AI-Powered Intrusion Detection with SHAP Explainability and Feedback Loop: A Modular Pipeline for Cyber Threat Classification

Raj Kumar Myakala
CVS Health
Washington D.C., USA
ORCID: 0009-0003-0798-708X

Akhil Reddy Jagirapu
University of North Texas
Texas, Usa
ORCID: 0009-0008-1793-6872

Vinithya Reddy Podduturi
World Bank
Washington D.C., USA
ORCID: 0009-0003-1450-4509

Praveen Kumar Nagata
Trine University
Texas, Usa
ORCID: 0009-0008-6550-760X

Sampath Lavudya
Campbellsville University
Texas, Usa
ORCID: 0009-0006-2315-7117

Sanjeev Kumar Pedhapally
University of North Texas
Texas, Usa
ORCID: 0009-0002-7551-5474

***Abstract*—Intrusion Detection Systems (IDS) are fundamental to protecting digital infrastructure against increasingly sophisticated cyber threats. While conventional machine learning approaches for IDS often suffer from limitations including static decision-making pipelines, inherent model opacity, and difficulty in adapting to novel attack vectors, this work introduces a novel, modular, interpretable, and adaptive AI pipeline for robust threat classification. Leveraging the well-established CICIDS2017 benchmark dataset, we construct an end-to-end framework that integrates high-performance Random Forest and XGBoost classifiers with SHAP (SHapley Additive exPlanations) for comprehensive feature attribution. Crucially, the pipeline incorporates a feedback-driven closed-loop mechanism enabling dynamic model retraining and adaptation to evolving threat landscapes. This integrated architecture not only achieves exceptionally high classification performance, with the XGBoost model yielding 99.96% accuracy and a 0.92 Macro F1-score across 15 diverse attack types, but also significantly enhances the transparency, maintainability, and real-world adaptability required for modern security operations. We provide detailed insights into model behavior and data characteristics through extensive SHAP analysis, t-SNE clustering visualizations, and a comprehensive class-wise performance breakdown, demonstrating the system's effectiveness in handling significant class imbalance and improving generalization. The resulting pipeline is engineered to be deployable, inherently interpretable, and ready for integration into production-scale cybersecurity environments, offering a significant step towards trustworthy and resilient AI-powered IDS.**

***Index Terms*—Intrusion Detection System, Cybersecurity, Explainable AI, SHAP, Feedback Loop, XGBoost, Random Forest, CICIDS2017, Threat Classification, t-SNE, Adaptive Systems**

I. Introduction

The escalating scale, complexity, and velocity of **cyber-attacks** necessitate advanced **intrusion detection systems** (IDS) as a cornerstone of secure network infrastructure. While machine learning has shown significant promise in cybersecurity, many **ML**-based IDS approaches are hindered by limitations: reliance on static models that fail to adapt to novel threats, "black box" decision-making, and rigid, non-adaptive pipelines [2], [3]. Addressing these gaps is critical for developing effective and trustworthy solutions.

This paper presents a **novel, modular, and adaptive AI-driven framework for intrusion detection** that integrates high-performance multiclass classification, actionable explainability, and continuous adaptation. Our key contributions are:

1) An end-to-end modular ML pipeline with robust preprocessing, feature scaling, and multiclass classification for network threat detection.
2) SHAP-based explainability [1] providing transparent, per-instance insights to support analyst decision-making.
3) A feedback loop for dynamic model updates and retraining from real-time or expert-labeled data, maintaining relevance against evolving threats [4].

Using the **CICIDS2017** dataset as a benchmark, our XGBoost model achieves **99.96% accuracy** and a **0.92 Macro F1-score**. We compare performance with Random Forest and illustrate model behavior via SHAP summaries and t-SNE clustering.

By combining accuracy, interpretability, and adaptability, our framework offers a practical, scalable, and trustworthy IDS for modern cybersecurity environments, demonstrating that human-in-the-loop feedback and explainability are essential for long-term reliability.

II. Related Work

Machine learning has significantly advanced IDS research, but static deployment, opaque decision-making, and limited adaptability remain key challenges. Integrating explainability and adaptability is increasingly recognized as vital [2].

For interpretability, Gaspar et al. [1] applied SHAP and LIME to analyze multi-layer perceptron outputs, while Ludwig [3] used SHAP visualizations for network traffic features.

These improve transparency but are often isolated from adaptive system design.

Adaptation-focused works, such as Roshan and Zafar [4], explored **Kernel SHAP** for optimizing anomaly detection and hinted at retrainable pipelines. However, these approaches rarely combine robust interpretability with a complete, modular architecture for continuous deployment.

Thus, despite progress in classification, explainability, and adaptability, a gap remains for a unified, inherently interpretable, and operationally adaptive AI pipeline for IDS. Our framework addresses this by integrating high-accuracy classifiers, comprehensive **SHAP** explanations, and a feedback loop for sustained performance in dynamic threat environments.

III. Methodology

This section details the dataset, the architecture of our proposed AI-driven intrusion detection pipeline, and the applied preprocessing, feature analysis, and model training methodologies.

A. Dataset and Preprocessing

We used the publicly available **CICIDS2017** dataset [5] from the Canadian Institute for **Cybersecurity (CIC)** at the University of New Brunswick, a benchmark simulating realistic network traffic with both benign activity and diverse modern attack types (e.g., DDoS, brute force, infiltration, web attacks).

The dataset, provided as daily CSV logs, was combined into a unified set with 5,950,088 rows and 86 features. Packet-level statistics **(e.g., Flow Duration, Total Fwd Packets, Packet Length Std)** formed the core features. Non-numeric identifiers (*Flow ID*, *Source/Destination IP*, *Timestamp*) were removed. Missing and infinite values were replaced, and rows with unresolved NaNs were dropped, yielding 80 numeric features. The target variable *Label* contains 15 classes with significant imbalance, dominated by the *BENIGN* class.

Labels were integer-encoded, and features standardized (zero mean, unit variance) using StandardScaler to ensure balanced scaling for model training.

B. System Architecture

Our modular IDS pipeline (Figure 1) consists of data ingestion, preprocessing, feature scaling, model inference, explainability generation, and adaptive retraining.

After preprocessing and scaling, the **Trained Model** (Random Forest or XGBoost) produces a predicted label and confidence score. The **SHAP Module** generates per-instance feature attributions. The **Feedback Loop** allows analysts or automated systems to flag misclassifications, which are stored for periodic retraining that are enabling adaptation to emerging threats without redesigning the pipeline.

C. Feature Engineering and Selection

We applied **t-SNE** to visualize sample flow clustering and Pearson correlation analysis to detect highly correlated features ($r > 0.95$). While these correlations suggest possible dimensionality reduction, this study retained all 80 features to assess full-set performance.

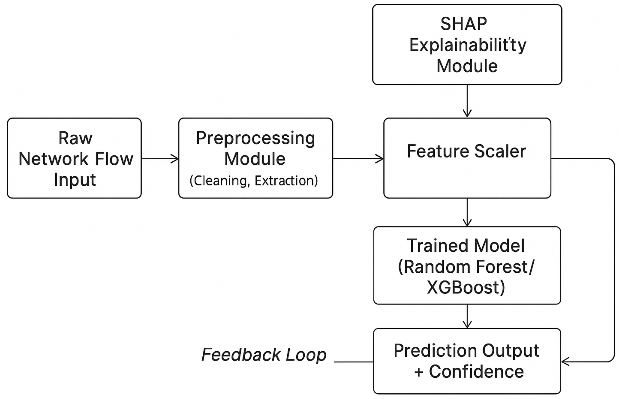


Fig. 1. End-to-end AI-based intrusion detection system with SHAP explainability and feedback-driven retraining loop.

SHAP values quantify feature contributions to each prediction. Global summaries reveal generally influential features **(e.g., *Flow Duration*, *Packet Length Std*, *Destination Port*)**, while individual plots assist analysts in validating model outputs, especially for anomalous traffic.

D. Model Design and Training

Two ensemble algorithms were evaluated: Random Forest and XGBoost, selected for strong tabular-data performance and SHAP compatibility. An 80/20 stratified train-test split preserved class ratios.

XGBoost was configured with 100 estimators, learning rate 0.1, and max depth 6. Random Forest used 100 trees with default entropy splitting. Hyperparameter tuning was not the focus of this initial study. Class imbalance was addressed through macro-averaged Precision, Recall, and F1-score evaluation rather than explicit resampling.

During inference, both models output predicted class and confidence. The SHAP module runs in parallel, and flagged flows in the feedback loop are incorporated into later retraining cycles to improve detection of challenging or **rare attack types**.

IV. Experiments and Results

This section presents the experimental setup, evaluation methodology, and the results obtained from evaluating our proposed AI-driven intrusion detection pipeline. We analyze the characteristics of the **CICIDS2017** dataset, evaluate the performance of the trained models, and utilize SHAP and t-SNE to provide insights into model behavior and data structure.

A. Dataset Characteristics Analysis

Understanding the underlying data distribution is crucial for interpreting model performance, especially in **cybersecurity** datasets known for severe class imbalance. We analyzed the composition and feature relationships within the processed **CICIDS2017** dataset.

Table I provides a summary of the dataset's key statistics. With nearly 6 million rows and 15 unique classes, the dataset offers a comprehensive representation of network traffic. However, it exhibits extreme class imbalance, notably the overwhelming dominance of the BENIGN class (over 2.2 million samples) compared to the most minority class, Heartbleed (only 11 samples). This results in an **imbalance ratio exceeding 200,000:1**, posing a significant challenge for training models to detect rare attack types.

TABLE I
DATASET SUMMARY

Metric	Value
Total Rows	5,950,088
Total Classes	15
Majority Class (BENIGN)	2,271,320 samples
Minority Class (Heartbleed)	11 samples
Imbalance Ratio	$> 200,000 : 1$

The class distribution is further visualized to highlight this disparity. Figure 2 shows a heatmap of flow counts per class, and Figure 3 provides a bar chart, both clearly illustrating the severe underrepresentation of most attack classes relative to BENIGN traffic.

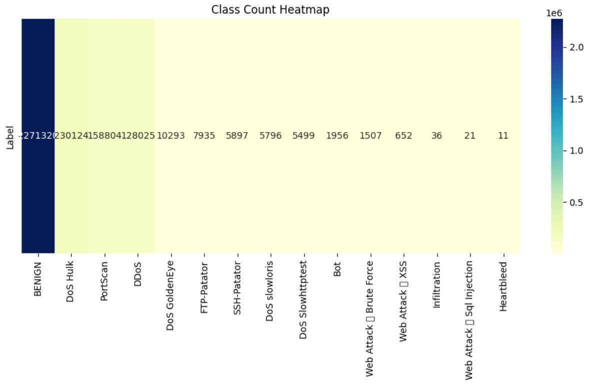


Fig. 2. Class Count Heatmap (Log-scaled colorbar for imbalance visualization)

To understand potential redundancy among features, we computed a Pearson correlation heatmap for the 80 numerical features used in training. As depicted in Figure 4, several features exhibit high pairwise correlation where r > 0.95, such as ***Total Length of Bwd Packets***, ***Subflow Bwd Bytes***, and *Total Backward Packets*.

While this suggests opportunities for dimensionality reduction, the models presented in this paper were trained on the full 80 features.

B. Model Performance Evaluation

To assess the effectiveness of the **Random Forest** and **XGBoost classifiers** within our pipeline, we evaluated their performance on the unseen **20%** test set using a comprehensive set of classification metrics. Given the significant class imbalance, we focus on Precision, Recall, and F1-score, particularly emphasizing macro-averaged metrics which treat all classes equally, thus providing a more reliable indicator of performance on minority classes compared to weighted averages or overall accuracy alone. Overall Accuracy is also reported as a general performance indicator.

Confusion matrices provide a detailed breakdown of correct and incorrect classifications for each class. Figure 5 shows the confusion matrix for the Random Forest classifier, and Figure 6 shows the confusion matrix for the XGBoost classifier. These matrices visually highlight where misclassifications occur, allowing for identification of classes that are **easily confused or poorly detected**.

Detailed classification reports for both models, including precision, recall, and F1-score for each of the 15 classes, are presented in Table II and Table III.

TABLE II
RANDOM FOREST: CLASSIFICATION METRICS SUMMARY

Class	Precision	Recall	F1-score
BENIGN	1.00	1.00	1.00
Bot	0.98	0.95	0.97
DDoS	1.00	1.00	1.00
DoS Hulk	1.00	1.00	1.00
PortScan	1.00	1.00	1.00
...	...	...	...
Accuracy	0.9993		
Macro Avg	0.93	0.88	0.90

TABLE III
XGBOOST: CLASSIFICATION METRICS SUMMARY

Class	Precision	Recall	F1-score
BENIGN	1.00	1.00	1.00
Bot	0.97	0.99	0.98
DDoS	1.00	1.00	1.00
DoS GoldenEye	1.00	1.00	1.00
DoS Hulk	1.00	1.00	1.00
DoS Slowhttptest	0.99	0.99	0.99
DoS slowloris	1.00	1.00	1.00
FTP-Patator	1.00	1.00	1.00
Heartbleed	1.00	1.00	1.00
Infiltration	1.00	0.71	0.83
PortScan	1.00	1.00	1.00
SSH-Patator	1.00	1.00	1.00
Web Attack – Brute Force	0.87	0.86	0.87
Web Attack – Sql Injection	1.00	0.25	0.40
Web Attack – XSS	0.72	0.75	0.74
Accuracy	1.00		
Macro Avg	0.97	0.90	0.92

A summary comparison of the overall performance metrics is presented in Table IV.

As shown, both models achieve exceptionally **high overall accuracy (above 99.9%)**. XGBoost slightly outperforms Random Forest in both overall accuracy (0.9996 vs 0.9993) and, more significantly for this imbalanced dataset, in Macro F1-score **(0.92 vs 0.90)**. The higher Macro F1-score for XGBoost

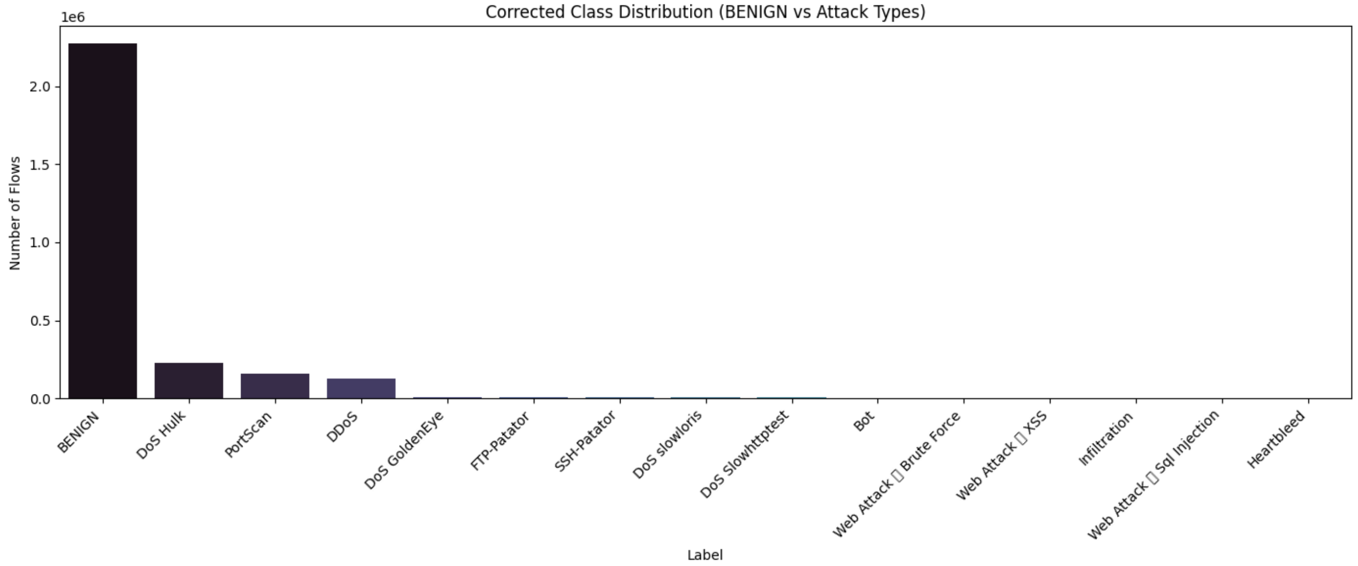


Fig. 3. Corrected Class Distribution: BENIGN vs. Attack Types

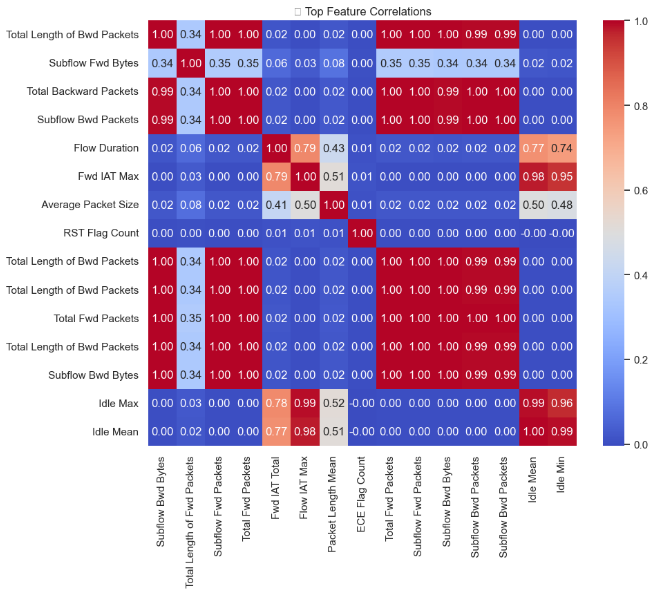


Fig. 4. Top Feature Correlation Heatmap (Pearson Coefficients)

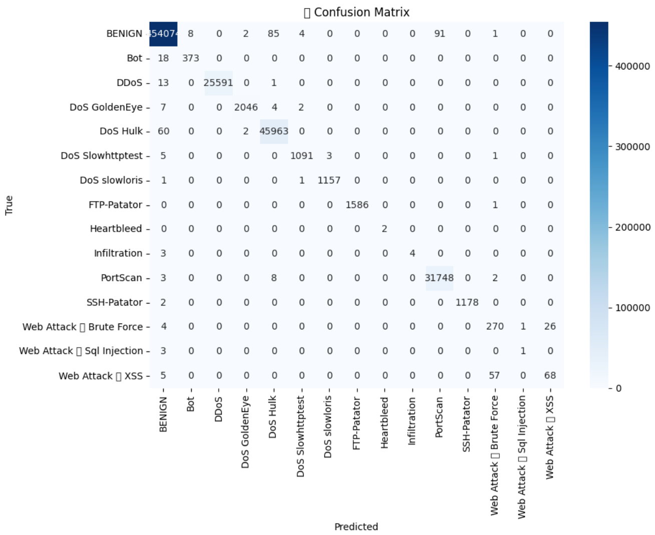


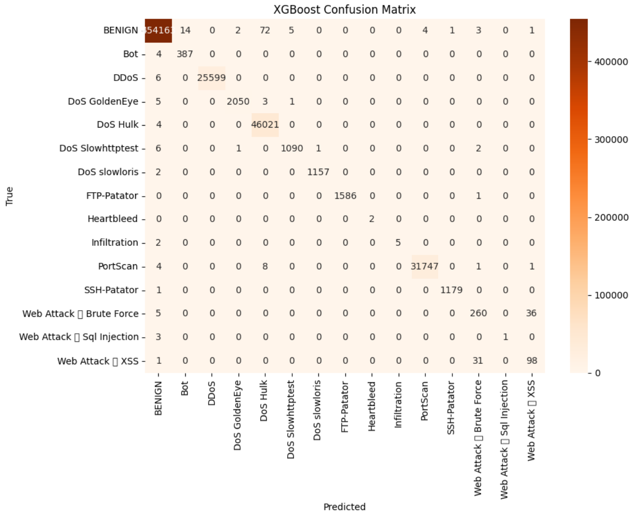
Fig. 5. Confusion Matrix – Random Forest Classifier

TABLE IV
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Macro F1-score
Random Forest	0.9993	0.90
XGBoost	0.9996	0.92

indicates its better balanced performance across all classes, including the minority ones.

Analyzing the detailed classification reports (Table II and Table III), both models achieve perfect or near-perfect scores for many attack types and the majority BENIGN class. However, challenges remain with certain minority classes. For instance, XGBoost shows lower Recall and F1-score for 'Infiltration' (0.71 Recall, 0.83 F1), 'Web Attack – Sql Injection' (0.25 Recall, 0.40 F1), and **'Web Attack – XSS'** (0.75 Recall, 0.74 F1). The low recall for **'Web Attack – Sql Injection'** is particularly notable, indicating that the model misses a significant portion of these instances. These results highlight the inherent difficulty in detecting extremely rare attack types, even with powerful classifiers, and underscore the potential value of the feedback loop mechanism to specifically target and improve detection for such challenging cases in operational deployment.



	BENIGN	Bot	DDoS	DoS GoldenEye	DoS Hulk	DoS Slowhttptest	DoS slowloris	FTP-Patator	Heartbleed	Infiltration	PortScan	SSH-Patator	Web Attack □ Brute Force	Web Attack □ Sql Injection	Web Attack □ XSS
BENIGN	454161	14	0	2	72	5	0	0	0	0	4	1	3	0	1
Bot	4	387	0	0	0	0	0	0	0	0	0	0	0	0	0
DDoS	6	0	25599	0	0	0	0	0	0	0	0	0	0	0	0
DoS GoldenEye	5	0	0	2050	3	1	0	0	0	0	0	0	0	0	0
DoS Hulk	4	0	0	0	46021	0	0	0	0	0	0	0	0	0	0
DoS Slowhttptest	6	0	0	1	0	1090	1	0	0	0	0	0	2	0	0
DoS slowloris	2	0	0	0	0	0	1157	0	0	0	0	0	0	0	0
FTP-Patator	0	0	0	0	0	0	0	1586	0	0	0	0	1	0	0
Heartbleed	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Infiltration	2	0	0	0	0	0	0	0	0	5	0	0	0	0	0
PortScan	4	0	0	0	8	0	0	0	0	0	31747	0	1	0	1
SSH-Patator	1	0	0	0	0	0	0	0	0	0	0	1179	0	0	0
Web Attack □ Brute Force	5	0	0	0	0	0	0	0	0	0	0	0	260	0	36
Web Attack □ Sql Injection	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Web Attack □ XSS	1	0	0	0	0	0	0	0	0	0	0	0	31	0	98

Fig. 6. Confusion Matrix – XGBoost Classifier

C. Model Interpretability and Data Visualization

To provide transparency into model decision-making and visualize the data structure, we employed SHAP analysis and t-SNE dimensionality reduction.

SHAP (SHapley Additive exPlanations) was used to quantify the contribution of each feature to the model's output. Figure 7 presents a global **SHAP summary** plot across all classes for the XGBoost model, illustrating the features with the highest average impact on the model's predictions.

Globally influential features identified by SHAP include *Destination Port*, *Packet Length Variance*, and *Average Packet Size*. Understanding these global feature importances helps in grasping which traffic characteristics are generally most indicative of malicious activity.

These class-specific plots provide actionable insights for security analysts. For example, analyzing the SHAP values for a predicted attack instance helps an analyst understand **why** the model flagged that specific flow as malicious based on its feature values, facilitating faster investigation and response. This interpretability is a key benefit for deploying AI in security-critical environments.

To visually assess the inherent separability of different traffic classes in a reduced dimension, we applied t-SNE dimensionality reduction to a sample of **5,000** flows from the dataset. Figure 8 presents the t-SNE plot, which reveals distinct clusters corresponding to BENIGN traffic and several prominent attack types.

The clear clustering observed in the **t-SNE visualization** supports the effectiveness of using these network flow features for classification and provides a visual confirmation of the underlying structure that the machine learning models exploit to distinguish between different traffic types.

D. Precision-Recall Curve Analysis

Given the highly imbalanced nature of the **CICIDS2017** dataset with majority classes like *BENIGN* which are dominating the use of traditional accuracy metrics alone can be misleading. To provide a more informative evaluation of model performance, especially on minority attack types, we computed the macro-averaged Precision-Recall (PR) curves for both Random Forest and XGBoost classifiers.

The XGBoost model demonstrates a noticeable advantage in handling class imbalance more effectively. Specifically, the macro-averaged area under the Precision-Recall curve (AUC-PR) for XGBoost was **0.9692**, compared to **0.9492** for Random Forest.

This improvement in macro-AUC suggests that XGBoost offers better average performance across all 15 classes, in-

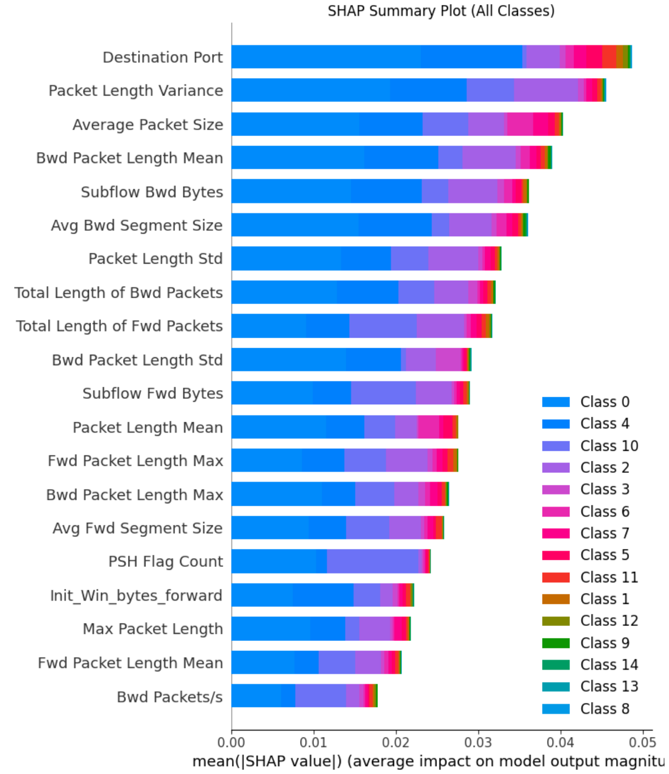


Fig. 7. Global SHAP Summary Plot (All Classes)

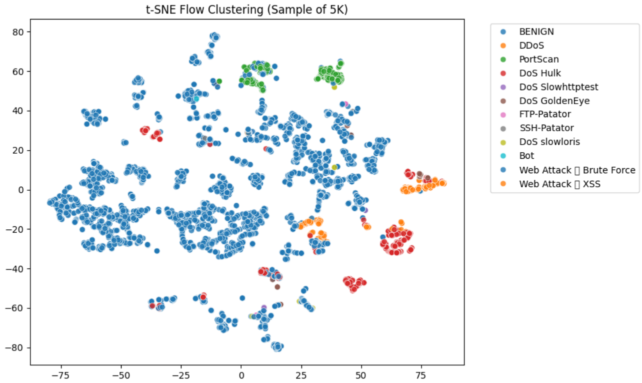


Fig. 8. t-SNE Flow Clustering (Sample of 5K)

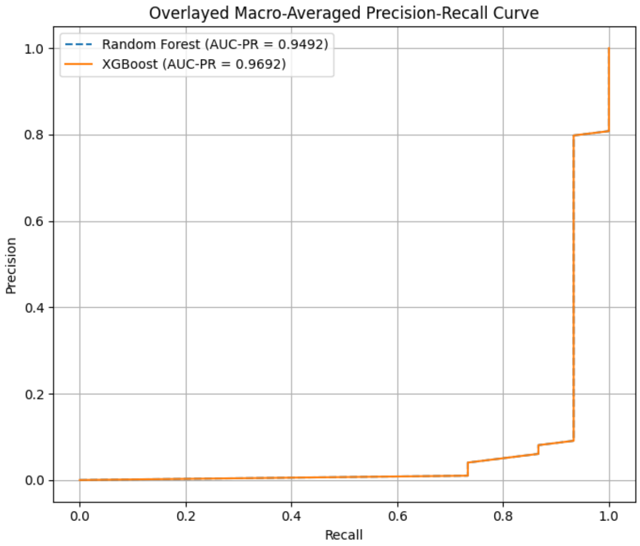


Fig. 9. Overlayed Macro-Averaged Precision-Recall Curve — Random Forest vs XGBoost

cluding rare attack types like **Heartbleed** and **Infiltration**. These findings reinforce the model's robustness in real-world intrusion detection settings where detecting minority threats is crucial.

V. Discussion

The results on **CICIDS2017** validate the effectiveness of our modular AI-driven pipeline for detecting a wide range of network intrusions. XGBoost achieved the **highest overall accuracy (0.9996)** and **Macro F1-score (0.92)**, showing better balanced performance across all classes compared to Random Forest, particularly under severe class imbalance.

Minority classes such as **Infiltration**, **Web Attack – SQL Injection**, and **Web Attack – XSS** remain challenging due to their rarity. The integrated feedback loop is designed to address this by enabling analysts to flag misclassifications for targeted retraining, improving detection over time.

SHAP-based explainability provides transparency, revealing both global and class-specific influential features (e.g., **Flow Duration**, **Packet Length Variance**, **Destination Port**). This aids in alert verification, security decision-making, and trust-building, and can guide feedback loop prioritization. t-SNE visualizations further confirmed the separability of benign and attack traffic clusters, supporting the models' discriminative capabilities.

From a deployment perspective, the pipeline's modular design, reliance on open-source tools **(XGBoost, SHAP, scikit-learn)**, and feedback loop enhance portability, scalability, and adaptability to evolving threats. However, limitations include the dataset's 2017 origin, ongoing class imbalance challenges, and the need for engineering optimizations, such as low-latency processing, model compression, and integration with SIEM platforms which are used for production-scale use.

VI. Conclusion and Future Work

We introduced a modular AI-driven IDS framework that combines high classification accuracy, interpretability, and adaptability. Using CICIDS2017, the XGBoost-based system achieved **99.96% accuracy** and a **0.92 Macro** F1-score across **15 traffic classes** despite severe imbalance. Key innovations include a feedback loop for analyst-driven retraining, SHAP-based explanations for both global and rare-class insights, and validation via macro-averaged PR curves and t-SNE clustering.

Future work will focus on:

- **Real-time Streaming:** Low-latency intrusion detection on live data streams.
- **Continual Learning:** Seamless model updates to address zero-day threats without full retraining.
- **Adversarial Robustness:** Protection against evasion tactics targeting ML-based IDS.
- **Production Deployment:** Evaluation in live environments with SIEM integration, optimizing latency, throughput, and operational efficiency.

These directions aim to transition this framework from research to production, delivering resilient, interpretable, and adaptive IDS solutions for modern cybersecurity landscapes.

Acknowledgment

The authors would like to thank the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick for providing the CICIDS2017 dataset used in this study. We also express our gratitude to our respective institutions such as CVS Health, the World Bank, and affiliated universities for their support and encouragement throughout this research.

Special thanks to our colleagues and peer reviewers who offered valuable feedback on model selection, data visualization, and explainability techniques. Their insights significantly improved the quality and clarity of this work.

This research was conducted independently and did not receive specific funding from any agency in the public, commercial, or not-for-profit sectors.

References

[1] D. Gaspar, P. Silva, and C. Silva, ”Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron,” in *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, 2021.

[2] V. Zibi and M. I. C. Obagbuwa, ”A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems,” *Journal of Cybersecurity*, vol. 9, no. 1, 2023.

[3] S. Ludwig, ”Explainability of Cybersecurity Threats Data Using SHAP,” in *2022 IEEE International Conference on Big Data (Big Data)*, 2022.

[4] K. Roshan and A. Zafar, ”Using Kernel SHAP XAI Method to Optimize the Network Anomaly Detection Model,” in *2023 International Journal of Information Security Science*, vol. 12, no. 1, 2023.

[5] Canadian Institute for Cybersecurity, ”Intrusion Detection Evaluation Dataset (CICIDS2017),” University of New Brunswick. Available: https://www.unb.ca/cic/datasets/ids-2017.html