Intrusion Detection: Boruta Feature Selection and Semi-Supervised Outlier Clustering with Multi-Dataset Evaluation

Agni Isador Harsapranata
Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia
agniisador@gmail.com (corresponding author)

Eko Sediyono
Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia
eko@uksw.edu

Hindriyanto Dwi Purnomo
Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia
hindriyanto.purnomo@uksw.edu

Received: 27 August 2025 | Revised: 14 October 2025 | Accepted: 29 October 2025

 DOI: https://doi.org/10.48084/etasr.14351

ABSTRACT

Intrusion Detection Systems (IDSs) remain essential as network attacks continue to increase in both volume and sophistication. This study presents a unified, dataset-agnostic preprocessing framework that integrates Boruta-based feature selection with class-wise semi-supervised clustering for outlier reduction before classification. The proposed pipeline standardizes encoding and scaling, prevents label leakage, selects relevant features, filters noise, and maps labels to a binary normal/intrusion classification task. The framework is evaluated on three benchmark datasets, NSL-KDD, UNSW-NB15, and CIC-IDS2017, using five representative classifiers: Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), Convolutional Neural Network (CNN), and Long Short-Term Memory (LTSM), all under a consistent experimental protocol. Ablation studies and paired statistical significance tests are conducted to quantify the individual effects of feature selection and outlier filtering. Results on CIC-IDS2017 demonstrate that the entire pipeline yields consistent and often statistically significant improvements over a simplified baseline. On NSL-KDD, performance gains are model-dependent, whereas on UNSW-NB15, the framework remains competitive with the baseline. Overall test accuracies range from 90.7% to 99.96%, with the best-performing models achieving an AUC-ROC of approximately 1.00. These findings indicate that combining Boruta with semi-supervised outlier reduction provides an effective and generalizable preprocessing strategy for IDS, particularly in heterogeneous network traffic environments.

Keywords-intrusion detection; network security; boruta; outlier reduction; machine learning

I. INTRODUCTION

The rapid advancement of computer network technology has resulted in increasingly complex information systems. Alongside this progress, various security threats have emerged, elevating risks to both organizational and individual digital assets. Network intrusions have become more frequent and diverse, encompassing attacks such as Distributed Denial-of-Service (DDoS), botnets, social engineering, phishing, malware, man-in-the-middle, Cross-Site Scripting (XSS), and SQL injection [1]. Cybercriminals continuously exploit new techniques to identify vulnerabilities in existing systems, often intending to steal sensitive information from organizations or individuals.

Intrusion Detection Systems (IDSs) are essential in mitigating these risks. IDS tools monitor network traffic, identify anomalies, and alert administrators to potential security breaches [2]. Machine Learning (ML)-based intrusion detection research has addressed persistent challenges such as high-dimensional feature spaces, dataset distribution shifts, and noisy or outlier samples through approaches including feature selection, normalization, and resampling strategies [3]. This study presents a unified and adaptive preprocessing pipeline that integrates Boruta feature selection with per-attack-type semi-supervised clustering-based outlier filtering prior to classification. The proposed pipeline is designed for heterogeneous datasets and prevents information leakage by separating labels from features, applying consistent encoding

and scaling, performing feature selection, reducing noise, and mapping labels to a binary normal/intrusion classification task prior to model training. The main contributions of this work are summarized as follows:

- A unified, dataset-agnostic preprocessing pipeline that incorporates encoding, scaling, feature selection, and noise filtering, aimed at enhancing robustness and cross-dataset generalization in IDS.
- Stage-wise integration of Boruta and semi-supervised clustering. The combination of Boruta with per-class outlier reduction improves feature relevance and training data quality prior to model training.
- Comprehensive multi-dataset evaluation. The proposed framework is assessed using a consistent protocol across NSL-KDD, UNSW-NB15, and CIC-IDS2017 datasets with five representative classifiers, Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), Convolutional Neural Network (CNN), and Long Short-Term Memory Network (LSTM), representing tree-based, linear, and Deep Learning (DL) paradigms.
- Ablation and statistical significance tests. Detailed ablation experiments and paired statistical tests are conducted to isolate and evaluate the impact of each preprocessing component (feature selection and outlier filtering).

In [4], the Enhanced Gorilla Troops Optimizer (EGTO) was used to improve the feature selection process prior to classification. This study introduced an Interrelated Dynamic Biased Feature Selection Model using EGTO (IDBFS-EGTO) to generate a set of feature vectors for intrusion detection. The EGTO method utilizes a collection of operators to achieve a more stable balance between exploitation and exploration, resulting in an ML model that achieved 98.4% accuracy in intrusion detection and 98.6% accuracy with EGTO-based optimization [4]. The IDCS-ELIBWO (Enhanced Intrusion Detection in Cybersecurity using Ensemble Learning with Improved Beluga Whale Optimization) technique [5] employed an ensemble learning classifier comprising three DL architectures: Deep Belief Network (DBN), Gated Recurrent Unit (GRU), and LSTM. The Improved Beluga Whale Optimization (IBWO) algorithm was then used for hyperparameter tuning. Extensive experiments demonstrated that the IDCS-ELIBWO method achieved a validation accuracy of 99.77%, indicating its superior performance in intrusion detection tasks. In [6], a two-layer intrusion classification framework utilized in the first layer a CNN-BiLSTM model to jointly process network traffic data, majority-class attacks, and aggregated minority-class attacks. The second layer enhanced minority-class attack classification through stacking ensemble learning. This model was evaluated on three datasets—CIC-IDS2017, NSL-KDD, and the Mississippi Gas Pipeline Industrial Network Dataset—to ensure generalization and real-world applicability, achieving overall detection accuracies of 99%, 99%, and 95%, respectively.

Feature selection plays a crucial role in building effective intrusion classification models, with each approach presenting distinct advantages and limitations. This study introduces a novel framework that differs from prior research by integrating Boruta-based feature selection with per-attack-type semi-supervised clustering for outlier reduction. The framework maps samples into intrusion and normal using multiple classifiers, including RF, LR, GB, CNN, and LSTM. Evaluation is performed using accuracy, precision, recall, F1-score, and AUC-ROC under a standardized data split and metric protocol. In summary, whereas previous studies typically addressed feature selection or outlier handling in isolation, the proposed pipeline explicitly combines both processes and validates their effectiveness and generalizability across three heterogeneous benchmark datasets.

II. METHODOLOGY

The proposed preprocessing pipeline is designed for uniform handling of diverse datasets, including NSL-KDD, UNSW-NB15, and CIC-IDS2017. Figure 1 illustrates the overall configuration of the proposed method.

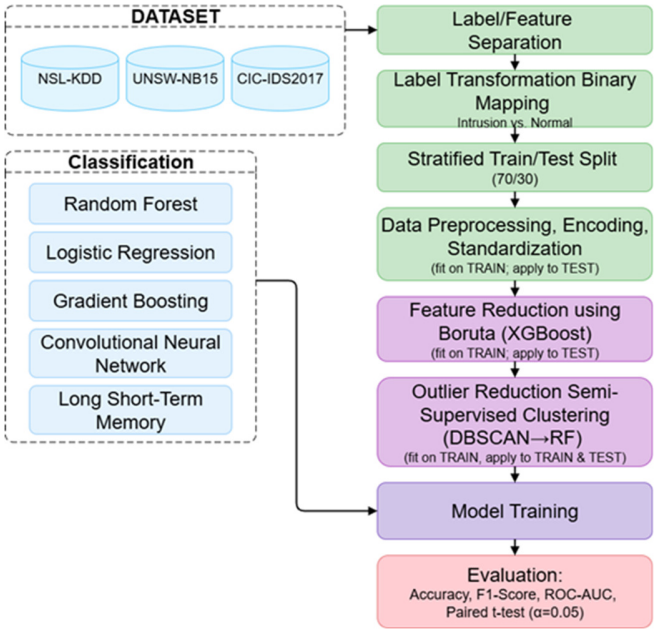

Fig. 1. Proposed model for IDS.

A. Dataset Description

Datasets play a critical role in developing effective machine learning models. Each dataset generally consists of three primary components: features (attributes), labels, and instances. Features represent the measurable characteristics of the data—such as connection duration, transfer bytes, and protocol type—which are used to describe network behavior.

1) NSL-KDD

The NSL-KDD dataset is derived from the KDD-CUP99 dataset [7, 8]. It contains 43 features utilized for both intrusion and normal traffic classification. The dataset was developed by the Canadian Institute for Cybersecurity and introduced in 2009. In [9], this dataset was employed in a classification framework that achieved an accuracy of 98.47%.

2) UNSW-NB15

The UNSW-NB15 dataset was generated in 2015 using the IXIA PerfectStorm tool at the Cyber Range Lab, UNSW Canberra. This dataset combines contemporary normal network activity with modern synthetic attack behaviors [10-12] and has since become one of the most widely used benchmarks for intrusion detection research. This study utilized 49 features. In [13], this dataset was employed in a classification model that combined mutual information and the Boruta algorithm for feature selection.

3) CIC-IDS2017

The CIC-IDS2017 dataset contains realistic packet capture (PCAP) network traffic summarized using CICFlowMeter. This tool extracts over 80 flow-based features from network traffic and stores them in CSV format [14]. In [15], an intrusion detection model based on an RF classifier achieved approximately 99.3% classification accuracy on this dataset [15].

B. Boruta-Based Feature Selection

Feature selection aims to identify the most critical subset of features that significantly influence model performance. It is a fundamental step in developing ML–based predictive models, as it improves both accuracy and interpretability. Algorithm 1 outlines the Boruta-based feature selection process employed in this study.

Algorithm 1: Boruta-based feature selection

Step 1: Initialize Boruta

Prepare the feature matrix $X \in R^{n \times m}$ (scaled, if necessary) and the target vector $y \in R^n$.

Choose XGBoost Classifier as the estimator.

Step 2: Add shadow features (random features).

For each original feature x_j, create a shadow feature by randomly permuting its values:

x_j^{shadow}=permute x_j (1)

Where permute shuffles the values in x_j independently of y.

The augmented dataset becomes:

$X^{aug} = [X \; X^{shadow}]$ (2)

Step 3: Train XGBoost on the augmented data.

Fit XGBoost on X^{aug} and y.

Compute feature importance for all features (original and shadow) using a metric such as "gain":

$I(x_j)$ = Importance from XGBoost for x_j (3)

$I(x_j^{shadow})$ = Importance from XGBoost for x_j^{shadow} (4)

Step 4: Compare feature importance.

Find the maximum importance among all shadow features:

$I_{max}^{shadow} = \max_j I(x_j^{shadow})$ (5)

For each original feature x_j , compare:

If $I(x_j) > I_{max}^{(shadow)}$ and statistically significant (6)

Mark x_j as Accepted.

Elseif $I(x_j) < I_{max}^{shadow}$ and statistically significant (7)

Mark x_j as Rejected.

Otherwise:

Mark x_j as Tentative.

Step 5: Select important features.

Gather all x_j where $I(x_j)$ is statistically greater than I_{max}^{shadow} :

$F_{\text{Selected}} = \{x_j \mid I(x_j) > I_{max}^{shadow} \; and \; significant\}$ (8)

Iterate until convergence or maximum iteration

Remove all rejected features.

Retain accepted features as the final selection.

For tentative features, repeat Steps 2-5 with new shadow features until all are marked as accepted or rejected, or the maximum iteration limit is reached.

Boruta is an effective feature selection technique for reducing dataset dimensionality, thereby improving interpretability and enhancing the predictive performance of machine learning models.

C. Outlier Reduction via Semi-Supervised Clustering

Outlier reduction through semi-supervised clustering is employed to eliminate anomalies within each attack category. Algorithm 2 illustrates the clustering stages used to minimize outliers for each class in the dataset. This approach combines DBSCAN (unsupervised) and RF (supervised) due to their efficiency when applied to large datasets. DBSCAN identifies dense regions as clusters, while data points outside these regions are considered noise or outliers. DBSCAN is applied only to sampled data, whereas RF generalizes these results to all data classes, offering faster and more stable predictions on large-scale datasets.

Algorithm 2: Outlier Reduction Process via Semi-Supervised Clustering

Step 1: Data extraction by class

Given a labeled dataset (X, y) with $X \in \mathrm{R}^{n \times m}$ and $y \in \mathcal{C}$, select data for class c:

$X_{sub} = \{x_i \in X \mid y_i = c\}, \; y_{sub} = \{y_i \mid y_i = c\}$ (9)

where c is the class being processed.

Step 2: Outlier detection using DBSCAN

Apply DBSCAN to X_{sample} (a sampled subset of X_{sample}):

$l^{DB} = \mathrm{DBSCAN}_\theta(X_{sample})$ (10)

where $l_i^{DB} \in -1,0,1,2,\ldots,k$ (-1 means outlier/ noise, k is the number of clusters).

Step 3: Convert DBSCAN labels to binary

Define a binary label:

$$y_i^{\{DB\}} = \begin{cases} 1, & l_i^{\{DB\}} \neq -1 \\ 0, & l_i^{\{DB\}} = -1 \end{cases} \quad (11)$$

Step 4: RF training

Train an RF model $\mathcal{F}$ using the sample data and binary labels:

$$\mathcal{F} = \text{RFClassifier}.fit(X_{sample}, y^{DB}) \quad (12)$$

Step 5: Inlier prediction on all class data

Apply $\mathcal{F}$ to all X_{sub}:

$$\hat{y}^{inlier} = \mathcal{F}.predict(X_{sub}) \quad (13)$$

Extract inlier data and labels:

$$X_{inlier} = \{x_j \in X_{sub} \mid \hat{y}^{inlier} = 1\} \quad (14)$$

$$y_{inlier} = \{y_j \mid x_j \in X_{inlier}\} \quad (15)$$

Step 6: Save inlier data

Store $(X_{inlier}; y_{inlier})$ for next steps or export.

D. Classification Algorithms

- Random Forest (RF) is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the class that represents the majority vote of these trees. It enhances model robustness and reduces overfitting, making it suitable for high-dimensional classification problems [16].
- Gradient Boosting (GB) is a powerful ensemble algorithm that builds a series of decision trees sequentially. Each tree corrects the errors made by the previous ones using gradient descent optimization on a defined objective function. This results in superior predictive performance compared to standard decision trees [17].
- Logistic Regression (LR) is a linear classification model commonly used for binary classification tasks. It applies the logistic (sigmoid) function to estimate the probability that a given input belongs to a particular class. Its simplicity, interpretability, and computational efficiency make it an effective baseline classifier [18].
- Convolutional Neural Network (CNN): CNNs are deep learning models designed to process data with a grid-like topology, such as images or time-series data. By leveraging convolutional, pooling, and fully connected layers, CNNs capture spatial hierarchies of features and effectively reduce feature dimensionality. This makes CNNs particularly well-suited for intrusion detection tasks requiring automated feature extraction.
- Long Short-Term Memory (LSTM) is a type of Recurrent Neural Networks (RNNs) capable of learning long-term dependencies in sequential data through the use of memory cells and gating mechanisms. LSTMs mitigate the vanishing gradient problem and are particularly effective for modeling temporal sequences, patterns, and time-series intrusion detection, where previous inputs significantly influence future predictions.

E. Evaluation Metrics

Evaluation metrics are used to assess the robustness and limitations of the proposed model, enabling necessary refinements to achieve optimal performance. The following metrics are employed in this study:

- Accuracy is the ratio of correctly predicted samples to the total number of samples:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (16)$$

- Precision is the ratio of true positive predictions to the total positive predictions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (17)$$

- Recall is the ratio of true positive predictions to the total actual positive instances:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (18)$$

- F1-score is the harmonic mean of precision and recall, balancing the two measures:

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (19)$$

- The Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) evaluates the trade-off between the true positive rate (TPR) and the false positive rate (FPR), providing a comprehensive view of classification performance across various thresholds.

$$AUC = \int_0^1 TPR(FPR)\, dFPR \quad (20)$$

A statistical significance test is conducted using the McNemar exact two-sided test. For accuracy:

$$b = \#\{i : \hat{y}_i^A \neq y_i, \hat{y}_i^B = y_i\} \quad (21)$$

where B indicates correct and A denotes wrong.

$$c = \#\{i : \hat{y}_i^A = y_i, \hat{y}_i^B \neq y_i\} \quad (22)$$

Then the exact p-value is:

$$X \sim \text{Bin}\left(b + c, \frac{1}{2}\right) \quad (23)$$

$$p_{Acc} = 2\min\{\Pr[X \leq \min(b,c)], \Pr[X \geq \max(b,c)]\}$$

For p-F1 (F1-score), define the F1 of a model on a set S as:

$$\text{F}_1(\text{S}) = \frac{2\text{TP(S)}}{2\text{TP(S)+FP(S)+FN(S)}} \quad (24)$$

F. Ablation Study

An ablation study is conducted to evaluate the contribution of each component within the proposed framework. The proposed pipeline was tested under the following configurations:

- Without the Boruta feature selection.
- Without semi-supervised outlier removal.

- Without both.
- With the full pipeline.

A comparative analysis of these configurations quantifies the individual and combined impacts of Boruta-based feature selection and semi-supervised outlier removal on overall model performance.

III. RESULTS AND DISCUSSION

On the CIC-IDS2017 dataset, both components independently enhanced model performance, while their combination yielded the best overall results. Across all datasets, the contribution of each component varies depending on the dataset and model used (see Table I). Removing either component resulted in a noticeable decline in both accuracy and F1-score, confirming their complementary roles within the proposed pipeline.

TABLE I. ABLATION STUDY: PERFORMANCE OF DIFFERENT PIPELINE CONFIGURATIONS ON CIC-IDS2017

Full pipeline (Boruta + outlier removal)					
	Accuracy %	**Precision %**	**Recall %**	**F1-score %**	**AUC-ROC %**
RF	99.91	99.91	99.91	99.91	100.00
LR	92.07	93.74	92.07	92.07	98.19
GB	99.63	99.63	99.63	99.63	99.64
CNN	97.92	98.00	97.92	97.94	99.79
LSTM	96.41	96.56	96.41	96.45	99.36
w/o Boruta (with outlier removal)					
RF	99.91	99.91	99.91	99.91	99.99
LR	92.07	93.74	92.07	92.45	98.17
GB	99.62	99.62	99.62	99.62	99.93
CNN	98.02	98.11	98.02	98.04	99.80
LSTM	95.82	95.82	95.82	95.82	99.10
w/o outlier removal (with Boruta)					
RF	99.91	99.91	99.91	99.91	99.99
LR	91.90	93.67	91.90	92.30	97.94
GB	99.63	99.63	99.63	99.63	99.95
CNN	97.55	97.71	97.55	97.59	99.70
LSTM	94.68	94.68	94.68	94.56	98.37
Baseline (No Boruta, No outlier removal)					
RF	99.91	99.91	99.91	99.91	99.99
LR	91.69	93.56	91.69	92.12	97.92
GB	99.62	99.62	99.62	99.62	99.93
CNN	97.44	97.56	97.44	97.47	99.68
LSTM	94.26	94.82	94.26	94.41	98.39

Across the three intrusion-detection benchmarks, the Boruta+clustering (full) framework yields statistically significant differences in 12 out of 15 model dataset combinations (α =0.05; p_{Acc} via McNemar, p_{F1} via paired bootstrap). On CIC-IDS2017, the full framework significantly improved the performance of LR, GB, CNN, and LSTM models (RF showed no significant improvement in accuracy).

On the NSL-KDD dataset, the effects were model-dependent, since LR and CNN performed significantly better than the simplified baseline, while LSTM showed significant improvement using the full framework. On UNSW-NB15, all classifiers exhibited statistically significant differences that favored the simplified baseline by substantial margins. The exact two-sided McNemar p for accuracy and paired bootstrap result (B=10,000 resamples, 95% confidence interval) for F1-score are reported, computed on the same test instances. Table III indicates statistically significant results with $p < 0.05$.

TABLE II. CLASSIFICATION PERFORMANCE OF THE PROPOSED FRAMEWORK ON THREE DATASETS

NSL-KDD					
	Accuracy %	**Precision %**	**Recall %**	**F1-score %**	**AUC-ROC %**
RF	99.96	99.96	99.96	99.96	100
LR	98.12	98.12	98.12	98.06	99.67
GB	99.58	99.58	99.58	99.57	99.99
CNN	99.41	99.42	99.41	99.41	99.98
LSTM	99.20	99.20	99.20	99.20	99.97
UNSW-NB15					
RF	94.90	94.91	94.90	94.90	99.17
LR	90.97	91.06	90.97	91.00	97.81
GB	93.29	93.29	93.29	93.29	98.62
CNN	91.07	91.12	91.07	90.97	97.86
LSTM	90.69	90.88	90.69	90.53	97.83
CIC-IDS2017					
RF	99.91	99.91	99.91	99.91	100
LR	92.07	93.74	92.07	92.07	98.19
GB	99.63	99.63	99.63	99.63	99.64
CNN	97.92	98.00	97.92	97.94	99.79
LSTM	96.41	96.56	96.41	96.45	99.36

TABLE III. SIGNIFICANCE TABLE

CIC-IDS2017			
Classifier	p_{Acc}	p_{F1}	**Significance**
RF	0.0662	0.0476	No
LR	0.0000	0.0000	Yes
GB	1.458E-09	0.0000	Yes
CNN	0.0000	0.0000	Yes
LSTM	0.0000	0.0000	Yes
NSL-KDD			
RF	0.3750	0.1820	No
LR	0.0003	0.0000	Yes
GB	0.2717	0.2140	No
CNN	2.353E-11	0.0000	Yes
LSTM	7.9E-37	0.0000	Yes
UNSW-NB15			
RF	0.0000	0.0000	Yes
LR	0.0000	0.0000	Yes
GB	0.0000	0.0000	Yes
CNN	0.0000	0.0000	Yes
LSTM	0.0000	0.0000	Yes

The proposed framework was compared against several recent state-of-the-art IDSs. Most recent IDS studies focus exclusively on either feature selection or outlier handling, with few offering a unified, adaptive preprocessing pipeline that maintains generalization across multiple datasets. The main novelties of the proposed framework are as follows:

- Integration: Systematic combination of Boruta-based feature selection and semi-supervised outlier removal within a single workflow.
- Generalizability: Comprehensive cross-dataset validation using three heterogeneous datasets (NSL-KDD, UNSW-NB15, and CIC-IDS2017).
- Ablation analysis: Detailed evaluation of the contribution of each component to overall performance.

- Transparency: Emphasis on interpretability and reproducibility through open-source implementation and unified experimental scripts.

Table IV presents a comparison of the proposed framework and recent state-of-the-art (SOTA) intrusion detection system (IDS) methods.

TABLE IV. COMPARISON OF THE PROPOSED FRAMEWORK WITH RECENT STATE-OF-THE-ART IDS METHODS

Ref.	Dataset(s)	Core method	Gap/limitation
[9]	NSL-KDD; KDD-Cup99	Adaptive Walrus Optimization (feature selection) + EANFIS; ACST encryption	Focus on classification and cloud; older datasets; no Boruta; no semi-supervised outlier filtering
[13]	UNSW-NB15; CIC-IDS2017	MI-Boruta feature selection + stacked ensemble (RF, CatBoost, XGBoost, MLP)	Only two datasets; no semi-supervised outlier filtering
[18]	UNSW-NB15; others	Federated anomaly detection with differential privacy	Privacy focus; no Boruta or clustering; federated only
[19]	IoTID20	Ensemble feature selection (Variance, MI, Chi-square, ANOVA, L1) + RFE; GRU, CNN	Single dataset; supervised DL; no Boruta; no semi-supervised outlier filtering
[20]	NSL-KDD; BoT-IoT	MissForest imputation + CALR feature selection + GA-LSTM with attention + EPSS	No Boruta; no semi-supervised outlier clustering; IoT-focused
[21]	NSL-KDD; UNSW-NB15; CIC-IDS2017	Optimized XGBoost + OSNN with hyperparameter tuning and imbalance handling	Fully supervised; no Boruta; no semi-supervised outlier filtering; cross-dataset generalization not analyzed
[22]	Vehicular CAM/DSRC	Probabilistic ML IDS with PVRS metric	EV/V2V-specific; no general IDS datasets

In contrast, most SOTA approaches:

- Employ only filter-based or wrapper-based feature selection without any outlier removal mechanism [19].
- Apply outlier or anomaly detection without advanced feature selection strategies [23].
- Rely heavily on DL models without rigorous ablation or multi-dataset benchmarking [20].
- Evaluate performance on a single dataset, limiting generalizability [18].

The proposed framework matches or exceeds recent SOTA results on CIC-IDS2017, while on UNSW-NB15, several models exhibit better results with the baseline configuration. Nevertheless, the key contribution of this work lies in its unified, statistically validated pipeline, supported by thorough ablation and significance testing across diverse datasets.

IV. CONCLUSION

This study introduced a modular intrusion detection framework that integrates Boruta-based feature selection with class-wise, semi-supervised outlier removal and evaluated its performance on the NSL-KDD, UNSW-NB15, and CIC-IDS2017 datasets. Across benchmarks, the proposed framework consistently improved performance on CIC-IDS2017 for most classifiers (particularly LR, GB, CNN, and LSTM), exhibited model-dependent effects on NSL-KDD, and remained competitive with the baseline on UNSW-NB15. These results suggest partial cross-dataset generalization, with the most pronounced benefits observed in heterogeneous and noisier traffic data.

Ablation studies further demonstrate that both Boruta feature selection and outlier removal components contribute significantly to performance gains on CIC-IDS2017, while their influence varies across other datasets and models. Paired statistical significance tests confirm that the observed improvements are statistically valid and also highlight model dataset pairs showing comparable or baseline-favored performance. From a deployment perspective, the framework is advantageous due to its modular design, allowing components to be selectively activated based on operational constraints and dataset characteristics. However, this study is limited by its dependence on publicly available benchmarks, the offline batch evaluation setting, fixed hyperparameters and thresholds, and the computational overhead introduced by the feature selection and clustering processes.

REFERENCES

[1] E. S. Shombot, G. Dusserre, R. Bestak, and N. B. Ahmed, "An application for predicting phishing attacks: A case of implementing a support vector machine learning model," *Cyber Security and Applications*, vol. 2, 2024, Art. no. 100036, https://doi.org/10.1016/j.csa.2024.100036.

[2] C. M. Nalayini, J. Katiravan, S. Geetha, and C. J. I. Eunaicy, "A novel dual optimized IDS to detect DDoS attack in SDN using hyper tuned RFE and deep grid network," *Cyber Security and Applications*, vol. 2, 2024, Art. no. 100042, https://doi.org/10.1016/j.csa.2024.100042.

[3] T. B. Shana, N. Kumari, M. Agarwal, S. Mondal, and U. Rathnayake, "Anomaly-based intrusion detection system based on SMOTE-IPF, Whale Optimization Algorithm, and ensemble learning," *Intelligent Systems with Applications*, vol. 27, Sept. 2025, Art. no. 200543, https://doi.org/10.1016/j.iswa.2025.200543.

[4] A. Grandhi and S. K. Singh, "Interrelated dynamic biased feature selection and classification model using enhanced gorilla troops optimizer for intrusion detection," *Alexandria Engineering Journal*, vol. 114, pp. 312–330, Feb. 2025, https://doi.org/10.1016/j.aej.2024.10.100.

[5] F. Alhayan *et al.*, "Design of advanced intrusion detection in cybersecurity using ensemble of deep learning models with an improved beluga whale optimization algorithm," *Alexandria Engineering Journal*, vol. 121, pp. 90–102, May 2025, https://doi.org/10.1016/j.aej.2025.02.069.

[6] J. Wang *et al.*, "A Two-Layer Network Intrusion Detection Method Incorporating LSTM and Stacking Ensemble Learning," *Computers, Materials & Continua*, vol. 83, no. 3, pp. 5129–5153, 2025, https://doi.org/10.32604/cmc.2025.062094.

[7] "NSL-KDD." Canadian Institute for Cybersecurity, 2009, [Online]. Available: https://www.unb.ca/cic/datasets/nsl.html.

[8] S. Choudhary and N. Kesswani, "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT," *Procedia Computer Science*, vol. 167, pp. 1561–1573, 2020, https://doi.org/10.1016/j.procs.2020.03.367.

[9] K V K. Chithanya and L. V. Reddy, "Automatic intrusion detection model with secure data storage on cloud using adaptive cyclic shift transposition with enhanced ANFIS classifier," *Cyber Security and Applications*, vol. 3, Dec. 2025, Art. no. 100073, https://doi.org/10.1016/j.csa.2024.100073.

[10] N. Moustafa, J. Slay, and G. Creech, "Novel Geometric Area Analysis Technique for Anomaly Detection Using Trapezoidal Area Estimation

on Large-Scale Networks," *IEEE Transactions on Big Data*, vol. 5, no. 4, pp. 481–494, Dec. 2019, https://doi.org/10.1109/TBDATA.2017.2715166.

[11] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," *Information Security Journal: A Global Perspective*, vol. 25, no. 1–3, pp. 18–31, Apr. 2016, https://doi.org/10.1080/19393555.2015.1125974.

[12] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, Canberra, Australia, Nov. 2015, pp. 1–6, https://doi.org/10.1109/MilCIS.2015.7348942.

[13] A. M. Alsaffar, M. Nouri-Baygi, and H. M. Zolbanin, "Shielding networks: enhancing intrusion detection with hybrid feature selection and stack ensemble learning," *Journal of Big Data*, vol. 11, no. 1, Sept. 2024, Art. no. 133, https://doi.org/10.1186/s40537-024-00994-7.

[14] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization:," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, Madeira, Portugal, 2018, pp. 108–116, https://doi.org/10.5220/0006639801080116.

[15] I. H. Hassan, M. Abdullahi, M. M. Aliyu, S. A. Yusuf, and A. Abdulrahim, "An improved binary manta ray foraging optimization algorithm based feature selection and random forest classifier for network intrusion detection," *Intelligent Systems with Applications*, vol. 16, Nov. 2022, Art. no. 200114, https://doi.org/10.1016/j.iswa.2022.200114.

[16] G. O. Anyanwu, C. I. Nwakanma, J. M. Lee, and D. S. Kim, "Novel hyper-tuned ensemble Random Forest algorithm for the detection of false basic safety messages in Internet of Vehicles," *ICT Express*, vol. 9, no. 1, pp. 122–129, Feb. 2023, https://doi.org/10.1016/j.icte.2022.06.003.

[17] O. H. Abdulganiyu, T. A. Tchakoucht, A. E. H. Alaoui, and Y. K. Saheed, "Attention-driven multi-model architecture for unbalanced network traffic intrusion detection via extreme gradient boosting," *Intelligent Systems with Applications*, vol. 26, June 2025, Art. no. 200519, https://doi.org/10.1016/j.iswa.2025.200519.

[18] A. Alabdulatif, "GuardianAI: Privacy-preserving federated anomaly detection with differential privacy," *Array*, vol. 26, July 2025, Art. no. 100381, https://doi.org/10.1016/j.array.2025.100381.

[19] T. Q. Al-Ghadi, S. Manickam, I. D. M. Widia, E. R. N. Wulandari, and S. Karuppayah, "Leveraging federated learning for DoS attack detection in IoT networks based on ensemble feature selection and deep learning models," *Cyber Security and Applications*, vol. 3, Dec. 2025, Art. no. 100098, https://doi.org/10.1016/j.csa.2025.100098.

[20] D. M. Dhanvijay, M. M. Dhanvijay, and V. H. Kamble, "Cyber intrusion detection using ensemble of deep learning with prediction scoring based optimized feature sets for IOT networks," *Cyber Security and Applications*, vol. 3, Dec. 2025, Art. no. 100088, https://doi.org/10.1016/j.csa.2025.100088.

[21] F. S. Alsubaei, "Smart deep learning model for enhanced IoT intrusion detection," *Scientific Reports*, vol. 15, no. 1, July 2025, Art. no. 20577, https://doi.org/10.1038/s41598-025-06363-5.

[22] D. Kosmanos *et al.*, "A novel Intrusion Detection System against spoofing attacks in connected Electric Vehicles," *Array*, vol. 5, Mar. 2020, Art. no. 100013, https://doi.org/10.1016/j.array.2019.100013.

[23] F. J. Abdullayeva, "Distributed denial of service attack detection in E-government cloud via data clustering," *Array*, vol. 15, Sept. 2022, Art. no. 100229, https://doi.org/10.1016/j.array.2022.100229.