, Pune, India

Abstract — Cyber threats such as denial-of-service attacks, scanning, brute-force attempts, and suspicious network activity create risks for users and organizations. Machine-learning-based intrusion detection can help identify patterns in network traffic, but a supervised model may miss attack types that are not represented in its training data. In addition, a prediction without a clear explanation can be difficult for a security analyst to trust. This paper proposes a small, explainable hybrid framework that combines a supervised classifier, anomaly detection, and feature-level explanations. XGBoost is proposed for classifying known traffic patterns, Isolation Forest for flagging unusual records, and SHAP for explaining selected classifier predictions. A Streamlit dashboard is planned to display predictions, anomaly flags, and evaluation results. The framework will be evaluated using a public intrusion-detection dataset such as CIC-IDS2017. The main research question is whether adding anomaly analysis and explanations improves the usefulness of a basic classifier, while keeping false alarms visible. This paper describes a research proposal and evaluation plan; no experimental performance values are claimed before implementation. SHAP provides feature-level explanations to help users understand the factors influencing classifier predictions. A Streamlit dashboard is planned to display classification results, anomaly flags, and evaluation metrics. The framework will be evaluated using the CIC-IDS2017 dataset, with precision, recall, F1-score, and false-positive rate used to assess its performance. The study aims to investigate whether combining supervised classification, anomaly detection, and explainability provides more useful threat analysis than a basic classifier alone. Experimental results will be reported after implementation and testing.

Keywords— Cybersecurity, intrusion detection, machine learning, XGBoost, Isolation Forest, explainable AI, SHAP, anomaly detection.

I. INTRODUCTION

Cybersecurity has become increasingly important due to the growing number and complexity of cyberattacks. Traditional security systems mainly depend on predefined rules and known attack signatures, which can make it difficult to identify new and evolving threats. Artificial Intelligence (AI) and Machine Learning (ML) provide useful techniques for detecting cyber threats and improving network security [1], [2]

Recent research has explored different AI-based approaches for cyber threat detection. Hmimou et al. proposed a multi-agent system using Large Language Models (LLMs) to detect and correlate cybersecurity threats from different sources [1]. Khan et al. reviewed machine learning approaches for cognitive cybersecurity, including threat detection, anomaly detection, and vulnerability analysis [2]

Research has also investigated AI-based methods for detecting cyberattacks in cyber-physical systems and improving cybersecurity protection [3]. AI-enabled threat detection approaches have further highlighted the potential of intelligent techniques to support security monitoring and cyber threat

mitigation [4]. In addition, research on machine-learning-based intrusion detection has examined the importance of incorporating cyber threat intelligence and updating detection models as threats evolve [5]

Despite these developments, challenges remain in detecting suspicious network activities, identifying unusual behaviour, and explaining why a system classifies network traffic as malicious. Therefore, this paper proposes an explainable hybrid AI framework for cyber threat detection. The framework uses XGBoost to classify known attack categories, Isolation Forest to identify anomalous network behaviour, and SHAP to explain the factors influencing model predictions. A dashboard is designed to display detection results and explanations, helping users understand potential cyber threats. The proposed framework will be evaluated using a network intrusion detection dataset and suitable performance metrics.

II. LITERATURE REVIEW

Recent research has explored several artificial intelligence and machine learning techniques for improving cybersecurity and intrusion detection. The selected studies focus on threat

correlation, machine learning-based attack detection, explainable AI, and dynamic cyber threat intelligence.

1. Multi-Agent System for Cybersecurity Threat Detection

Hmimou et al. [1] proposed a multi-agent system for cybersecurity threat detection and correlation using Large Language Models (LLMs). Their work focuses on analysing and correlating security threats using a multi-agent approach. It highlights the importance of combining security information to obtain a better understanding of cyber incidents. The proposed project follows a different approach by focusing on network traffic classification, anomaly detection, and explainability using machine learning techniques.

2. Cognitive Cybersecurity Analysis Using Machine Learning

Khan et al. [2] presented a comprehensive survey of cognitive cybersecurity analysis using machine learning approaches. Their study reviews the role of machine learning in cybersecurity and discusses its applications in threat identification and security analysis. This research provides a foundation for investigating intelligent threat detection techniques. Building on this direction, the proposed framework combines supervised classification and unsupervised anomaly detection to analyse network traffic.

3. AI-Based Cyberattack Detection in Cyber-Physical Systems

Abdullahi et al. [3] investigated and compared AI-based approaches for cyberattack detection in cyber-physical systems. Their work examines the application of artificial intelligence techniques to identify cyberattacks in interconnected systems. This study demonstrates the relevance of AI-based methods in cybersecurity. In comparison, the proposed framework focuses on network traffic analysis using XGBoost and Isolation Forest, with the aim of identifying known traffic classes and flagging unusual network behaviour.

4. AI-Enabled Threat Detection and Cyber Threat Mitigation

Dhanushkodi and Thejas [4] discussed the application of artificial intelligence for advanced threat detection and cyber threat mitigation. Their study explores AI-based approaches for strengthening cybersecurity and highlights relevant developments in intelligent security analysis. The proposed framework extends this direction by incorporating SHAP-based explanations for XGBoost predictions, helping users understand which input features influence the classification results.

5. Dynamic Cyber Threat Intelligence for Intrusion Detection

Lin et al. [5] proposed an approach for integrating Cyber Threat Intelligence (CTI) into machine learning-based intrusion detection through dynamic model updates. Their work addresses the need to adapt detection models as threat information changes. This research highlights the challenge of maintaining effective detection as attack patterns evolve. The current project focuses on classification and anomaly flagging, while dynamic threat intelligence integration is considered a possible future enhancement.

III. RESEARCH GAP

Although existing research investigates AI-based cyberattack detection, threat correlation, explainable security analysis, and dynamic threat intelligence, there remains a practical opportunity to investigate how these capabilities can be combined within a single, interpretable network traffic analysis framework.

The following research gaps are considered in this study:

- **Combined Detection Approach:** The project investigates the integration of supervised classification and unsupervised anomaly detection to analyse known traffic classes and flag potentially unusual network behaviour.
- **Interpretability of Predictions:** The project incorporates SHAP to explain the features influencing XGBoost classification predictions, making the detection process easier to interpret.
- **Identification of Unusual Traffic:** The framework uses Isolation Forest to flag anomalous network traffic that may not fit the patterns learned by the classifier. Such flags require further analysis and do not independently confirm a cyberattack.
- **Practical Result Visualization:** The project provides a Streamlit dashboard to present predicted traffic classes, anomaly flags, and model evaluation metrics in an accessible format.
- **Performance Evaluation:** The framework will be evaluated using the CIC-IDS2017 dataset. Precision, recall, F1-score, and false-positive rate will be used to assess its performance. Comparison with a baseline classifier will help determine whether adding anomaly detection provides measurable benefits.

Review of Related Work

The five papers selected for this review represent different directions in AI-assisted cybersecurity: multi-agent correlation, broad machine-learning surveys, comparisons of AI methods,

AI-enabled threat mitigation, and dynamic threat-intelligence updates.

Proposed System and Methodology

The system is designed as a research prototype for offline analysis of a public dataset. It is not intended to monitor a real organization's network or to take automatic defensive action. The workflow begins with data preparation, then trains two models, generates explanations, and presents the outputs in a dashboard.

Dataset and Preparation

CIC-IDS2017 is a publicly used intrusion-detection dataset that contains labelled network traffic for different benign and attack activities. Before using it, the student should review the dataset licence and documentation, choose a manageable subset if needed, and record the exact files and classes used.

Dataset versions and preprocessing choices can affect results, so they must be described clearly. Preparation steps will include checking column types, identifying missing or infinite values, removing duplicate records only when justified, converting labels into a consistent form, and separating features from the target label. The train/test split should be created before fitting models. Any scaling, encoding, feature selection, or resampling must be learned from the training data only to avoid information leakage. If records from the same flow or capture period can appear in both sets, the split strategy should be reviewed carefully.

XGBoost Classifier

XGBoost is a gradient-boosted decision-tree method. In this project it is proposed as the supervised model because it can learn non-linear relationships among tabular traffic features and can support binary or multi-class classification. The target could be benign versus attack, or a set of attack categories if the selected subset contains enough examples. A simple baseline, such as a decision tree or logistic regression, should be evaluated first so the results have a point of comparison. The model will be trained only on the training partition. Hyperparameters such as tree depth, learning rate, and number of estimators can be tuned using a validation set or cross-validation. The final test set should be kept untouched until the evaluation stage. Class imbalance should be reported because a high accuracy value alone can hide poor detection of rare attack classes.

Isolation Forest Anomaly Detector

Isolation Forest is an unsupervised anomaly-detection algorithm that isolates unusual observations using random

partitions. It can assign an anomaly score or label to a traffic record. The training setup must be stated clearly—for example, whether it is fitted only on the benign training records or on a broader training sample. The threshold or contamination setting should be chosen on training/validation data, not adjusted to improve test results. The anomaly result will be shown separately from the XGBoost class. If a record is flagged by Isolation Forest but classified as benign by XGBoost, the dashboard should label it as “needs review” rather than “confirmed attack.” This wording helps prevent users from treating every unusual record as malicious

IV. EXPLAINABILITY, DASHBOARD, AND EVALUATION

1. SHAP-Based Explanation

SHAP (SHapley Additive exPlanations) can estimate how input features contribute to a model output. For a selected traffic record, the dashboard may display the top positive and negative feature contributions for the predicted class. This can help a reviewer understand which features influenced the classifier, but it does not prove that the model is correct or that a feature caused an attack. The explanation method, output class, and any approximation settings should be documented.

2. Dashboard Design

- Upload or select a prepared CSV file containing the expected traffic features.
 - Display the predicted class and confidence/probability where supported by the model.
 - Show the anomaly flag and score separately from the classifier result.
 - Display a SHAP plot or ranked feature contributions for selected records.
 - Show evaluation metrics and a confusion matrix for the test set.
 - Allow the user to export a results table for the project report.
- The dashboard is a presentation layer. It should not execute network scans, block traffic, or take actions against external systems. For a college demonstration, use a saved dataset and show how the system processes records in a controlled offline environment.

3. Experimental Design

The main experiment should compare at least three configurations: (1) a baseline classifier, (2) XGBoost alone, and (3) XGBoost with the anomaly flag available as a second signal. The third configuration should not silently change the classifier's predictions; instead, report how the additional flag changes review workload and the number of suspicious records surfaced. If a combined decision rule is added, define it before the final test and evaluate it separately.

Risks and Limitations

- Dataset bias: results on a public benchmark may not transfer to current real-world networks.
- Label quality: errors or simplifications in dataset labels can affect training and evaluation.
- Class imbalance: overall accuracy may be misleading when some attacks are rare.
- False positives: anomaly detection can flag legitimate but unusual activity.
- Explainability limits: SHAP describes model behaviour, not the true cause of an event.
- Scope: the proposed prototype is offline and does not provide live packet capture, automated blocking, or validated production security.

How the Research Gap Will Be Addressed

The gap will be addressed through an explicit comparison rather than by simply combining algorithms. First, a baseline and XGBoost classifier will be measured on the same test set. Second, anomaly flags will be applied as an additional review signal. Third, the project will count how many additional attack-labelled records are surfaced and how many benign records are also flagged. Finally, SHAP explanations will be shown for representative true positives, false positives, and false negatives. This makes the trade-off transparent and gives the evaluator evidence to discuss.

V. CONCLUSION

The expected deliverables are a trained supervised model, an anomaly detector, an explanation view, a dashboard, and an evaluation report. These are planned outputs, not results already achieved. After experiments, the student should replace placeholders with measured metrics, screenshots, and a discussion of errors. If the hybrid workflow does not improve detection or produces too many false alarms, that finding should be reported honestly; a negative or mixed result can still be useful when the experiment is reproducible. The proposed framework studies a practical way to combine supervised cyber-threat classification, anomaly analysis, and explainable predictions. Its main value is the evaluation of complementary signals in one accessible workflow. The design deliberately avoids treating every anomaly as an attack and avoids automated mitigation. Future work may investigate dynamic threat-intelligence updates, concept-drift monitoring, multi-source correlation, testing on another dataset, and careful evaluation in a controlled environment.

REFERENCES

1. Y. Hmimou, M. Tabaa, A. Khiat, and Z. Hidila, "A Multi-Agent System for Cybersecurity Threat Detection and Correlation Using Large Language Models," *IEEE Access*, vol. 13, pp. 150199–150215, 2025, doi: 10.1109/ACCESS.2025.3602681.
2. W. Khan, K. Ashoka, M. S. Abdul Razak, M. V. Manoj Kumar, and R. Naseer, "A Comprehensive Survey on Cognitive Cyber Security Analysis Using Machine Learning Approaches," *IEEE Access*, vol. 13, pp. 169314–169326, 2025, doi: 10.1109/ACCESS.2025.3614388.
3. M. Abdullahi et al., "Comparison and Investigation of AI-Based Approaches for Cyberattack Detection in Cyber-Physical Systems," *IEEE Access*, vol. 12, pp. 31988–32004, 2024, doi: 10.1109/ACCESS.2024.3370436.
4. K. Dhanushkodi and S. Thejas, "AI Enabled Threat Detection: Leveraging Artificial Intelligence for Advanced Security and Cyber Threat Mitigation," *IEEE Access*, vol. 12, pp. 173127–173136, 2024, doi: 10.1109/ACCESS.2024.3493957.
5. Y.-D. Lin, Y.-H. Lu, R.-H. Hwang, Y.-C. Lai, D. Sudyana, and W.-B. Lee, "Evolving ML-Based Intrusion Detection: Cyber Threat Intelligence for Dynamic Model Updates," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 3, pp. 605–622, 2025, doi: 10.1109/TMLCN.2025.3564587.
6. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Proc. 4th International Conference on Information Systems Security and Privacy (ICISSP)*, 2018, pp. 108–116. Dataset information: <https://www.unb.ca/cic/datasets/ids-2017.html>.
7. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
8. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. IEEE International Conference on Data Mining*, 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17. [9]
9. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.