IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 May 2026, accepted 6 June 2026, date of publication 15 June 2026, date of current version 19 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3703481

RESEARCH ARTICLE

SHAP-Guided Feature Selection for Cross-Dataset Generalization in Network Intrusion Detection Systems

CAN KILIÇ[1] AND GÖKHAN ŞENGÜL[2]
[1]Department of Software Engineering, Atılım University, 06830 Ankara, Türkiye
[2]Department of Computer Engineering, Atılım University, 06830 Ankara, Türkiye
Corresponding author: Can Kılıç (kiliccan.business@gmail.com)

ABSTRACT Flow-based machine learning intrusion detection systems (IDS) often achieve near-perfect performance when trained and tested on a single benchmark dataset; nonetheless, their ability to generalize across datasets is a crucial and mostly unresolved challenge. This study analyzes the cross-dataset generalization behavior of an explainable, flow-based IDS trained on CICIDS2017 and externally evaluated on the CSE-CIC-IDS2018 dataset, which represents a more realistic network environment with varying attack implementations, traffic compositions, and background services. Two frequently used ensemble models, Random Forest and XGBoost, are trained solely on flow-level metadata without packet payload examination. After removing non-behavioral identifiers (Flow ID, Source IP, Destination IP, and Timestamp) and harmonizing feature schemas, the datasets are aligned into a unified 80-dimensional feature space extracted with CICFlowMeter. SHAP (TreeSHAP) is used to calculate global feature importance and create multiple explainability-driven feature subsets, such as model-specific Top-20 sets, a COMMON-10 intersection, and a UNION-30 superset. Although both models attain near-perfect accuracy and weighted F1-scores on CICIDS2017 (macro-F1 $\approx$ 0.90), when evaluated on CSE-CIC-IDS2018, macro-F1 drops to 0.127 for Random Forest and 0.119 for XGBoost, despite high overall accuracy, indicating a strong bias toward majority classes under domain shift conditions. SHAP-guided feature reduction provides a measurable but limited improvement for Random Forest, increasing macro-F1 from 0.127 to 0.166, while an additional port-removal ablation further improves macro-F1 to 0.207. In contrast, no significant cross-dataset improvement is observed for XGBoost. An additional practical observation is that SHAP-guided feature rankings remain highly stable across sample sizes: class-balanced subsets of approximately 400 flows (50 samples per class) produce highly similar Top-20 rankings to those obtained from 10,000 flows (1250 samples per class), supporting the feasibility of computationally efficient explainability. Overall, the results show that explainability-driven feature analysis improves transparency, compactness, and feature prioritization; however, it does not fully resolve the broader distributional shift challenges that limit cross-dataset generalization in flow-based intrusion detection systems.

INDEX TERMS Cross-dataset generalization, explainable artificial intelligence, flow-based traffic analysis, network intrusion detection, random forest, SHAP, XGBoost.

I. INTRODUCTION

Network intrusion detection systems (NIDS) play an important role in protecting modern computer networks from increasingly complex cyber threats. As encrypted communication has become the primary type of Internet traffic, traditional payload-based inspection methods have lost access to application-layer content. Protocols like HTTPS, TLS, and VPN-based tunneling preserve confidentiality while limiting deep packet inspection capabilities. As a result,

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

flow-based intrusion detection systems that use statistical metadata, such as flow durations, inter-arrival times, byte counts, and transport-layer attributes, have emerged as a viable and privacy-preserving option for large-scale traffic monitoring [1], [11], [14].

Machine learning (ML)-based NIDS have received significant attention due to their ability to automatically learn discriminative patterns from flow-level data without relying on manually crafted signatures. Publicly available benchmark datasets, particularly CICIDS2017 [1], [13] and its successor CSE-CIC-IDS2018 [2], have enabled rapid development and evaluation of ML-based intrusion detection models. Numerous studies report near-perfect classification performance when models are trained and tested on the same dataset, with ensemble methods such as Random Forest and XGBoost frequently achieving accuracy and weighted F1-scores above 99% in intra-dataset settings [8], [9], [15], [18].

However, high performance under controlled evaluation conditions does not necessarily translate to real-world robustness. Prior work has repeatedly shown that models trained on a single benchmark dataset may suffer substantial degradation when evaluated on different datasets or operational environments [3], [17], [19]. This phenomenon is commonly attributed to domain shift, where differences in traffic composition, background services, attack implementations, and data collection procedures alter the statistical properties of the feature space. Even when feature definitions remain consistent across datasets, distributional shifts can cause models to over-rely on dataset-specific artifacts rather than stable behavioral patterns.

Therefore, cross-dataset generalization is one of the most critical and under-addressed challenges in operational intrusion detection. Studies have shown that models trained on CICIDS2017 often exhibit sharp drops in macro-level performance when evaluated on CSE-CIC-IDS2018, frequently defaulting to benign-dominant predictions and failing to detect minority attack classes [3], [19]. This limitation is especially concerning in realistic deployment scenarios, where intrusion detection systems must operate in environments that differ from the training distribution and must detect emerging or zero-day attack variants. As highlighted by Sommer and Paxson [17], reliance on static benchmark datasets may lead to overly optimistic performance assessments that do not reflect deployment realities.

In addition to generalization challenges, the increasing complexity of machine learning models has prompted questions about interpretability and trustworthiness in cybersecurity applications. Security analysts demand transparent and explainable decision-making methods to audit warnings, validate model behavior, and identify probable failure factors. Explainable Artificial Intelligence (XAI) approaches have gained popularity in intrusion detection research [7], [12]. Among these approaches, SHAP (SHapley Additive Explanations) [10] has emerged as one of the most extensively accepted frameworks due to its solid theoretical foundation in cooperative game theory and ability to provide consistent, instance-level feature attribution.

SHAP assigns contribution scores to individual features for each prediction and allows aggregation of local explanations into global importance rankings. For tree-based ensemble models such as Random Forest [8] and XGBoost [9], the TreeSHAP algorithm enables efficient and exact computation of feature contributions [10]. Accordingly, these models are selected due to their strong performance on tabular flow-based data and their compatibility with exact SHAP (TreeSHAP) explanations, enabling consistent and interpretable feature attribution. Recent studies have used SHAP primarily to improve interpretability in intrusion detection contexts, including IoT security and deep learning-based models [4], [5], [6], [7]. More recent work has also started exploring explainable cross-domain evaluation settings for ML-based intrusion detection systems [25]. However, most of this literature focuses on post-hoc explanation within a single dataset, often under binary classification settings. Systematic evaluation of SHAP-driven feature analysis under external dataset deployment conditions remains relatively limited. Unlike most existing studies that use SHAP primarily for post-hoc interpretability within a single dataset, this work evaluates SHAP-guided feature subsets under external dataset deployment conditions, with a focus on multi-class classification and generalization behavior.

Beyond interpretability, SHAP has the potential to influence model design and feature selection. By finding features that regularly contribute to classification judgments, SHAP-guided analysis can assist in reducing reliance on false correlations and dataset-specific artifacts. However, certain highly ranked flow attributes, particularly port-related features, may still encode dataset-specific service behavior and therefore transfer poorly across different network environments. Consequently, actual evidence on whether SHAP-guided feature selection improves cross-dataset generalization is limited and sometimes conflicting, especially in multi-class intrusion detection contexts.

Motivated by these limitations, this study does not propose a domain adaptation technique; instead, it systematically evaluates the effect of explainability-driven feature selection under cross-dataset domain shift. Specifically, SHAP-guided feature subsets are derived from CICIDS2017 and used to analyze how feature selection influences generalization performance when models are deployed on CSE-CIC-IDS2018 without retraining or domain adaptation. This experimental setup reflects a realistic deployment scenario in which models must operate under distributional shift.

In addition to evaluating classification performance, this study further analyzes the stability of SHAP-guided global feature importance rankings under different sampling strategies. By comparing rankings obtained from small class-balanced subsets with those derived from larger samples, the analysis examines whether explainability-driven feature analysis can be performed efficiently while maintaining highly overlapping Top-k feature rankings. Reduced feature subsets,

including model-specific Top-20 sets, a COMMON-10 intersection, and a UNION-30 union set, are constructed based on SHAP rankings and systematically evaluated under cross-dataset conditions.

Rather than proposing a new detection model, this study focuses on analyzing the behavior of explainability-driven feature selection under domain shift and quantifying its effect on cross-dataset generalization. Accordingly, the contribution of this work lies in systematic empirical analysis and explainability-driven evaluation rather than architectural novelty.

The main contributions of this work are as follows:

(C1) A controlled cross-dataset evaluation framework in which models are trained on CICIDS2017 and deployed on CSE-CIC-IDS2018 without retraining, reflecting an external-dataset generalization scenario under domain shift conditions.

(C2) An explainability-driven feature selection protocol based on TreeSHAP global importance rankings, resulting in multiple reduced feature subsets (Top-20, COMMON-10, and UNION-30) designed to analyze feature robustness and transferability across datasets.

(C3) An empirical investigation of SHAP feature importance stability across sample sizes, suggesting that small, class-balanced subsets can approximate global feature rankings obtained from significantly larger samples with substantially reduced computational cost.

(C4) A comparative analysis of Random Forest and XGBoost under domain shift, highlighting model-dependent differences in how SHAP-guided feature reduction affects cross-dataset generalization performance.

Overall, this study aims to clarify the practical benefits and limitations of explainability-driven feature selection in realistic intrusion detection settings and to contribute to a more rigorous understanding of cross-dataset robustness in flow-based NIDS.

II. MATERIALS AND METHOD

This section covers the datasets, preprocessing methods, machine learning models, and evaluation protocol used in the study. The methodology is intended to isolate the effect of SHAP-guided feature selection on cross-dataset generalization by establishing a controlled experimental setup in which all models are trained under the same conditions. Attention is given to dataset harmonization, class imbalance handling, and explainability-driven feature stability analysis, allowing for a fair and reproducible assessment of model performance under domain shift. Rather than treating SHAP as a performance optimization mechanism, the methodology frames explainability as a diagnostic tool for analysing feature stability and generalization behaviour under domain shift.

A. DATASETS AND FEATURE SPACE ALIGNMENT

This study uses two benchmark intrusion detection datasets provided by the Canadian Institute for Cybersecurity: CICIDS2017 [1] and CSE-CIC-IDS2018 [2]. CICIDS2017 is only used for model training and internal evaluation, whereas CSE-CIC-IDS2018 is set up as an external test set to evaluate cross-dataset generalization in an external dataset deployment evaluation scenario.

Both datasets were created using CICFlowMeter and contain flow-level statistical features collected from network traffic without packet payload inspection. To ensure structural consistency, four non-behavioral identifiers were removed from both datasets: Flow ID, Source IP, Destination IP, and Timestamp. Following schema harmonization and feature name reconciliation, a single 80-dimensional feature space common to both datasets was generated. This alignment ensures that any reported performance decrease on CSE-CIC-IDS2018 derives from domain shift rather than feature mismatch.

During preprocessing, column names and label fields were standardized to ensure consistent formatting across all CSV files. Duplicate header rows appearing inside the datasets were removed, and empty strings were converted into missing values. Rows containing unresolved missing values were removed during the cleaning process. For CSE-CIC-IDS2018, missing values in Src Port, Dst Port, and Flow Byts/s were filled with zero to preserve feature-space consistency between datasets. In addition, infinite values generated by flow-rate calculations were replaced with zero to ensure numerical stability during training and evaluation. Final validation checks confirmed that the cleaned datasets contained no remaining missing, empty, or infinite values.

To ensure label consistency between CICIDS2017 and CSE-CIC-IDS2018, a label harmonization process was applied by grouping semantically similar attack types into broader categories. Specifically, all DoS-related attacks (DoS Hulk, DoS GoldenEye, DoS slowloris, and DoS Slowhttptest) were merged under the ''DoS'' class. Similarly, all DDoS variants (including DDoS, DDoS-HOIC, DDoS-LOIC-HTTP, and DDoS-LOIC-UDP) were grouped into a single ''DDoS'' category. Brute-force attacks such as FTP-Patator and SSH-Patator were combined under the ''Brute Force'' class. All web-based attacks (Web Attack – Brute Force, Web Attack – XSS, and Web Attack – SQL Injection) were merged into a unified ''Web Attack'' category. Other attack types, including Bot, PortScan, and Infiltration, were retained as separate classes. The Heartbleed class was removed due to its extremely limited number of samples in CICIDS2017 and the lack of representation in CSE-CIC-IDS2018. This harmonization process and the resulting unified label structure are summarized in Table 1, while the final harmonized class distributions for both datasets are presented in Table 2.

B. DATA SPLITTING AND CLASS IMBALANCE HANDLING

CICIDS2017 shows an extraordinary class imbalance, with benign traffic dominating the dataset and some attack classes receiving very little support. To maintain natural class

TABLE 1. Attack label original sample distributions.

Attack Type	Sample Count (2017)	Sample Count (2018)
BENIGN	2,272,688	13,425,831
DoS Hulk	230,124	461,912
PortScan	158,930	-
DDoS	128,027	-
DDOS attack-HOIC	-	686,012
DdoS attacks-LOIC-HTTP	-	576,191
DDOS attack-LOIC-UDP	-	1,730
DoS GoldenEye	10,293	41,508
DoS slowloris	5,796	10,990
DoS Slowhttptest	5,499	139,890
FTP-Patator / FTP-BruteForce	7,938	193,354
SSH-Patator / SSH-Bruteforce	5,897	187,589
Bot	1,966	286,191
Web Attack – Brute Force	1,507	611
Web Attack – XSS	652	230
Web Attack – SQL Injection	21	87
Infiltration	36	161,096
Heartbleed	11	-

TABLE 2. Final 8-Class label distribution for CICIDS2017 and CSE-CIC-IDS2018.

Attack Type	Sample Count (2017)	Sample Count (2018)
Benign	2,272,688	13,425,831
DoS	251,712	654,300
DDoS	128,027	1,263,933
PortScan	158,930	-
Brute Force	13,835	380,943
Bot	1,966	286,191
Web Attack	2,180	928
Infiltration	36	161,096

distributions during assessment, the dataset was stratified into 64% training, 16% validation, and 20% internal test subsets.

To reduce majority-class dominance during training, the Synthetic Minority Over-Sampling Technique (SMOTE) [24] was applied exclusively to the training split. This restriction ensures that oversampling does not introduce information leakage across evaluation splits and preserves the validity of cross-dataset performance comparisons. Minority attack classes were expanded to appropriate sample numbers without disturbing the hierarchy of existing attack types, while validation and test sets were preserved without oversampling to better reflect realistic traffic conditions. This technique reduces evaluation bias and reflects realistic deployment conditions.

SMOTE oversampling was applied to the minority Bot, Infiltration, and Web Attack classes using target sample sizes of 4500, 1500, and 5000 instances, respectively. All oversampling operations were performed exclusively on the CICIDS2017 training split using a fixed random seed (random_state = 42).

C. MACHINE LEARNING MODELS AND HYPERPARAMETERS

Ensemble-based learning approaches have consistently performed well on structured, tabular network traffic data, especially in flow-based intrusion detection settings. Unlike deep neural architectures, which frequently require extensive feature engineering and large-scale tuning, tree-based ensemble models achieve a good balance between predictive accuracy, robustness to noisy statistical data, and interpretability. Given the goal of this study, which is to investigate explainability-driven feature selection under cross-dataset domain shift, it is crucial to select models that are both high-performing and fully compatible with exact SHAP computation.

For this reason, two complementary ensemble models were chosen: Random Forest (RF) [8], a bagging-based variance reduction framework, and XGBoost (XGB) [9], a gradient boosting optimization framework. These models differ fundamentally in how they construct ensembles and handle bias-variance trade-offs, allowing for a controlled comparison of SHAP-guided feature reduction across various learning processes.

To ensure that performance variations are completely due to feature selection rather than model adjustment, hyperparameters were intentionally fixed across all experimental settings. This controlled design examines the impact of SHAP-guided feature subsets on in-domain and cross-dataset generalization performance.

1) RANDOM FOREST

Random Forest (RF) was selected in this study as a robust bagging-based ensemble model suitable for structured flow-level intrusion detection data. RF constructs a collection of decision trees and aggregates their predictions through majority voting for classification tasks [8]. The core idea behind RF is bagging (bootstrap aggregating), where each tree is trained independently on a bootstrap sample drawn with replacement from the original training data. Consequently, different trees observe different subsets of the data distribution, which promotes model diversity and reduces variance.

In addition to bootstrap sampling, RF adds more randomness to tree construction by selecting a random subset of the features at each split rather than evaluating all available features. This randomized feature selection decreases inter-tree correlation and prevents the ensemble from depending too heavily on a limited number of dominating predictors. By averaging many weakly correlated decision trees, RF significantly reduces variance while preserving low bias, resulting in better generalization on unknown data. Essentially, the final class prediction of a Random Forest classifier is obtained by aggregating individual tree predictions via majority voting [8].

A key strength of RF lies in its robustness to noisy, redundant, and highly correlated features. Flow-based network traffic datasets typically contain statistical attributes such as packet length distributions, inter-arrival time metrics,

directional byte counts, and port-related features that exhibit strong inter-feature correlations and dataset-specific regularities. RF mitigates the adverse effects of such redundancy by distributing feature usage across multiple trees, thereby reducing sensitivity to spurious correlations and noise. Unlike boosting-based techniques that iteratively focus on residual errors, RF relies on ensemble averaging, which generally lowers the risk of severe overfitting under distributional shift [8].

An additional practical advantage of RF is the availability of out-of-bag (OOB) error estimation. Since each tree is trained on a bootstrap sample, approximately one-third of the training instances are excluded from that tree's construction. These samples can be used to estimate generalization performance without requiring a separate validation set [8]. Although explicit validation and test splits are employed in this work to ensure controlled cross-dataset evaluation, OOB estimation further highlights RF's suitability for real-world deployment.

From an explainability perspective, RF is fully compatible with the TreeSHAP algorithm [10]. Because RF is a tree-based ensemble, TreeSHAP can compute exact Shapley values efficiently by exploiting the internal tree structure. This enables both instance-level explanations and stable aggregation of global feature importance across large-scale flow-based intrusion detection datasets. The compatibility between RF and TreeSHAP is particularly important in this study, where SHAP-guided feature subsets are systematically evaluated under cross-dataset domain shift.

Overall, Random Forest provides a variance-reducing ensemble framework that is robust to noisy and correlated features, resistant to overfitting, and natively compatible with SHAP-guided explainability. These characteristics make RF an appropriate and interpretable baseline for analyzing feature stability and cross-dataset generalization in flow-based network intrusion detection systems [8].

2) XGBoost

XGBoost (XGB) was selected as a complementary boosting-based ensemble model to contrast with the bagging-oriented structure of Random Forest. While RF reduces variance through parallel tree aggregation, XGBoost constructs trees sequentially, where each new tree is trained to correct the residual errors of the previously learned ensemble [9]. XGBoost's repeated refinement method allows it to model complex nonlinear interactions and subtle feature dependencies that bagging-based algorithms may not fully capture.

XGBoost trains the model progressively. At each boosting round, a new decision tree is created to reduce the residual errors of the previous ensemble. The training method aims to reduce the loss function, which quantifies the difference between predicted and true labels. Unlike traditional gradient boosting algorithms, XGBoost has additional limitations to regulate model complexity, such as restricting tree depth and leaf weights [9]. These limitations help to keep the model from getting too complex and lower the risk of overfitting.

Another important aspect of XGBoost is its utilization of both gradient and second-order (Hessian) information when creating trees. It analyzes how quickly the mistake varies in addition to its direction (first-order gradient). This enables the model to select better splits, enhance convergence speed, and achieve more stable training than classic gradient boosting approaches [9].

XGBoost was also created with computational efficiency in mind. It features parallel tree construction, fast sparse input processing, and memory-aware optimization algorithms, making it ideal for large-scale, multidimensional tabular datasets. These characteristics have led to its widespread use in structured data modeling tasks, such as flow-based intrusion detection systems.

From an explainability aspect, XGBoost is completely compatible with the TreeSHAP algorithm [10]. Because it is a tree-based ensemble, accurate Shapley values may be computed efficiently using the internal tree structure. This permits consistent feature attribution across tests and allows for a systematic comparison of full-feature and SHAP-reduced configurations during domain transition.

In this study, XGBoost is used as a high-capacity gradient boosting baseline to assess the effect of SHAP-guided feature reduction. Its capacity to capture fine-grained feature interactions, together with built-in regularization and TreeSHAP compatibility, make it ideal for determining if explainability-driven feature selection improves cross-dataset generalization performance [9].

To achieve a controlled and fair comparison of feature configurations, both models' hyperparameters were intentionally held constant throughout all tests. This design decision separates the impact of SHAP-guided feature reduction from any performance changes caused by model tuning. Using two fundamentally different ensemble paradigms, bagging-based Random Forest and boosting-based XGBoost, under identical experimental constraints, this study allows for a structured analysis of how explainability-driven feature selection interacts with distinct learning mechanisms during cross-dataset domain shift. The following part explains the SHAP framework and describes how feature attributions were computed and examined to create the reduced feature subsets used in this study.

D. SHAP COMPUTATION AND FEATURE STABILITY ANALYSIS

In cross-dataset intrusion detection, understanding which features drive model predictions is essential for diagnosing performance degradation under domain shift. For this purpose, SHAP (SHapley Additive Explanations) [10] was employed as the primary feature attribution method in this study.

SHAP provides instance-level feature contributions based on cooperative game theory and enables consistent aggregation of local explanations into global importance rankings [10]. Unlike impurity-based importance measures, SHAP quantifies the marginal contribution of each feature

to the model's output. This property is particularly important in flow-based intrusion detection, where correlated statistical attributes and dataset-specific patterns may otherwise distort feature importance interpretation.

For tree-based models such as Random Forest and XGBoost, SHAP values were computed using the TreeSHAP algorithm, which leverages the structure of decision trees to compute feature contributions efficiently [10]. This makes SHAP computation feasible for large-scale datasets such as CICIDS2017. In this study, global feature importance rankings were computed by averaging the absolute SHAP values across all samples within each subset.

To evaluate the stability of these rankings, SHAP values were computed for two class-balanced subsets of CICIDS2017 containing 400 flows (50 samples per class) and 10,000 flows (1250 samples per class). These subsets were constructed using stratified sampling to ensure equal representation from each class. Despite the substantial difference in sample size, both configurations produced highly similar global feature rankings. Random Forest produced an identical Top-20 feature subset across both sampling configurations (20/20 overlap), while XGBoost retained 17 out of 20 Top-ranked features, with only minor variations among lower-ranked attributes. These findings suggest that smaller balanced subsets can reasonably approximate global SHAP rankings while reducing computational cost in large-scale intrusion detection analysis.

Based on the global SHAP rankings, multiple reduced feature subsets were constructed to evaluate explainability-driven feature selection under cross-dataset domain shift. These subsets include model-specific Top-20 feature sets, a COMMON-10 subset representing the intersection of RF and XGB Top-20 features, and a UNION-30 subset formed by combining both Top-20 lists.

E. CROSS-DATASET EVALUATION PROTOCOL

All models were trained solely on CICIDS2017 and tested in-domain (validation and internal test splits) and cross-dataset on CSE-CIC-IDS2018, with no retraining or adaptation. This setup represents a cross-dataset domain-shift evaluation scenario in which models must generalize to new traffic patterns, background services, and attack implementations.

Model performance was evaluated using the accuracy, macro-F1, weighted-F1, and confusion matrices. Because of the substantial class imbalance and domain shift, macro-F1 was chosen as the key metric of generalization performance, while accuracy and weighted-F1 were reported for completeness. In addition, precision, recall, macro-F1, and confusion matrix analyses were used to further characterize minority-class degradation under cross-dataset conditions.

F. HYPERPARAMETERS

The hyperparameters of the models were kept fixed across all experiments to isolate the effect of SHAP-guided feature selection. All experiments were conducted using a fixed random seed (random_state = 42) to ensure consistency and reproducibility across runs.

For Random Forest, the following parameters were used: n_estimators = 300, max_depth = None, min_samples_split=10, min_samples_leaf=4, and max_features="sqrt".

For XGBoost, the following parameters were used: objective = "multi:softprob", eval_metric = "mlogloss", learning_rate = 0.1, max_depth = 8, subsample = 0.8, colsample_bytree = 0.8, and num_boost_round = 300.

III. RESULT

This section reports the empirical findings of the proposed experimental framework, including SHAP-guided feature importance analysis, baseline performance under full feature configurations, and the cross-dataset effects of explainability-driven feature reduction.

A. SHAP-GUIDED FEATURE IMPORTANCE RESULTS

This section presents the global SHAP-guided feature importance results from models trained on CICIDS2017. TreeSHAP was used to calculate SHAP values and identify the most influential flow-level features contributing to multi-class intrusion detection decisions.

To assess the reliability of SHAP-guided feature rankings, SHAP values were computed for two class-balanced sample sizes: 400 flows (50 samples per class) and 10,000 flows (1250 samples per class). The Top-20 feature rankings produced by the Random Forest and XGBoost models showed highly consistent Top-20 feature structures across both sample sizes, with only minor variations in lower-ranked features.

The similarity between the two Random Forest SHAP rankings was quantified using Spearman's rank correlation coefficient ($\rho = 0.991$, $p < 0.001$), indicating extremely strong agreement between the two sampling configurations. Random Forest also produced an identical Top-20 feature subset across both sample sizes (20/20 overlap), while XGBoost retained 17 out of 20 top-ranked features, with only minor variations among lower-ranked attributes.

Table 3 lists the top 20 SHAP-ranked features for the Random Forest model. The highest-ranked features mainly include packet length statistics, directional byte volumes, and transport-layer attributes, including port information and TCP window sizes.

Table 4 shows the top 20 SHAP-ranked features for the XGBoost model. In addition to packet size and port-related attributes, XGBoost prioritizes temporal and rate-based data, such as inter-arrival time statistics and packet rates.

Comparing the RF and XGBoost SHAP rankings shows that several features appear in both Top-20 lists. These overlapping features were used to construct the COMMON-10 subset, while the UNION-30 subset was formed by combining the model-specific Top-20 feature sets.

B. BASELINE PERFORMANCE WITH THE FULL FEATURE SET

This subsection describes the baseline performance of Random Forest and XGBoost models trained on the entire

TABLE 3. Feature importance rankings for random forest.

Rank	Feature	Importance
1	Destination Port	0.015536
2	Backward Packet Length Mean	0.010384
3	Subflow Forward Bytes	0.009675
4	Total Length of Forward Packets	0.009634
5	Backward Packet Length Standard Deviation	0.009183
6	Average Packet Size	0.009060
7	Backward Segment Size Average	0.009058
8	Packet Length Variance	0.008229
9	Packet Length Mean	0.008206
10	Packet Length Standard Deviation	0.007609
11	Backward Packet Length Maximum	0.007137
12	Forward Segment Size Average	0.006735
13	Initial Backward Window Bytes	0.006684
14	Forward Packet Length Mean	0.006493
15	Forward Packet Length Maximum	0.006267
16	Source Port	0.005907
17	Subflow Backward Bytes	0.005678
18	Total Length of Backward Packets	0.005537
19	Packet Length Maximum	0.005514
20	Forward Header Length	0.004923

TABLE 4. Feature importance rankings for XGBoost.

Rank	Feature	Importance
1	Destination Port	1.293027
2	Initial Backward Window Bytes	0.514623
3	Initial Forward Window Bytes	0.504486
4	Source Port	0.439053
5	Forward Segment Size Minimum	0.370087
6	Forward Inter-Arrival Time Minimum	0.312961
7	Subflow Forward Bytes	0.260599
8	Backward Packets Per Second	0.249778
9	Flow Bytes Per Second	0.203952
10	Backward Packet Length Mean	0.197894
11	Flow Inter-Arrival Time Minimum	0.185461
12	Backward Packet Length Standard Deviation	0.155528
13	Forward Header Length	0.147951
14	Packet Length Mean	0.141228
15	Average Packet Size	0.138872
16	Flow Duration	0.137775
17	Forward Packet Length Maximum	0.121138
18	Forward Inter-Arrival Time Total	0.110721
19	PSH Flag Count	0.107222
20	Flow Inter-Arrival Time Maximum	0.106007

80-feature space. All models were trained using CICIDS2017 and tested in-domain and across datasets with identical experimental settings.

Table 5 highlights the in-domain performance statistics for CICIDS2017, which include validation and internal test splits. Both models achieved near-perfect in-domain performance, with accuracy values exceeding 0.999 and macro-F1 scores reaching up to 0.9941 for XGBoost on the internal test set. However, despite these strong in-domain results, substantial degradation was observed during cross-dataset evaluation.

Table 6 demonstrates the cross-dataset performance results obtained by applying the same models to CSE-CIC-IDS2018. Macro-F1 dropped from near-perfect in-domain values to 0.1273 for Random Forest and 0.1192 for XGBoost during cross-dataset evaluation, despite accuracy remaining around 0.83, indicating a strong bias toward majority-class predictions.

TABLE 5. Baseline in-domain performance on CICIDS2017 (Full 80-Feature set).

Model	Split	Accuracy	Macro-F1	Weighted-F1
Random Forest	Validation	0.9993	0.8997	0.9993
Random Forest	Internal Test	0.9993	0.9885	0.9993
XGBoost	Validation	0.9998	0.9049	0.9998
XGBoost	Internal Test	0.9998	0.9941	0.9998

TABLE 6. Baseline cross-dataset performance on CSE-CIC-IDS2018 (Full 80-Feature set).

Model	Accuracy	Macro-F1	Weighted-F1
Random Forest	0.8310	0.1273	0.7554
XGBoost	0.8310	0.1192	0.7553

Figure 1 presents the confusion matrix obtained for Random Forest on CSE-CIC-IDS2018 using the full 80-feature configuration. The matrix shows that, despite high overall accuracy, the model exhibits substantial misclassification across minority attack classes under cross-dataset conditions.

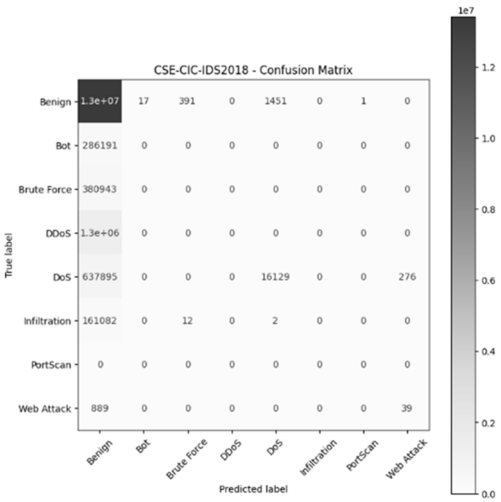

FIGURE 1. Confusion matrix for random forest on CSE-CIC-IDS2018 (Full 80 Features).

C. PERFORMANCE WITH SHAP-GUIDED FEATURE REDUCTION AND ABLATION ANALYSIS (RANDOM FOREST)

This subsection examines the performance of Random Forest models trained on SHAP-guided feature subsets, including RF Top-20, COMMON-10, UNION-30, RF Top-20-NoPort, and a manually constructed non-SHAP feature subset (Non-SHAP-20).

Table 7 summarizes the in-domain performance findings for CICIDS2017. Reducing the feature space resulted in small differences in accuracy and macro-F1 when compared to the complete feature set.

TABLE 7. In-domain performance of random forest with SHAP-GUIDED feature subsets (CICIDS2017).

Feature Set	Split	Accuracy	Macro-F1	Weighted-F1
RF Top-20	Validation	0.9949	0.8567	0.9951
RF Top-20	Internal Test	0.9949	0.9159	0.9951
COMMON-10	Validation	0.9948	0.8569	0.9951
COMMON-10	Internal Test	0.9949	0.9167	0.9951
UNION-30	Validation	0.9996	0.9033	0.9996
UNION-30	Internal Test	0.9996	0.9910	0.9996

Table 8 presents the cross-dataset performance results on CSE-CIC-IDS2018. Compared to the full 80-feature baseline (macro-F1 = 0.1273), the RF Top-20 configuration improved macro-F1 to 0.1663, while the COMMON-10 subset achieved 0.1653. Additional ablation experiments further showed that removing port-related features from the RF Top-20 subset increased macro-F1 to 0.2079, with accuracy rising to 0.8538. Confusion matrix analysis further showed that removing port-related features reduced dominant benign-class collapse and improved detection for certain attack categories such as DoS and Brute Force under cross-dataset evaluation. In contrast, the manually constructed Non-SHAP-20 subset achieved only limited performance improvement (macro-F1 = 0.1223), indicating that explainability-guided feature selection provided more effective feature prioritization under domain shift conditions. The UNION-30 configuration produced the lowest cross-dataset performance among the reduced SHAP subsets, suggesting that increasing feature count alone does not necessarily improve generalization.

TABLE 8. Cross-dataset performance of random forest with SHAP-GUIDED feature subsets (CSE-CIC-IDS2018).

Feature Set	Accuracy	Precision	Recall	Macro-F1	Weighted-F1
Full 80-Features	0.8310	0.2340	0.1333	0.1273	0.7554
RF Top-20	0.8364	0.3610	0.1574	0.1663	0.7660
RF Top-20-NoPort	0.8538	0.3555	0.2003	0.2079	0.7930
Non-SHAP-20	0.8303	0.1791	0.1295	0.1223	0.7559
COMMON-10	0.8362	0.3998	0.1568	0.1653	0.7656
UNION-30	0.8306	0.2094	0.1268	0.1170	0.7544

Confusion matrix analysis presented in Figure 2 further showed that removing port-related features reduced dominant benign-class collapse and improved detection for certain attack categories such as DoS and Brute Force under cross-dataset evaluation.

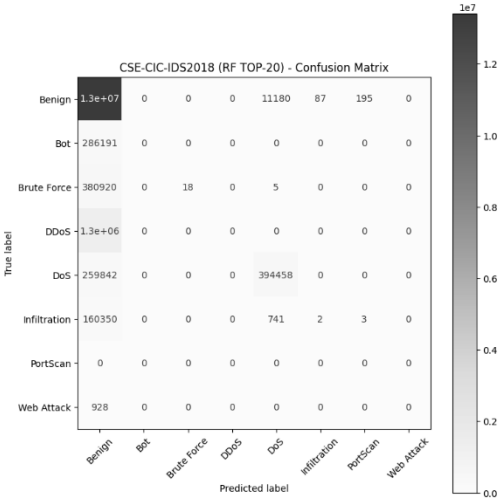

FIGURE 2. Confusion matrix for random forest on CSE-CIC-IDS2018 (RF TOP-20-NoPort).

D. PERFORMANCE WITH SHAP-GUIDED FEATURE REDUCTION (XGBoost)

This subsection evaluates the performance of XGBoost models trained on SHAP-guided feature subsets, including XGB Top-20, COMMON-10, and UNION-30.

Table 9 summarizes the in-domain performance findings for CICIDS2017. Across all feature combinations, XGBoost maintains high accuracy and weighted F1 scores.

TABLE 9. In-domain performance of XGBoost with SHAP-Guided feature subsets (CICIDS2017).

Feature Set	Split	Accuracy	Macro-F1	Weighted-F1
XGB Top-20	Validation	0.9997	0.9033	0.9997
XGB Top-20	Internal Test	0.9998	0.9937	0.9998
COMMON-10	Validation	0.9946	0.8528	0.9950
COMMON-10	Internal Test	0.9946	0.9163	0.9950
UNION-30	Validation	0.9997	0.9035	0.9997
UNION-30	Internal Test	0.9998	0.9937	0.9998

Table 10 presents the cross-dataset performance findings for CSE-CIC-IDS2018. For XGBoost, macro-F1 values remained within a narrow range between 0.1135 and 0.1192 across all feature configurations, indicating that SHAP-guided feature reduction did not substantially improve cross-dataset generalization for XGBoost under the evaluated conditions. Unlike Random Forest, XGBoost did not benefit from reduced SHAP-derived feature subsets under cross-dataset evaluation, suggesting that the impact of explainability-guided feature reduction may depend on the underlying model architecture.

E. COMPARATIVE SUMMARY OF CROSS-DATASET RESULTS

This subsection summarizes cross-dataset performance findings for all models and feature settings.

TABLE 10. Cross-dataset performance of XGBoost with SHAP-GUIDED feature subsets (CSE-CIC-IDS2018).

Feature Set	Accuracy	Macro-F1	Weighted-F1
Full 80-Features	0.8310	0.1192	0.7553
XGB Top-20	0.8306	0.1171	0.7544
COMMON-10	0.8301	0.1135	0.7531
UNION-30	0.8305	0.1165	0.7542

Table 11 compares accuracy, macro-F1, and weighted-F1 scores on CSE-CIC-IDS2018.

TABLE 11. Comparative cross-dataset performance summary on CSE-CIC-IDS2018.

Model	Feature Set	Accuracy	Macro-F1	Weighted-F1
Random Forest	Full (80)	0.8310	0.1273	0.7554
Random Forest	Top-20	0.8364	0.1663	0.7660
Random Forest	Top-20-NoPort	0.8538	0.2079	0.7930
Random Forest	Non-SHAP-20	0.8303	0.1223	0.7559
Random Forest	COMMON-10	0.8362	0.1653	0.7656
Random Forest	UNION-30	0.8306	0.1170	0.7544
XGBoost	Full (80)	0.8310	0.1192	0.7553
XGBoost	Top-20	0.8306	0.1171	0.7544
XGBoost	COMMON-10	0.8301	0.1135	0.7531
XGBoost	UNION-30	0.8305	0.1165	0.7542

Table 11 provides a comparative summary of cross-dataset performance across all models and feature configurations. Overall, SHAP-guided feature reduction resulted in measurable improvements for Random Forest, particularly under the RF Top-20-NoPort configuration, which achieved the highest macro-F1 score (0.2079) among all evaluated settings. In contrast, the manually constructed Non-SHAP-20 subset did not improve performance over the full feature baseline. XGBoost showed no consistent performance gains across reduced feature subsets, indicating that the impact of explainability-guided feature selection may depend on the underlying model architecture and feature interactions under domain shift conditions.

IV. DISCUSSION

This study investigated whether SHAP-guided feature selection can enhance cross-dataset generalization in flow-based intrusion detection systems trained on CICIDS2017 and deployed on CSE-CIC-IDS2018. The findings provide useful insights into domain shift, explainability-driven feature selection, and model-dependent generalization behavior.

A key observation is that near-perfect in-domain performance on CICIDS2017 is not equivalent to reliable cross-dataset detection. Despite reaching near-perfect accuracy and weighted F1-scores during in-domain evaluation, both Random Forest and XGBoost experience a significant drop in macro-F1 when applied to CSE-CIC-IDS2018. This demonstrates that high accuracy scores under domain shift are mostly due to benign traffic rather than true threat detection capacity. Similar findings have been reported in recent cross-dataset research, which show that models trained on a single curated dataset frequently fail to generalize new traffic scenarios and attack implementations [3], [17], [19], [26].

The confusion matrix analysis demonstrates that, during domain shift, models act as extremely confident benign detectors, misclassifying most of the attack flows as benign or mapping them to a minority group of attack types. This behavior is consistent with previous research indicating that flow-based intrusion detection models implicitly acquire dataset-specific traffic regularities rather than attack-invariant behavioral patterns [4], [11], [17]. Differences in background services, attack intensities, traffic composition, and temporal characteristics between CICIDS2017 and CSE-CIC-IDS2018 are anticipated to increase this effect [1], [2], [13].

SHAP-guided feature importance analysis provides additional insight into this occurrence. While SHAP successfully finds highly influential features within the source domain, many of these features are closely linked to static traffic characteristics like packet length distributions and port utilization. These traits are effective for discrimination within a single dataset but are susceptible to distributional shifts between datasets. This confirms previous explainable intrusion detection studies, which show that feature relevance does not always indicate feature robustness during domain change [4], [7], [12].

Additional ablation experiments further revealed that removing port-related features from the RF Top-20 subset substantially improved cross-dataset macro-F1 performance. Although port information appeared among the most influential SHAP-ranked attributes, these features may partially encode environment-specific service configurations rather than stable attack-related behavior. Their removal improved robustness under domain shift, suggesting that feature importance within a source dataset does not necessarily imply cross-dataset transferability.

This work makes an essential contribution by observing that SHAP feature ranks remain almost constant across sample sizes. Computing SHAP values on around 400 class-balanced flows resulted in almost identical Top-20 feature sets to those obtained from 10,000 flows. Minor rank changes happened within the Top-20 list, but the feature makeup remained intact. This finding shows that global SHAP-guided feature attribution may be done quickly without exhaustive sampling, considerably lowering the computing cost of large-scale intrusion detection systems. This builds on previous SHAP-guided IDS investigations by giving actual evidence for lightweight explainability without compromising ranking reliability [10], [12].

The effect of SHAP-guided feature reduction on cross-dataset generalization is highly model dependent. Compact feature subsets, such as RF Top-20 and COMMON-10, improve macro-F1 on CSE-CIC-IDS2018 in a moderate but

consistent way. This indicates that deleting dataset-specific or noisy characteristics allows Random Forest to focus on more reliable behavioral signals. Similar findings have been reported in explainable feature selection experiments, where smaller feature spaces increased robustness by minimizing overfitting to source-domain artifacts [4], [5], [21].

In contrast, the manually constructed Non-SHAP-20 subset failed to improve cross-dataset performance over the full feature baseline, indicating that explainability-guided feature prioritization provided more effective feature selection than manual heuristic reduction. However, XGBoost does not benefit from SHAP-guided feature reduction during cross-dataset testing. While decreased feature subsets maintain near-perfect in-domain performance, they do not increase macro-F1 with domain shift. This distinction can be due to XGBoost's reliance on fine-grained temporal and rate-based characteristics, which are extremely sensitive to changes in traffic generation and attack execution. Boosting-based models may thus overfit modest source-domain patterns even when trained on properly selected features, a drawback that has been identified in previous comparative IDS investigations [3], [6]. These findings suggest that explainability-guided feature reduction may interact differently with bagging and boosting-based learning mechanisms under domain shift.

While alternative explainability techniques such as LIME and Integrated Gradients are widely used in intrusion detection research, they are less suitable for explainability-driven feature selection in the context of tree-based, cross-dataset intrusion detection systems. LIME provides local, instance-level explanations that are highly sensitive to sampling variability and therefore do not yield stable global feature rankings required for constructing robust feature subsets across datasets [22]. Integrated Gradients, in contrast, is primarily designed for differentiable deep learning architectures and does not naturally extend to ensemble tree-based classifiers such as Random Forest and XGBoost without architectural adaptation [23]. In comparison, SHAP offers theoretically grounded and consistent global feature attributions that are directly compatible with tree-based ensemble models, making it more appropriate for stability-oriented explainability and feature analysis under domain shift [10], [12].

Overall, these findings show that explainability-driven feature selection enhances interpretability and model compactness but is insufficient to effectively resolve domain shift in flow-based intrusion detection. While SHAP can help uncover influential features and provide limited robustness advantages for some model families, generalization across datasets requires procedures other than feature selection. These procedures may include domain adaptation, drift-aware feature selection, or learning algorithms that prioritize attack-invariant behavioral patterns rather than dataset-specific statistics. Features that appear highly important within a source dataset may still capture unstable or environment-specific traffic characteristics that fail to generalize across deployment environments.

As a result, this research contributes to a better understanding of how explainable machine learning interacts with cross-dataset intrusion detection. It shows both the practical utility and the limitations of SHAP-guided feature selection, emphasizing that explainability is an effective diagnostic tool for establishing robust, real-world network intrusion detection.

This study also has some limitations. The experiments are based on two benchmark datasets, and the results may vary under different network environments or traffic conditions. In addition, domain adaptation techniques were not included, since the main goal was to focus on the effect of SHAP-guided feature selection under domain shift.

Future work can explore combining SHAP-guided feature selection with domain adaptation methods or drift-aware approaches. It would also be useful to investigate models that focus on more stable, attack-related patterns and attack-invariant feature representations to improve cross-dataset generalization.

V. CONCLUSION

This study examined the limitations of explainability-driven feature selection under cross-dataset domain shift in flow-based intrusion detection systems. Although Random Forest and XGBoost achieved near-perfect in-domain performance, both models exhibited substantial degradation in minority attack detection when deployed across datasets.

The findings demonstrate that SHAP provides stable and computationally efficient global feature attribution while also helping identify dataset-specific features that transfer poorly across environments. Additional ablation experiments showed that certain highly ranked port-related features may encode service-specific traffic behavior rather than attack-invariant characteristics. However, feature reduction alone does not resolve broader structural domain shift limitations. The observed model-dependent robustness gains indicate that explainability can improve interpretability, compactness, and feature prioritization, but cannot substitute for dedicated generalization mechanisms.

Overall, robust cross-dataset intrusion detection requires approaches beyond feature selection, including adaptation strategies and learning paradigms that emphasize attack-invariant behavioral representations.

REFERENCES

[1] Canadian Institute for Cybersecurity. *CICIDS2017 Dataset*. Accessed: Jun. 17, 2026. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html

[2] Canadian Institute for Cybersecurity. *CSE-CIC-IDS2018 Dataset*. Accessed: Jun. 17, 2026. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html

[3] M. Cantone, C. Marrocco, and A. Bria, "On the cross-dataset generalization of machine learning for network intrusion detection," 2024, *arXiv:2402.10974*.

[4] M. Sarhan, S. Layeghy, and M. Portmann, "Evaluating standard feature sets towards increased generalisability and explainability of ML-based network intrusion detection," 2021, *arXiv:2104.07183*.

[5] X. Chen, M. Liu, Z. Wang, and Y. Wang, "Explainable deep learning-based feature selection and intrusion detection method on the Internet of Things," *Sensors*, vol. 24, no. 16, p. 5223, 2024.

[6] M. Mahmoud, Y. O. Youssef, and A. Abdel-Hamid, "An explainable intelligent intrusion detection system," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2024, pp. 1–6.
[7] A. Samed and S. Sagiroglu, "Explainable artificial intelligence models in intrusion detection systems," *Eng. Appl. Artif. Intell.*, vol. 144, Mar. 2025, Art. no. 110145.
[8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[10] S. M. Lundberg and S.-I. Lee, "SHAP: A unified approach to interpreting model predictions," in *Proc. 31st Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.
[11] A. Thakkar and R. Lohiya, "A review of the advancement in intrusion detection datasets," *Proc. Comput. Sci.*, vol. 167, pp. 636–645, Apr. 2020.
[12] C. I. Nwakanma, L. A. C. Ahakonye, J. N. Njoku, J. C. Odirichukwu, S. A. Okolie, C. Uzondu, C. C. Ndubuisi Nweke, and D.-S. Kim, "Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review," *Appl. Sci.*, vol. 13, no. 3, p. 1252, Jan. 2023, doi: 10.3390/app13031252.
[13] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy*, Funchal, Portugal, 2018, pp. 108–116.
[14] M. Ring, S. Wunderlich, D. Grüdl, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1491–1530, Feb. 2019.
[15] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *J. Inf. Secur. Appl.*, vol. 50, Feb. 2020, Art. no. 102419.
[16] E. Hodo, X. Bellekens, A. Hamilton, P. L. Dubouilh, E. Iorkyase, C. Tachtatzis, and R. Atkinson, "Threat analysis of IoT networks using artificial neural network intrusion detection system," *Tsinghua Sci. Technol.*, vol. 23, no. 3, pp. 298–306, 2018.
[17] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE Symp. Secur. Privacy*, Oakland, CA, USA, May 2010, pp. 305–316.
[18] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–26, 2019.
[19] R. Jameel, K. Marwah, S. M. Idrees, and M. Nowostawski, "Evaluating the generalization gaps of intrusion detection systems across DoS attack variants," *J. Cybersecurity Privacy*, vol. 5, no. 4, p. 85, Oct. 2025.
[20] M. A. Ambusaidi, X. He, P. Nanda, and Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," *IEEE Trans. Comput.*, vol. 65, no. 10, pp. 2986–2998, Oct. 2016.
[21] C. E. L. Asry, I. Benchaji, S. Douzi, and B. E. L. Ouahidi, "Enhancing cybersecurity: A high-performance intrusion detection approach through boosting minority class recognition," *PLoS ONE*, vol. 20, no. 3, Mar. 2025, Art. no. e0317346.
[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144.
[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 3319–3328.
[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[25] S. Layeghy and M. Portmann, "Explainable cross-domain evaluation of ML-based network intrusion detection systems," *Comput. Electr. Eng.*, vol. 108, May 2023, Art. no. 108692, doi: 10.1016/j.compeleceng.2023.108692.
[26] A. Bhilwarawala, L. Rongmei, H. Sharma, A. Jena, K. Singh, J. Piri, and R. Dey, "BRIDGE and TCH-net: Heterogeneous benchmark and multi-branch baseline for cross-domain IoT botnet detection," 2026, *arXiv:2604.11324*.

CAN KILIÇ was born in Ankara, Türkiye, in 1999. He received the B.Sc. degree in computer engineering from Bilkent University, Ankara, and the M.Sc. degree in software engineering from Atılım University, Ankara.

His research interests include network intrusion detection systems, explainable artificial intelligence (XAI), cross-dataset generalization in machine learning, cybersecurity analytics, and SHAP-guided feature selection and robustness analysis in flow-based intrusion detection systems.

GÖKHAN ŞENGÜL was born in Ankara, Türkiye, in 1976. He received the B.S. degree from the Department of Electronic Engineering, Ankara University, Ankara, and the M.S. and Ph.D. degrees from the Department of Electrical and Electronics Engineering, Hacettepe University, Ankara, in 2002 and 2008, respectively. He is currently a Professor with the Department of Computer Engineering, Atılım University, Ankara. His current research interests include image processing, artificial intelligence, machine learning, and deep learning.