Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

Improving Lateral-Movement Intrusion Detection in Virtualized Networks using SHAP Feature Selection, SMOTE, and a Voting Ensemble Classifier

Avin Maulana*[1], Syaiful Anam[2], Hilmi Aziz Bukhori[3]

[1,2,3]Mathematics Department, Brawijaya University, Indonesia

Email: [1]avin_maulana@ub.ac.id

Received : Aug 4, 2025; Revised : Aug 5, 2025; Accepted : Aug 24, 2025; Published : Aug 25, 2025

Abstract

Modern virtualized networks, such as those using VXLAN (Virtual eXtensible LAN), generate heavy east–west traffic, which can conceal the lateral movement of attackers. Detecting such infiltration attacks is challenging due to overlay encapsulation (e.g., VXLAN) and flat subnet architectures create blind spots for traditional IDS. This study aims to evaluate a robust methodology for addressing class imbalance in intrusion detection by integrating SHAP-driven feature selection with SMOTE in a voting ensemble. We conducted an ablation study on the CICIDS2017 Thursday-WorkingHours-Afternoon-Infiltration subset, which is highly imbalanced (36 infiltration flows vs. 288,566 benign flows), varying SHAP feature sets (Top-5 vs. Top-30), classification thresholds $\boldsymbol{\theta} \in (\mathbf{0.3}, \mathbf{0.5}, \mathbf{0.7})$, and SMOTE (Synthetic Minority Over-sampling Technique) balancing. The ensemble combined XGBoost, Random Forest, and Logistic Regression, and was evaluated with ROC-AUC, precision, recall, and F1-score. Results indicate that using more SHAP-important features improves ROC-AUC and recall, while SMOTE substantially enhances minority-class detection. The best configuration is Top-30 SHAP features with SMOTE at $\boldsymbol{\theta} = \mathbf{0.7}$, achieved ROC-AUC = 0.976 and F1-score = 0.78, whereas using fewer features or omitting SMOTE significantly reduced recall and F1-score. This synergy of interpretable feature selection and synthetic oversampling establishes a practical methodology for intrusion detection in highly imbalanced, modern virtualized environments. The novelty lies in demonstrating that SHAP + SMOTE integration yields both transparency and resilience, directly addressing encapsulation challenges in detecting stealthy lateral movement.

Keywords: *CICIDS2017, Class Imbalance, Intrusion Detection, Lateral Movement, SMOTE, Voting Ensemble*

1. INTRODUCTION

For decades, the Virtual Local Area Network (VLAN) was the standard for logical network segmentation [1]. However, the rise of large-scale virtualization and cloud services exposed its inherent limitations, most notably a scalability cap of 4,094 unique segments. To overcome these challenges, the industry developed Virtual Extensible LAN (VXLAN), a network virtualization technology that decouples the logical network (the overlay) from the physical network (the underlay) [1], [2]. While VXLAN enables scalable virtualization, it also introduces a security weakness: east-west traffic between virtual machines can be encapsulated and bypass centralized inspection, creating blind spots for intrusion detection.

Once an adversary has breached the network perimeter, the real attack begins. The initial point of entry serves as a beachhead from which the attacker explores the internal network, expands their access, and moves methodically towards their true target [3]. This process is known as lateral movement. It is a phase characterized by stealth and the deliberate use of legitimate system tools to evade detection, which in turn creates a significant data science challenge. Lateral movement involves internal compromises as attackers move within the network [4]; tracking this requires focusing on east-west traffic rather than the usual client-server (north-south) flows.

Intrusion Detection Systems (IDS) are vital for monitoring network traffic and flagging malicious actions, yet traditional signature-based approaches struggle with adaptability and produce high false positives [5], [6]. Machine learning and ensemble methods have emerged as promising alternatives, capable of learning complex traffic patterns and improving generalization [7]. Recent advances show that hybrid IDS, combining explainable AI (e.g., SHAP for feature selection) with resampling techniques like SMOTE, have achieved substantial improvements in imbalanced settings such as IoT networks [14], [17]. These results underscore the importance of interpretable and balanced IDS models that can both detect rare attacks and provide transparent explanations.

Although techniques such as micro-segmentation have been proposed, VXLAN overlays often still form large flat subnets where perimeter IDS sensors cannot observe internal flows. This makes the accurate detection of infiltration attacks, which simulate lateral movement, particularly challenging in modern virtualized environments.

To address this challenge, we investigate an ensemble IDS using explainable feature selection. We use the Canadian Institute for Cybersecurity Intrusion Detection Systems (CIC-IDS) 2017 dataset [8], which provides realistic network traffic with 80 flow features covering normal behavior and contemporary attacks [8], [9], [10] (specifically the *Thursday-Afternoon-Infiltration* file) because it contains realistic traffic and a small number of infiltration attacks (36 malicious flows vs 288,566 benign, highlighting class imbalance. CIC-IDS2017 is widely used in IDS research, so results generalize to practical scenarios.

Our proposed classifier is a Voting ensemble of XGBoost (eXtreme Gradient Boosting), Random Forest, and Logistic Regression, an approach known to leverage complementary strengths of different models. We apply SHAP (SHapley Additive exPlanations) to rank features by importance. We then evaluate two selection strategies: using only the Top-5 features vs Top-30 features. Additionally, we test different SHAP-importance thresholds (0.3, 0.5, 0.7) to see how strict feature cuts affect performance. To mitigate the severe class imbalance, we also evaluate scenarios with and without SMOTE oversampling. We focus on key metrics (ROC-AUC, precision, recall, and especially F1-score, which balances the last two). Prior work stresses that in highly imbalanced IDS settings, "precision and recall are more essential metrics" than raw accuracy; in fact, the *F1 score* is often treated as the primary performance measure.

Unlike prior studies that focus only on feature selection or only on resampling, we systematically combine both to assess their synergy in lateral-movement detection. On CICIDS2017 (infiltration), our best setting (Top-30 + SMOTE) voting ensemble yields high-precision, low-false-alarm alerts while retaining competitive recall in highly imbalanced, virtualized east–west traffic, which directly lowers analyst workload in real-time SOC operations while retaining actionable recall.

The contributions of this work are threefold: (1) a systematic ablation of SHAP-driven feature selection in an IDS context, (2) quantification of SMOTE's effect on detecting rare infiltration flows, and (3) empirical evidence that the synergy of SHAP + SMOTE significantly improves recall and F1-score while maintaining interpretability. This positions our study at the intersection of interpretable machine learning and cybersecurity, addressing blind spots in modern virtualized environments.

2. RESEARCH METHOD

2.1. Dataset and Preprocessing

We use the Canadian Institute for Cybersecurity (CIC) CICIDS2017 [8] dataset, focusing on the *Thursday-WorkingHours-Afternoon-Infiltration* capture. This subset is highly imbalanced, containing only 36 infiltration attack flows among approximately 288,566 benign flows, resulting in a positive-class prevalence of <0.01%. Such severe imbalance can bias classifiers toward always predicting the majority class [9], [10], [11].

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

To ensure high-quality inputs, we first remove duplicate entries, missing values (NaN), and infinite values. We then filter only the BENIGN and Infiltration labels and encode them as binary targets:

$$y = \begin{cases} 1, & Infiltration \\ 0, & BENIGN \end{cases}$$

This clean dataset is stratified into 70% training and 30% testing sets to preserve the rare-class distribution. Feature columns are normalized using Min–Max scaling, as required for combining tree-based and linear models in our ensemble.

2.2. Feature Selection with SHAP (SHapley Additive exPlanations)

CIC IDS 2017 Dataset, with its reach feature, having 80 features. To improve interpretability and reduce feature dimensionality, we apply SHAP (SHapley Additive exPlanations) for feature importance estimation. SHAP computes Shapley values for each feature, which quantify its marginal contribution to the model output considering all possible feature combinations. Unlike univariate ranking methods, SHAP inherently accounts for feature interactions, making it suitable for network traffic analysis with correlated statistics [11], [12], [13], [14]

SHAP is based on Shapley values from cooperative game theory, which ensures a fair distribution of feature importance by considering all possible feature combinations. This provides a comprehensive view of feature contributions, unlike methods that may overlook interactions or dependencies between features SHAP provides an interpretable ranking of features by how much they contribute to model outputs [13]. The method offers a clear and interpretable explanation of model predictions, which is crucial for domains like finance and healthcare, where understanding model decisions is as important as accuracy [12]. As mentioned in [15], SHAP-based feature selection helps in identifying the most relevant features, improving model interpretability and performance. Using more features generally increases ROC-AUC and recall, indicating better detection capabilities

We employ TreeSHAP (model-specific SHAP for tree ensembles) on a baseline XGBoost to estimate per-feature contributions and derive global importance by aggregating absolute Shapley values across samples. Formally, the Shapley value for feature j is shown in equation (1):

$$\varphi_j(f,x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)}{|F|!} \left(f_x(S \cup \{j\}) - f_x(S)\right) \quad (1)$$

Where F is full feature set and $f_x(S)$ is the model output using subset S. We compute the global ranking via $\mathrm{GI}(j) = \mathbb{E}_x\left[\left|\varphi_j(f,x)\right|\right]$ and construct Top-5 and Top-30 subsets for ablation. These subsets are then used in our ablation experiments to evaluate the trade-off between model simplicity and detection performance. In general, using more SHAP-selected features is expected to increase recall and ROC-AUC, particularly for rare-class infiltration flows.

2.3. Handling Class Imbalance with SMOTE

To address class imbalance, we experiment with SMOTE oversampling. SMOTE has been used with tree-boosting models like LightGBM, XGBoost, and CatBoost, resulting in high accuracy and F1 scores, averaging around 99%. This demonstrates the technique's effectiveness in improving the predictive capabilities of these models on imbalanced dataset [16]. In the context of IoT networks, SMOTE has been used to balance highly imbalanced datasets, resulting in improved accuracy of over 90% with machine learning classifiers like K-Nearest Neighbors, Decision Trees, and Random Forests. This highlights SMOTE's role in effectively mitigating attacks, especially in attack detection in network domain [17]. In our study, we apply SMOTE to the training data so that the infiltration class is up

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

sampled to parity (or a high ratio) with the majority [18]. The rationale is that SMOTE can increase recall on the rare class, as seen in prior IDS studies in [19], [20].

Our dataset exhibits extreme class imbalance, with only 36 infiltration flows compared to approximately 288,000 benign flows. Such imbalance can bias classifiers toward predicting only the majority class, resulting in low recall for the rare attack class. To mitigate this issue, we employ the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples of the minority class by interpolating between existing instances. SMOTE has been widely used with tree-boosting models such as LightGBM, XGBoost, and CatBoost, consistently achieving high accuracy and F1-scores on imbalanced datasets [22]. In IoT and network intrusion detection scenarios, SMOTE has been shown to improve minority-class recall and overall detection accuracy when combined with classifiers like K-Nearest Neighbors, Decision Trees, and Random Forests [23].

We apply SMOTE on the training split only with a target 1:1 minority-to-majority ratio. For each minority instance x, select one of its k-nearest minority neighbors x_{NN} and generate a synthetic sample by linear interpolation, as shown in equation (2):

$$x_{\text{new}} = x + \lambda(x_{NN} - x), \qquad \lambda \sim U(0,1) \quad (2)$$

This increases minority support and typically improves recall for rare "Infiltration" flows in imbalanced IDS settings. This approach avoids contaminating the test set and allows a clear comparison between with-SMOTE and without-SMOTE scenarios in our ablation study. Prior IDS research [21], [22], [23] suggests that this strategy can significantly increase recall on rare attack classes without sacrificing precision.

2.4. Ensemble Models

Our classifier is a VotingClassifier ensemble combining XGBoost, Random Forest, and Logistic Regression. Each base model is trained on the selected features subset. A VotingClassifier that integrates XGBoost, Random Forest, and Logistic Regression can capitalize on the strengths of each model. In studies, XGBoost has shown superior performance in terms of accuracy and recall, making it suitable for applications where identifying the minority class is critical, such as fraud detection and disease prediction [24], [25]. The XGBoost model optimizes the following objective/loss function [26] as follow in equation (3):

$$L(f_k) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (3)$$

Where $\hat{y}_i$ is predicted output for sample i, $l(y_i, \hat{y}_i)$ is the loss function, usually logistic loss for classification, f_k is the $k-th$ regression tree, K is the total number of tress, and $\Omega(f_k)$ is the regularization term that penalized the complexity of the model. The predictions for each step are updates as follow in equation (4):

$$\widehat{y^{(t)}}_i = \widehat{y^{(t-1)}}_i + \eta f_t(X_i) \qquad (4)$$

where η is learning rate, f_t is the new regression tree at step t, and X_i is the vector feature for sample i. Hence, we can write the models of XGBoost as follow in equation (5):

$$f_{XGB}(X) = \sum_{k=1}^{K} f_k(X) \qquad (5)$$

Random Forest can handle the complexity and imbalance [25], with a model as follow in equation (6):

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

$$f_{RF}(X) = \frac{1}{B}\sum_{b=1}^{B} T_b(X) \quad (6)$$

where $T_b(X)$ is the prediction of the b^{th} tree, and B is the total number of trees in the forest. On the other hand, Logistic Regression adds interpretability [21], [22]. For binary classification, Logistic Regression models the probability of the positive class as written in equation (7):

$$P(y = 1 \mid X) = f_{LR}(X) = \sigma\left(\mathbf{w}^{\top} X + b\right) \quad (7)$$

where X is input feature vector, $\mathbf{w}$ is weight vector, b is bias term, and $\sigma(\cdot)$ is sigmoid function mapping to $[0,1]$, also can be defined as $\sigma(z) = \frac{1}{1+e^{-z}}$.

In our study, we use ensemble method. Ensemble methods generally achieve higher and more robust performance than any single model, by leveraging their complementary strengths [7]. For example, [7] demonstrate that combining different algorithms (trees, gradient boosting, etc.) in a voting ensemble "can be effective because it leverages the strengths of each model". Likewise, some related work has used XGBoost and Logistic Regression together (via late fusion) for IDS, emphasizing that such combinations can improve both accuracy and interpretability with SHAP [23]. In our ensemble, we adopt soft voting, where the final prediction score is obtained by averaging the predicted probabilities from all base learners. The final ensemble in our study as follows in equation (8):

$$P(y = 1 \mid X) = \frac{f_{RF}(X) + f_{XGB}(X) + 2 \cdot f_{LR}(X)}{4} \quad (8)$$

Following our implementation, the ensemble combines the three base models with weights $[1,1,2]$, giving higher emphasis to Logistic Regression. The predicted probability for the infiltration class, is then, as written in equation (9):

$$\hat{y} = \begin{cases} 1, & P(y = 1 \mid X) \geq \theta \\ 0, & otherwise \end{cases} \quad (9)$$

where θ is classification threshold. In our study, we analyze the $\theta \in \{0.3, 0.5, 0.7\}$.

2.5. Experiment and Evaluation Matrices

Overall, five configurations were evaluated: (1) Baseline (all features, no SMOTE), (2) SHAP Top-5 (no SMOTE), (3) SHAP Top-5 + SMOTE, (4) SHAP Top-30 (no SMOTE), and (5) SHAP Top-30 + SMOTE. Each configuration was assessed under three classification thresholds, resulting in 15 experimental runs for comprehensive ablation analysis. We evaluate all models on a held-out test set (using stratified sampling to preserve class ratio), on a 70:30 train/test split, using stratified sampling to preserve the extreme class imbalance (36 infiltration vs ~288k benign flows). The flow of the experiment can be seen in Figure 1.

Performance is measured by ROC-AUC, precision, recall (true positive rate), and F1-score. We therefore report the standard definitions formula of Precision, Recall and F1, can be seen in equation (10), (11), (12), where TP, FP, FN is True Positive, False Positive and False Negative, respectively:

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (12)$$

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

These definitions are applied uniformly across all ablation settings and thresholds $\theta \in \{0.3,0.5,0.7\}$.

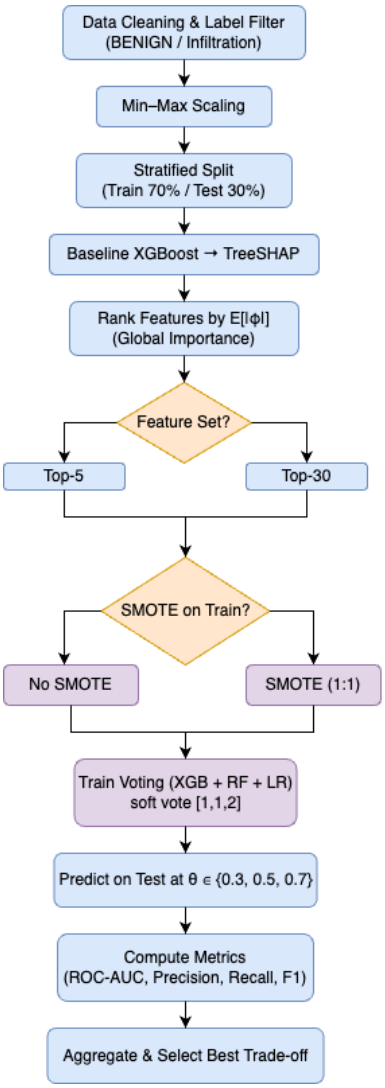

Figure 1. Experiment flow

ROC-AUC summarizes the ability of the model to distinguish between benign and infiltration classes across all thresholds. Precision reflects how many predicted attacks are truly infiltration events. Recall indicates how many actual infiltration events are correctly detected. F1-score is the harmonic mean of precision and recall, which is particularly relevant for imbalanced IDS datasets where both false positives and false negatives are critical. F1 Score is a widely used metric in evaluating IDS performance, especially in imbalanced datasets. It provides a balance between precision and recall, making it suitable for scenarios where false negatives and false positives are critical [6], [11], [14].

A high F1 means the model detects most attacks (high recall) without an excessive false-positive rate (high precision). In fact, it is mentioned in [14] note that “in practice, the performance measure of interest in IDPS is the F1 score, since raw accuracy is misleading on imbalanced data”. Thus, while we report ROC-AUC for completeness, we pay particular attention to precision, recall, and F1.

3. RESULT

We performed a full factorial experiment over feature selection (Top-5 vs Top-30), SMOTE (yes/no), and Classification thresholds (0.3, 0.5, 0.7). In all cases the VotingClassifier used XGBoost,

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

Random Forest, and Logistic Regression with default parameters (tuned marginally via cross-validation). Table 1 summarizes key metrics for representative settings (we omit trivial combinations where thresholding removes nearly all features).

3.1. Dataset Class Distribution

The Thursday-WorkingHours-Afternoon-Infiltration subset is extremely imbalanced, with only 36 infiltration flows versus 288,566 benign flows. The test split (30%) contains 11 infiltration flows and 75,826 benign flows, which we use as the basis for the confusion matrices in the Results section. This imbalance motivates the use of SMOTE and threshold ablation to improve minority detection.

3.2. SHAP-Based Feature Analysis Ranking

SHAP analysis on the baseline XGBoost model identified the most influential features for infiltration detection. The results of SHAP Feature Ranking shown in Figure 2:

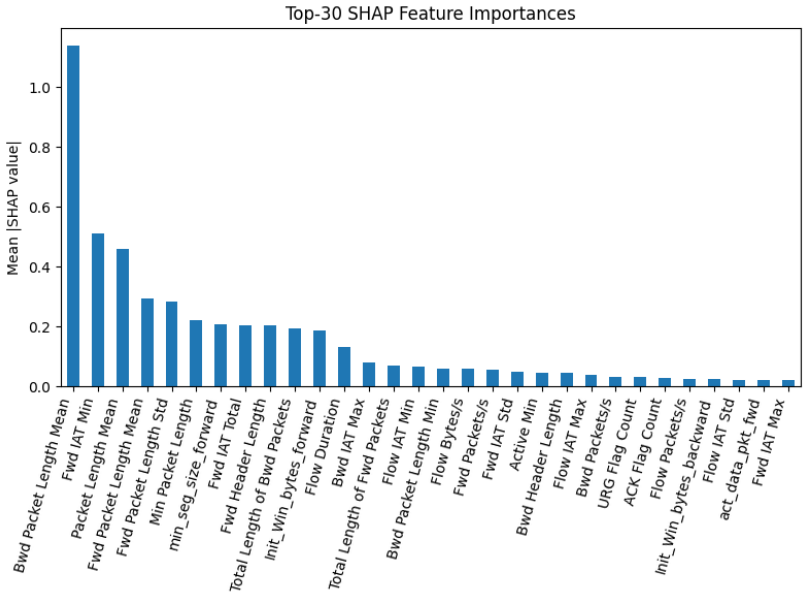

Figure 2. SHAP Feature Analysis Result

It can be seen from the figure that the Top 5 Features are as follow: ['Bwd Packet Length Mean', 'Fwd IAT Min', 'Packet Length Mean', 'Fwd Packet Length Mean', 'Fwd Packet Length Std']. While the Top-30 Features (including flow-based statistics and flag counts) show that infiltration traffic is strongly correlated with: Packet length statistics (mean, std, min); Forward/Backward inter-arrival times (IAT); Flow duration and bytes per second; TCP flags such as URG/ACK counts. This ranking supports the intuition that temporal patterns and packet size behavior are key indicators of stealthy lateral movement.

3.3. Experiment Result

We performed a full factorial experiment over feature selection (Top-5 vs Top-30), SMOTE (yes/no), and Classification thresholds $\theta \in (0.3, 0.5, 0.7)$. In all cases the VotingClassifier used XGBoost, Random Forest, and Logistic Regression with default parameters (tuned marginally via cross-validation). Table 1 show metrics for representative settings (we omit trivial combinations where thresholding removes nearly all features). As it can be seen from the Table 1, the best-performing configuration is SHAP Top-30 + SMOTE at threshold $= 0.7$, as it achieves ROC-AUC = 0.976 and F1 = 0.778 on the rare infiltration class.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

Table 1. Ablation Result Across Configuration

Model	Threshold	ROC-AUC	Precision	Recall	F1
Baseline (All)	0.3	0.9573	0.043	**0.818**	0.081
	0.5		0.233	0.636	0.341
	0.7		**1.000**	0.545	0.706
SHAP Top-5	0.3	0.8986	0.006	0.727	0.012
	0.5		0.033	0.636	0.062
	0.7		**1.000**	0.545	0.706
SHAP Top-5	0.3	0.9112	0.007	0.727	0.015
SMOTE	0.5		0.031	0.636	0.060
	0.7		0.750	0.545	0.632
SHAP Top-30	0.3	0.9340	0.020	0.727	0.039
	0.5		0.292	0.636	0.400
	0.7		**1.000**	0.545	0.706
SHAP Top-30	0.3	**0.9758**	0.034	**0.818**	0.065
SMOTE	0.5		0.364	0.727	0.485
	0.7		**1.000**	0.636	**0.778**

The Table 1, also can be represented as precision-recall trade-off plot, as shown in Figure 3. Figure 3 illustrates the precision-recall trade-off for the five model configurations when applied to infiltration detection on the severely imbalanced CIC-IDS 2017 dataset. For each model, the solid line depicts recall, measuring the ability to identify true attacks, while the dashed line represents precision, measuring the reliability of the alerts.

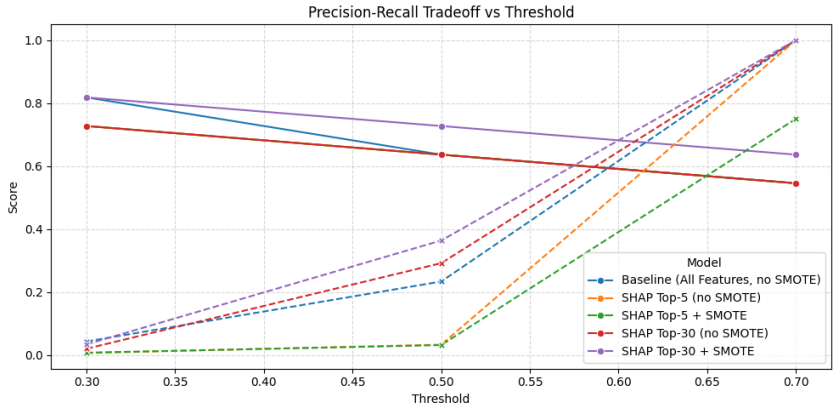

Figure 3. Precision-Recall Tradeoff

The plot highlights the critical operational dilemma in intrusion detection. At a low threshold of 0.3, models act as highly sensitive detectors. They achieve high recall (catching over 81% of attacks, with a maximum of 0.818 for the Baseline and SHAP Top-30 + SMOTE models), but this comes at the cost of extremely low precision, as also shown in Figure 4. Such a configuration would detect most threats but would also flood security analysts with an unmanageable number of false alarms. Conversely, at a high threshold of 0.7, the models become highly specific. While four configurations achieve perfect precision (1.000), this reliability comes at the dangerous cost of reduced recall, meaning a portion of attacks would be missed.

This analysis underscores the superior performance of models enhanced with the Synthetic Minority Over-sampling Technique (SMOTE) to combat the dataset's severe class imbalance. The SHAP Top-30 + SMOTE model (purple line) emerges as the most effective and practical configuration.

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

At the 0.7 threshold, it achieves a perfect precision of 1.000, effectively generating zero false alarms. Critically, it does so while maintaining a strong recall of 0.636, successfully detecting nearly 64% of all infiltration events. This superior balance yields the highest F1-Score of 0.778 (as shown in Table 1) and represents an ideal profile for a real-world security tool that minimizes analyst fatigue while maintaining robust threat detection.

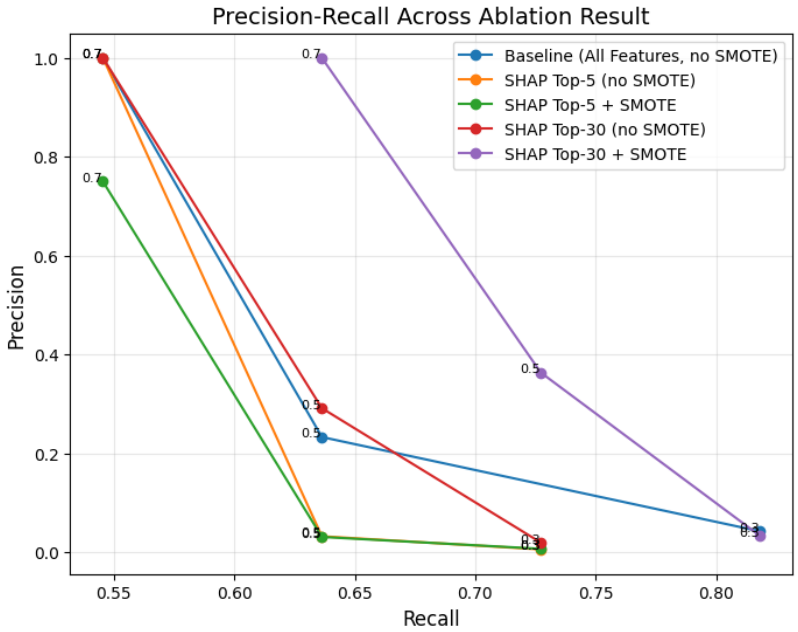

Figure 4. Precision-Recall Across Ablation Result

Figure 5 provides a conclusive summary of model performance by comparing the F1-Scores, a critical metric for the imbalanced intrusion detection task, across all configurations and thresholds. A clear trend emerges where the F1-Score consistently improves as the threshold increases from 0.3 to 0.7, indicating that the gains in precision at higher thresholds create a better overall model balance. The chart culminates in identifying the optimal configuration: the SHAP Top-30 + SMOTE model, when paired with a 0.7 threshold, achieves a distinctly superior F1-Score of 0.778. This result quantitatively confirms the conclusions from the precision-recall analysis, demonstrating that the synergy between advanced feature selection (SHAP) and robust class imbalance handling (SMOTE) produces the most effective and well-balanced model for this security application.

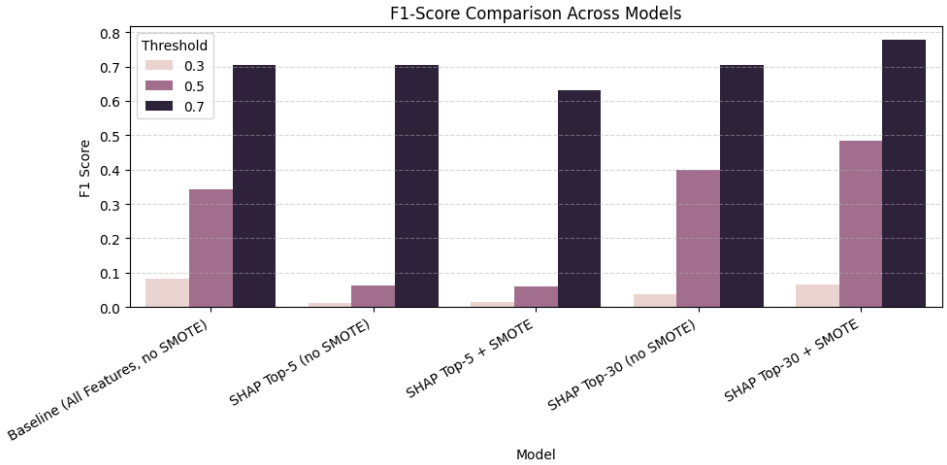

Figure 5. F1-Score Comparison Across Models

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

To provide operational insight, we add confusion matrices for the best configuration (SHAP Top-30 + SMOTE) across thresholds on the test set (BENIGN = 75,826; Infiltration = 11) as shown in Table 2. At $\theta = 0.7$, the ensemble outputs zero false positives ($FP = 0$) and only seven positive predictions, all of which are true attacks ($TP = 7$), hence $precision = 1.000$. This occurs because a higher decision threshold requires higher posterior confidence from the soft-voting ensemble (weights [1,1,2]); borderline benign flows stay below 0.7 and are not flagged. The trade-off is lower recall ($\frac{7}{11} = 0.636$) because some true attacks with scores between $0.3 - 0.7$ are no longer captured. Conversely, at $\theta = 0.3$ the detector becomes very sensitive ($TP = 9$) but admits many low-confidence alerts ($FP = 257$), sharply reducing precision to 0.0338 on an extremely imbalanced test set. With 75,826 negatives in the test set, even a small number of FP can sharply depress precision, especially at low thresholds.

This demonstrates the classic precision–recall trade-off in IDS: raising the decision threshold eliminates false positives but simultaneously degrades recall by allowing more attacks to slip through. These matrices complement Table 2 by making the operational trade-off explicit: security teams can tune θ depending on tolerance to false alarms (favoring higher θ) versus missed attacks (favoring lower θ). The $\theta = 0.7$ setting is attractive when minimizing analyst fatigue is paramount, while $\theta = 0.5$ offers a middle-ground with improved recall at the cost of some false positives.

Table 2. Confusion matrices for SHAP Top-30 + SMOTE across thresholds (test set)

Threshold	TN	FP	FN	TP	Precision	Recall	F1
0.3	75,569	257	2	9	0.034	**0.818**	0.065
0.5	75,812	14	3	8	0.364	0.727	0.485
0.7	75,826	0	4	7	**1.000**	0.636	**0.778**

4. DISCUSSIONS

Our experiments highlight the critical trade-offs involved in designing an effective intrusion detection system, particularly for the imbalanced CIC-IDS 2017 dataset. The results confirm that a multi-faceted approach, balancing feature selection, class imbalance, and threshold tuning, is necessary for optimal performance.

The most significant factor in model performance was the application of the Synthetic Minority Over-sampling Technique (SMOTE). Without it, models tuned for high precision via a high threshold (0.7) saw their recall and F1-scores collapse. With SMOTE, the classifier could effectively learn the patterns of the minority "Infiltration" class, greatly improving recall. This aligns perfectly with findings in the literature; many IDS studies on the CIC-IDS 2017 dataset report that resampling techniques like SMOTE are essential to mitigate bias and enhance the model's ability to distinguish minority attack classes [14], [19]. Empirically, studies that inject SMOTE into IDS pipelines (KDD, CSE-CIC-IDS2018) report improved minority detection [27]; our ablation echoes those gains when SMOTE is combined with SHAP-ranked features (Top-30). In our case, enabling SMOTE was the key factor that allowed the SHAP Top-30 + SMOTE model to maintain a high recall of 0.636 even at a threshold that enforced perfect precision. Severe class imbalance is a well-documented barrier in IDS; recent surveys reiterate that accuracy alone is misleading, and that recall/precision (or F-scores) should be foregrounded [28], matching our metric choices.

Threshold selection provides a tunable lever to manage operational priorities. Our results showed a clear trade-off: a low threshold (0.3) maximized recall at the cost of extremely low precision (flooding analysts with false alarms), while a high threshold (0.7) maximized precision at the cost of recall (missing potential attacks). While theory suggests a compromise is often best, our findings indicate that for this specific problem, the 0.7 threshold was optimal. It pushed precision to a perfect 1.0, and the SMOTE-enhanced model was robust enough to absorb this pressure without a catastrophic loss of recall,

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

ultimately yielding the highest F1-Score of 0.778. This demonstrates that for environments where the cost of false positives is very high, a high threshold can be viable if the model is properly balanced.

The feature selection comparison revealed that including more SHAP-informed features (Top-30 vs. Top-5) was beneficial. The 30-feature models consistently outperformed the 5-feature models in F1-score, suggesting the additional features carried meaningful signals for distinguishing attacks. This validates the use of SHAP for creating an effective and efficient feature set, one that is compressed from the original dataset but not so aggressively that it loses critical information. This approach is consistent with other state-of-the-art systems where explainable AI-driven feature selection is combined with ensemble classifiers to improve IDS performance and transparency [12], [13], [14].

Overall, the superior performance of the SHAP Top-30 + SMOTE configuration confirms that an ensemble classifier + SHAP-based feature selection + SMOTE is a highly effective recipe for imbalanced intrusion detection. The clear patterns visualized in the Precision-Recall curves provide actionable insight for real-world deployment, quantifying exactly how much recall is sacrificed for gains in precision at each decision point. This allows practitioners to tune the system explicitly to meet specific operational needs, balancing the risk of missed attacks against the cost of false alarms. Our findings are consistent with ensemble-based IDS literature that reports accuracy and robustness gains from combining heterogeneous learners. Beyond accuracy, interpretability is increasingly emphasized: works that pair SHAP explanations with IDS show how feature-attribution improves analyst trust and triage, and even explore LLM-generated narratives over SHAP outputs to make decisions auditably human-readable [29]; our SHAP-guided feature selection aligns with this direction.

Comparing to prior studies, our results are competitive with the current state of the art. When contextualized with recent literature, our work highlights the distinct advantages of a supervised classification approach for detecting specific rare attacks. A study by Panwar et al. [30] on the CICIDS-2017 dataset provides a compelling point of comparison. Using an anomaly detection framework, where the model is trained primarily on benign data to identify deviations, their Random Forest classifier with RFE feature selection achieved a near-perfect F1-Score of 0.9992 for the "Infiltration" class. This impressive result suggests the "Infiltration" traffic is highly distinct from normal traffic, making it easily flaggable as an anomaly.

While anomaly detection is a common approach, our study employed a different paradigm: a supervised Voting Classifier explicitly trained on both benign and "Infiltration" samples that were balanced by SMOTE. While our resulting F1-Score of 0.78 is more modest, our methodology achieved a perfect precision of 1.0. This distinction is critical: the anomaly detection approach proved effective at identifying outliers, but our supervised method was tuned to produce high-certainty alerts with zero false positives. For a security operations team focused on mitigating a known threat like lateral movement, every alert is actionable, demonstrating the unique value of a targeted, supervised learning approach for producing high-certainty alerts with zero false positives.

These results underscore the urgency of strengthening east-west monitoring in virtualized/VXLAN data centers, where encapsulated VM-to-VM flows can bypass perimeter inspection and create blind spots for traditional IDS. By pairing interpretable selection (SHAP) with class-imbalance handling (SMOTE), the proposed ensemble offers a practical route to high-certainty alerts under such constraints, even under extreme imbalance. This complements ensemble/XAI frameworks that emphasize modularity and deployability [31]

Limitations: The study acknowledges several limitations that suggest avenues for future research:

- Dataset Scope: The evaluation was confined to the CIC-IDS2017 "Infiltration" dataset. Future studies should examine other lateral-movement scenarios, such as internal reconnaissance, to test the generalizability of the findings.

- Model Specificity: A specific ensemble with fixed hyperparameters was used. Exploring different architectures, such as neural networks or other classifiers, could yield different results.
- Feature Set: Real-world virtualized networks may offer additional meta-features (e.g., VNI tags) that were not available in this dataset but could be incorporated to enhance detection.

Despite these limitations, the core findings, that SHAP and SMOTE are crucial for balancing the precision-recall. Trade-off are expected to generalize to other imbalanced security scenarios.

Ethical considerations: While SMOTE corrects minority under-representation, synthetic interpolation can alter local data geometry and amplify atypical patterns if overused [27], [32]. To reduce optimism and leakage risk, we apply SMOTE on training only and keep the test distribution intact; this mirrors best-practice cautions in IDS studies on imbalanced data. Future work should check whether oversampling biases SHAP attributions and perform fairness audits across sub-populations.

5. CONCLUSION

This study successfully demonstrated the effectiveness of a voting ensemble IDS for detecting rare infiltration attacks on the CICIDS2017 dataset. Through a systematic ablation study, we confirmed that combining SHAP-driven feature selection with SMOTE for data balancing is a highly effective strategy. Our best configuration, SHAP Top-30 + SMOTE with decision threshold $\theta = 0.7$, achieves $F1 = 0.778\,(77.8\%)$, $precision = 1.000$, and $recall = 0.636$ on the infiltration class. This reflects a deliberate operating point that prioritizes high-certainty alerts while maintaining competitive recall. Operationally, threshold tuning dramatically reduces false alarms in real-time settings: false positives drop from 257 at $\theta = 0.3$ to 14 at $\theta = 0.5$, and to 0 at $\theta = 0.7$. This yields high-certainty alerts while preserving competitive recall, improving analyst efficiency in virtualized/VXLAN environments where east-west traffic is highly imbalanced. By pairing interpretable selection (SHAP) with training-only oversampling (SMOTE) inside a soft-voting ensemble, the approach provides a practical, tunable IDS that balances transparency and performance. Future work will validate on live virtualized traffic and extending it to other types of network attacks.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Data Science Computer Lab, Department of Mathematics, Universitas Brawijaya for providing the support necessary to conduct this research.

REFERENCES

[1] M. Elmadani and S. O. Sati, "Data Center Lab Using VxLAN Data Plane and BGP-EVPN Control Plane," in *2023 4th International Conference on Data Analytics for Business and Industry, ICDABI 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 354–358. doi: 10.1109/ICDABI60145.2023.10629438.

[2] D. Li, Z. Yang, S. Yu, M. Duan, and S. Yang, "A Micro-Segmentation Method Based on VLAN-VxLAN Mapping Technology," *Future Internet*, vol. 16, no. 9, Sep. 2024, doi: 10.3390/fi16090320.

[3] D. Kushwaha *et al.*, "Lateral Movement Detection Using User Behavioral Analysis," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.13524

[4] C. Smiliotopoulos, G. Kambourakis, and C. Kolias, "Detecting lateral movement: A systematic survey," *Heliyon*, vol. 10, no. 4, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26317.

[5] U. Ahmed *et al.*, "Explainable AI-based innovative hybrid ensemble model for intrusion detection," *Journal of Cloud Computing*, vol. 13, no. 1, Dec. 2024, doi: 10.1186/s13677-024-00712-x.

[6] V. Shanmugam, R. Razavi-Far, and E. Hallaji, "Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches," *Electronics (Switzerland)*, vol. 14, no. 1, Jan. 2025, doi: 10.3390/electronics14010069.

[7] A. H. Farooqi, S. Akhtar, H. Rahman, T. Sadiq, and W. Abbass, "Enhancing Network Intrusion Detection Using an Ensemble Voting Classifier for Internet of Things," *Sensors*, vol. 24, no. 1, Jan. 2024, doi: 10.3390/s24010127.

[8] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy*, SciTePress, 2018, pp. 108–116. doi: 10.5220/0006639801080116.

[9] Z. I. Khan, M. M. Afzal, and K. N. Shamsi, "A Comprehensive Study on CIC-IDS2017 Dataset for Intrusion Detection Systems," 2024. [Online]. Available: https://irjaeh.com

[10] S. Borah and R. Panigrahi, "A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems," 2018. [Online]. Available: https://www.researchgate.net/publication/329045441

[11] M. S. Towhid, N. S. Khan, M. Hasan, and N. Shahriar, "Towards Effective Network Intrusion Detection in Imbalanced Datasets: A Hierarchical Approach."

[12] Q. Xu *et al.*, "SHAP-based Interpretable Models for Credit Default Assessment Using Machine Learning," in *Proceedings - 2024 14th International Conference on Software Technology and Engineering, ICSTE 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 213–217. doi: 10.1109/ICSTE63875.2024.00044.

[13] Ms. M. M. Kedar, "Exploring the Effectiveness of SHAP over other Explainable AI Methods," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 08, no. 06, pp. 1–5, Jun. 2024, doi: 10.55041/IJSREM35556.

[14] U. Ahmed *et al.*, "Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-81151-1.

[15] D. Spiekermann, T. Eggendorfer, and J. Keller, "Deep Learning for Network Intrusion Detection in Virtual Networks," *Electronics (Switzerland)*, vol. 13, no. 18, Sep. 2024, doi: 10.3390/electronics13183617.

[16] L. H. Li, R. Ahmad, R. Tanone, and A. K. Sharma, "STB: synthetic minority oversampling technique for tree-boosting models for imbalanced datasets of intrusion detection systems," *PeerJ Comput Sci*, vol. 9, 2023, doi: 10.7717/peerj-cs.1580.

[17] V. Surya and M. M. Selvam, "An Effective Machine Learning Approach for loT Intrusion Detection System based on SMOTE," in *6th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 905–911. doi: 10.1109/ICECA55336.2022.10009130.

[18] R. Kaur and N. Gupta, "Comprehending SMOTE Adaptations to Alleviate Imbalance in Intrusion Detection Systems," in *2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 976–982. doi: 10.1109/ICESC57686.2023.10193257.

[19] A. O. Widodo, B. Setiawan, and R. Indraswari, "Machine Learning-Based Intrusion Detection on Multi-Class Imbalanced Dataset Using SMOTE," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 578–583. doi: 10.1016/j.procs.2024.03.042.

[20] M. W. Nawaz, R. Munawar, A. Mehmood, M. M. U. Rahman, and Q. H. Abbasi, "Multi-class Network Intrusion Detection with Class Imbalance via LSTM & SMOTE," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.01850

[21] A. B. Hassanat, A. S. Tarawneh, S. S. Abed, G. A. Altarawneh, M. Alrashidi, and M. Alghamdi, "RDPVR: Random Data Partitioning with Voting Rule for Machine Learning from Class-Imbalanced Datasets," *Electronics (Switzerland)*, vol. 11, no. 2, Jan. 2022, doi: 10.3390/electronics11020228.

[22] T. Fulazzaky, A. Saefuddin, and A. M. Soleh, "Evaluating Ensemble Learning Techniques for Class Imbalance in Machine Learning: A Comparative Analysis of Balanced Random Forest,

Vol. 6, No. 4, August 2025, Page. 2693-2706
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5233

SMOTE-RF, SMOTEBoost, and RUSBoost," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 969–980, Dec. 2024, doi: 10.15294/sji.v11i4.15937.

[23] A. Hafid, M. Rahouti, and M. Aledhari, "Optimizing Intrusion Detection in IoMT Networks Through Interpretable and Cost-Aware Machine Learning," *Mathematics*, vol. 13, no. 10, May 2025, doi: 10.3390/math13101574.

[24] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," *Pattern Recognit Lett*, vol. 136, pp. 190–197, Aug. 2020, doi: 10.1016/j.patrec.2020.05.035.

[25] B. Septian Cahya Putra, I. Tahyudin, B. Adhi Kusuma, and K. Nur Isnaini, "Efektivitas Algoritma Random Forest, XGBoost, dan Logistic Regression dalam Prediksi Penyakit Paru-paru The Effectiveness of Random Forest, XGBoost, and Logistic Regression Algorithms in Predicting Lung Disease," 2024. [Online]. Available: https://www.kaggle.com/datasets/andot03bsrc/dataset-predic-terkena-penyakit-paruparu.

[26] G. Liu, "Leveraging Machine Learning for Telecom Banking Card Fraud Detection: A Comparative Analysis of Logistic Regression, Random Forest, and XGBoost Models," *Computers and Artificial Intelligence*, vol. 1, no. 1, pp. 13–27, Nov. 2024, doi: 10.70267/1cc7aw07.

[27] A. H. Ali, M. Charfeddine, B. Ammar, and B. Ben Hamed, "Intrusion Detection Schemes Based on Synthetic Minority Oversampling Technique and Machine Learning Models," in *Proceedings - 2024 IEEE 27th International Symposium on Real-Time Distributed Computing, ISORC 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ISORC61049.2024.10551335.

[28] K. Abhiram, H. Muthusamy, S. Ravindran, and V. Vijean, "A Comprehensive Survey of Intrusion Detection System Using Machine Learning and Deep Learning Approaches," in *10th International Conference on Advanced Computing and Communication Systems, ICACCS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1927–1932. doi: 10.1109/ICACCS60874.2024.10717043.

[29] A. Khediri, H. Slimi, A. Yahiaoui, M. Derdour, H. Bendjenna, and C. E. Ghenai, "Enhancing Machine Learning Model Interpretability in Intrusion Detection Systems through SHAP Explanations and LLM-Generated Descriptions," in *PAIS 2024 - Proceedings: 6th International Conference on Pattern Analysis and Intelligent Systems*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/PAIS62114.2024.10541168.

[30] S. S. Panwar, Y. P. Raiwani, and L. S. Panwar, "An Intrusion Detection Model for CICIDS-2017 Dataset Using Machine Learning Algorithms," in *2022 International Conference on Advances in Computing, Communication and Materials, ICACCM 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICACCM56405.2022.10009400.

[31] Thirumaraiselvi, Sreyaniketha, V. Rahul, and K. S. Tamilnilavan, "Enabling Robust Intrusion Detection in Network Traffic through an Integrated Machine Learning Framework," in *Proceedings - 2024 5th International Conference on Image Processing and Capsule Networks, ICIPCN 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 183–188. doi: 10.1109/ICIPCN63822.2024.00038.

[32] C. E. Ben Ncir, M. A. Ben HajKacem, and M. Alattas, "Enhancing intrusion detection performance using explainable ensemble deep learning," *PeerJ Comput Sci*, vol. 10, 2024, doi: 10.7717/PEERJ-CS.2289.