Cross-Dataset Generalization in Malware Traffic Classification: A Comparative Study with SHAP Explainability

Sharjeel Ibtisam
Faculty of Engineering and Technology
The Islamia University of Bahawalpur, Pakistan
sharjeel.ibtisam@gmail.com

Muhammad Uzair
Faculty of Engineering and Technology
The Islamia University of Bahawalpur, Pakistan
uzairch835@gmail.com

Aoun Muhammad
Faculty of Engineering and Technology
The Islamia University of Bahawalpur, Pakistan
aoun.muhammad@iub.edu.pk

Sana Tariq
Faculty of Engineering and Technology
The Islamia University of Bahawalpur, Pakistan
sana.tariq@iub.edu.pk

***Abstract*—Machine learning-based NIDS tend to perform well on the test sets included with their software, but don't work well in production environments. In this paper, the cross-dataset generalization in malware traffic classification is analyzed for the datasets of UNSW-NB15 and CTU-13. Five flow features were selected and used to train and test the classifiers namely Random Forest, Gradient Boosting and XGBoost classifiers as the features have semantic similarity to each other. The Kolmogorov-Smirnov (KS) test was employed to measure the domain shift between datasets and SHAP was used to measure feature importance and the stability of explanations between domains. The experimental results illustrate that both models have similar performance in the in-domain evaluation while there is a significant drop in performance for the out-of-domain evaluation on CTU-13. Gradient Boosting is the best performing algorithm over domain and KS analysis revealed that the distributions of all the features aligned were significantly different. Furthermore, SHAP analysis showed that the influence of features is very sensitive among different datasets indicating instability of learned representations. A clear indication of the importance of explicitly dealing with domain shift is seen in the results, which show the difficulty to deploy a machine learning intrusion detection system in an environment different from the environment it was trained in.**

***Index Terms*—Cross-dataset generalization, network intrusion detection, Random Forest, UNSW-NB15, CTU-13, SHAP explainability, feature stability**

I. Introduction

Today's network traffic is being increasingly encrypted by default due to privacy laws and security protocols (HTTPS, TLS 1.3, VPNs). While the encrypted information is safe, so is the ability to obscure malicious operations command-and-control communication, data exfiltration, and botnet coordination with the aid of conventional intrusion detection systems (IDS) that use deep packet inspection (DPI) or signature matching [1]–[3]. Machine learning have come out as a promising method for encrypted traffic analysis. Even though the traffic might be encrypted, the ML model can recognize it as either legitimate or malicious traffic. However, without any decryption, it is possible to identify the statistical and behavioral properties of a network flow, such as packet count, volume of bytes, and the time between the network packets (inter-arrival times) [4], [5]. The high accuracy (>95%) on benchmark datasets like CICIDS-2018 or UNSW-NB15 is reported in several studies [6]. However, there remains one big question that needs to be addressed: cross-dataset generalization. A well-working model that is successful in one data set will tend to fail in another data set that is collected independently and under different conditions and different time periods [7], [8]. This generalizability problem in different dissimilar network settings is one of the main challenges of practical deployment of intrusion detection systems based on ML [9]. But more important the most of ML-based IDS are black boxes. Security analysts are not in a position to comprehend why a model raises a flag on a flow as malicious [10] and this skews the confidence and hinders the deployment [11]. Explainable Artificial Intelligence (XAI) technologies like SHAP can explain predictions as a function of the input features, but no previous study has examined if the explanations are consistent across various network configurations [12].

This work makes the following main contributions:

1) We present a systematic cross-dataset evaluation of machine learning malware traffic classification methods by training them on UNSW-NB15 and testing them on CTU-13, showing the effect of domain shift for real world deployment.
2) We measure distribution differences at the feature level using Kolmogorov-Smirnov test and investigate the relationship between the feature stability and the performance of cross-domain models.
3) We study the consistency of feature importance across different datasets by analyzing the explainability of the

© 2026 TechShield Publications
DOI: https://doi.org/10.67535/tsp.000003.018
Published under the ICACNC 2026 Proceedings

learnt decisions with SHAP as a tool, to understand the consistency of learnt decision patterns.

4) Show that high in-domain performance does not necessarily result in high across-domain performance and highlight the importance of having domain-aware approaches for intrusion detection.

II. Related Work

Encrypted malware detection has been progressing at a fast rate, with methods such as unsupervised feature fusion, and graph-based transformers showing high in-domain accuracy [13], [14]. However, most of these studies do not take into account generalization of their results to other datasets that have been independent of each other [4]. Consistent generalization gap has been reported between datasets. When classifiers have been transferred between four public datasets, Cantone et al. (2024) showed that the accuracy was reduced to almost randomness [9]. In a recent study, Apruzzese et al. demonstrate that to facilitate performance in different network conditions, attack traffic must be available during training [15]. Explainability has been a focus for deployment of NIDS. SHAP has been used to analyse intrusion detection models to improve transparency [16]. Zeleke et al. (2025) used XGBoost and SHAP to the CTU-13; their model was found to be extremely accurate but did not hold true in any other data set [5]. In another study, Singh et al. (2025) leveraged SHAP for identifying anomalies, but without any cross-dataset evaluation [17]. To the best of our knowledge, there is no previous work that quantifies the stability of SHAP explanations when training a network on different versions of the same data set. In addition, no studies have directly correlated feature stability to cross-dataset performance in the field of malware traffic classification. Although most of the existing works address how to enhance model accuracy when using a single dataset, they tend to ignore how feature distributions in different network environments vary, resulting in a large degradation in performance when models are deployed on unseen data [4].

III. Methodology

A. Datasets

1) **UNSW-NB15:** The UNSW-NB15 dataset consists of about 2.5 million network flow records with 49 features extracted using the tools Argus and Bro. It has several attack categories including Fuzzers, DoS and Exploits. The official training split was followed, including binary labels (0 = normal, 1 = attack).
2) **CTU-13:** CTU-13 traffic over 13 scenarios + background normal traffic. It was collected in a different network environment, hence different statistical properties than the UNSW-NB15. Normal and attack data sets were loaded separately and combined.

B. Feature Alignment

Semantically similar features were aligned in order to facilitate cross-dataset evaluation between UNSW-NB15 and CTU-13. Column names were converted to lowercase and contents of columns were trimmed of spaces. Table I shows the semantic feature alignment between UNSW-NB15 and CTU-13. We kept five aligned flow features, as they were the only features that could be consistently mapped across UNSW-NB15 and CTU-13, allowing a controlled cross-dataset evaluation.Selected features were aligned with their defined descriptions from the different datasets. Descriptive statistical analysis was also undertaken to compare their distribution between datasets, as well as semantic matching. The aligned features are then analysed in order to quantify the degree of similarity and mismatch, using the following KS-test analysis.

TABLE I
Semantic Feature Alignment Between UNSW-NB15 and CTU-13

CTU-13 Feature	UNSW	Description
Flow Duration	dur	Flow duration
Total Fwd Packets	spkts	Number of forward packets
Total Bwd Packets	dpkts	Number of backward packets
Total Fwd Bytes	sbytes	Bytes sent forward
Total Bwd Bytes	dbytes	Bytes sent backward

C. Data Cleaning

The data sets were both preprocessed by discarding invalid values. The values at the limits of the range were set to NaN and values missing from a row were deleted. These sets were then re-indexed to make sure there was consistency.

D. Feature Engineering

Logarithmic transformation (log1p) was used on certain features such as flow duration and byte related features to reduce skewness and stabilize feature distributions. This is a compression of extreme values that are frequently found in network traffic.

E. Train/Validation Split

The UNSW-NB15 was split into a training set (80%) and a validation set (20%) by stratified sampling to ensure class balance. Only the CTU-13 dataset was used for test data in a cross domain setting.

F. Feature Scaling

The features were normalized using the RobustScaler which scales features using interquartile range and median. This method is extremely stable to outliers in the network traffic data. This scaler was only applied to the UNSW training set and applied to both the validation and the CTU-13 sets, so there is no information leakage from the target domain.

G. Kolmogorov-Smirnov Test

Using the two-sample Kolmogorov-Smirnov (KS) test, distribution difference between UNSW-NB15 and CTU-13 was quantified for each feature. The KS statistic is given as the largest difference between the cumulative distribution functions of the samples. The features that have smaller KS values are deemed to be relatively less variable between the data sets.

H. Model Training

In this work, three machine learning models were trained: Random Forest, Gradient Boosting and XGBoost. The hyperparameters for each model are summarized in Table. All experiments were performed using the same fixed random seed, to assure the reproducibility and consistency of the results.

TABLE II
HYPERPARAMETER SETTINGS OF THE EVALUATED MODELS

Model	Hyperparameters
Random Forest	`n_estimators=100, max_depth=None, max_features='sqrt'`
Gradient Boosting	`n_estimators=100, learning_rate=0.1, max_depth=3, loss='log_loss'`
XGBoost	`n_estimators=100, learning_rate=0.1, max_depth=6, gamma=0, reg_alpha=0, reg_lambda=1, subsample=1.0, colsample_bytree=1.0`

I. Evaluation Protocol

A strict evaluation protocol was adhered to:

- **In-domain evaluation:** Performance on UNSW validation set.
- **Cross-domain evaluation:** The results of a cross-domain evaluation are presented using the CTU-13 dataset.

This configuration is used to get a true picture of the ability to generalize.

J. Evaluation Metrics

Model performance was assessed by:

- Accuracy
- ROC-AUC
- Average Precision (Precision-Recall)

The model was evaluated on the basis of Accuracy, Precision, Recall, F1-score, ROC-AUC and Average Precision (AP).

K. Generalization Gap

A generalization gap was defined as:

$$\text{Gap} = \text{AUC}_{\text{validation}} - \text{AUC}_{\text{CTU}} \tag{1}$$

This measure shows the performance of models on data distributions that are not included in their training set.

L. Stable Feature Analysis

A KS statistic has been used to rank the features. The relatively stable features were chosen as the subset of smallest KS values. These features were used to train a separate model, which was used to assess the effect of these features.

M. Explainability (SHAP)

The analysis of feature contributions was performed using the software called SHAP (TreeExplainer). The data sets from both UNSW validation and CTU-13 were sampled to 500 samples. The base SHAP values for each domain were averaged to calculate mean absolute SHAP values for comparison.

N. Learning Curve Analysis

The model performance with an increasing training data size was analyzed through the learning curves, created by Stratified Cross-Validation.

O. Experimental Configuration

Table III presents the experimental configuration used in this study.

TABLE III
EXPERIMENTAL CONFIGURATION

Parameter	Value
Random Seed	42
Train/Validation Split	80/20 (Stratified)
Scaler	RobustScaler
KS Test Sample Size	Full dataset
SHAP Sample Size	500 samples per dataset
Cross-Validation	Stratified K-Fold
Evaluation Metrics	Accuracy, ROC-AUC, AP

IV. RESULTS

A. Domain Shift Analysis

The Kolmogorov-Smirnov (KS) test shows that there are significant differences in distribution between UNSW-NB15 and CTU-13 in all features ($p < 10^{-15}$). The results show that there is severe domain shift, especially for the features time and byte. The distributions of each of the aligned features are dramatically different between UNSW-NB15 and CTU-13 as shown in Fig. 1. The distribution of CTU-13 is more concentrated than that of the other, broader distribution, confirming statistical mismatch between datasets.

Statistical comparison was done between Gradient Boosting and XGBoost. The results of the McNemar's test for predictions from CTU-13 showed that there is a statistically significant difference between the two classifiers ($p < 0.001$), indicating that the difference in performance is not likely to be random. The Wilcoxon signed rank test of the 5-fold cross validation AUC scores did not show a significant difference ($p = 0.0625$). The results indicate that the models have a very different prediction behavior on the target dataset, but the in-domain cross validation performance is similar.

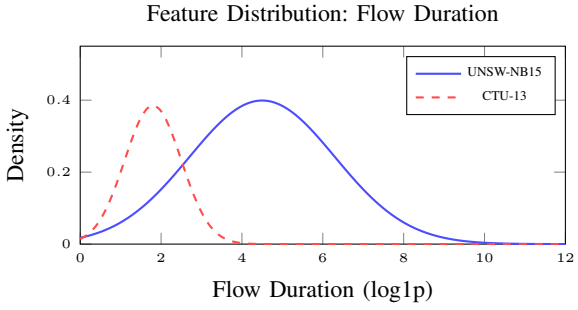

Fig. 1. Feature distribution comparison (UNSW vs CTU-13). CTU-13 is more concentrated while UNSW-NB15 has a broader distribution, confirming statistical mismatch.

Table IV shows the KS statistics for feature distribution shift.

TABLE IV
KS STATISTICS FOR FEATURE DISTRIBUTION SHIFT

Feature	KS Statistic
Flow Duration	0.965
Total Forward Bytes	0.804
Total Forward Packets	0.621
Total Backward Packets	0.437
Total Backward Bytes	0.257

B. In-Domain vs Cross-Domain Performance

There were two conditions evaluated:

- **In-domain:** UNSW validation set
- **Cross-domain:** CTU-13 dataset

All models deliver good in-domain performance (AUC ≈ 0.98), but performance decreases significantly at CTU-13. As one can see in Table V, the generalization gap is quite large in all models, with the largest gap being in the Random Forest model (0.4393) and the smallest gap being in the Gradient Boosting model (0.2978). This means that, even with the best-performing model, there is a large drop in performance when it is used on a new set, and that is the effect of a large domain shift. Additional cross-domain metrics reveal that XGBoost achieved highest precision (0.55), recall (0.58) and F1-score (0.54), followed by Gradient Boosting and Random Forest.

TABLE V
IN-DOMAIN AND CROSS-DOMAIN PERFORMANCE COMPARISON

Model	UNSW Acc.	CTU Acc.	AUC Gap
Random Forest	0.9330	0.5510	0.4393
Gradient Boosting	0.9325	0.6157	0.2978
XGBoost	0.9316	0.6768	0.3491

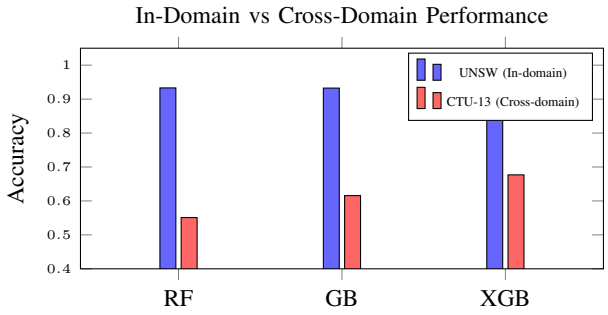

Fig. 2. In-domain vs cross-domain performance comparison. All models show significant accuracy drop on CTU-13.

C. ROC and Precision-Recall Analysis

Gradient Boosting achieves the highest AUC (0.6864), indicating better discriminative capability.

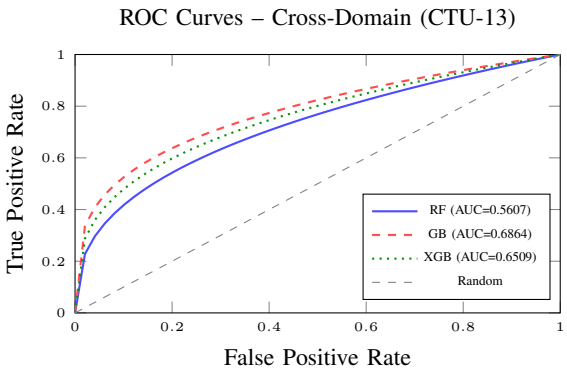

Fig. 3. ROC curves for cross-domain evaluation on CTU-13. Gradient Boosting achieves the highest AUC.

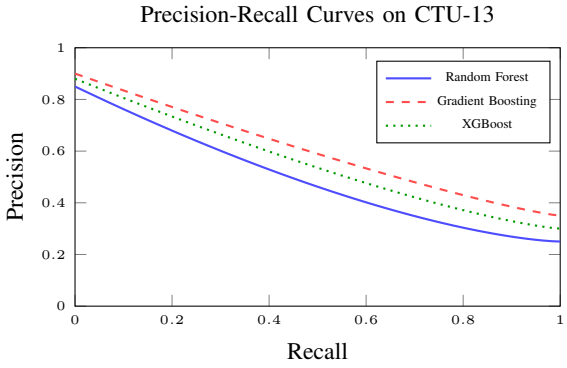

Fig. 4. Precision-Recall curves on CTU-13. Gradient Boosting maintains higher precision across recall levels.

D. Confusion Matrix Analysis

Key observations:

- Random Forest has low attack recall (0.25).
- Gradient Boosting has a balanced performance (recall ≈ 0.57).
- The high attack recall (0.80) comes with the drawback of introducing a large number of errors and more false positives.

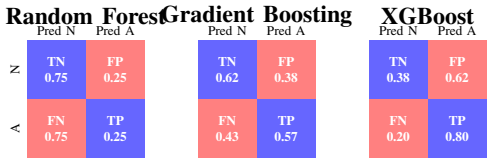

Fig. 5. Confusion matrices for all models on CTU-13. RF has low attack recall (0.25); XGBoost achieves high recall (0.80) at cost of more false positives.

E. Stable Feature Model

A smaller feature set (Total Forward Packets, Total Backward Packets) was used to train a feature model in order to test its ability to generalize better between datasets. The selection of these features was done because they are less sensitive to traffic volume variations compared to byte-based features,

and are often found in a heterogeneous network monitoring environments.

TABLE VI
STABLE FEATURE MODEL PERFORMANCE

Metric	UNSW Val	CTU Test
Accuracy	0.8192	0.3523
ROC-AUC	0.8824	0.2986

Even with less distribution shift, performance is still significantly lowered, showing that it is not enough to be stable for prediction.

F. Feature Importance

Features that are important are typically those that have high distribution mismatch, impacting generalization.

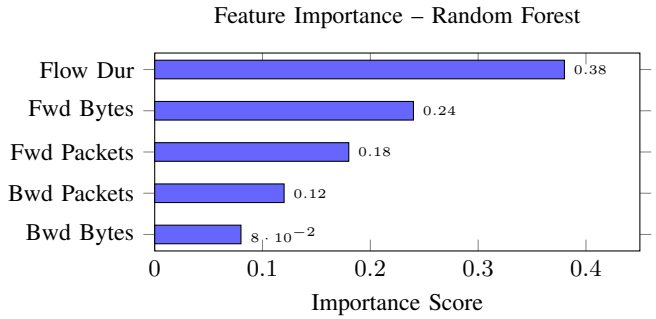

Fig. 6. Feature importance (Random Forest). Flow Duration and Forward Bytes dominate, and both have high KS statistics, confirming their role in generalization failure.

G. SHAP Analysis

SHAP results reveal that the features' contributions differ for different data sets, which indicates instability in learned representations.

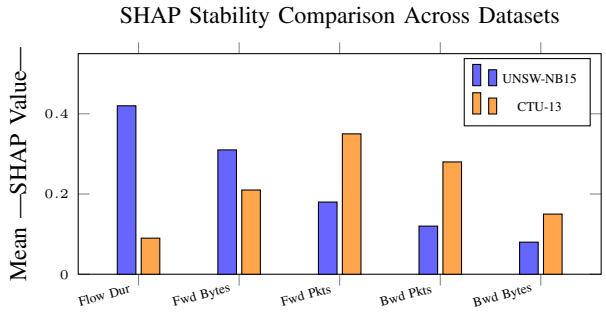

Fig. 7. SHAP stability comparison across datasets. Feature contributions differ substantially between UNSW-NB15 and CTU-13, indicating instability in learned representations.

only minimal explanation consistency across datasets was found by quantitative SHAP analysis utilising Spearman correlation, cosine similarity, and JSD. Both models showed significant shifts in feature contributions, demonstrating the influence of domain shift on model explanations, even if XGBoost outperformed Gradient Boosting in terms of cosine similarity.

TABLE VII
QUANTITATIVE SHAP STABILITY COMPARISON ACROSS DATASETS

Model	Spearman r	Cosine Similarity	Mean JSD
Gradient Boosting	0.7000	0.7586	0.4353
XGBoost	0.7000	0.8973	0.3890

H. Learning Curve

The learning curve shows that it has a good in-domain learning but not as much in the cross domain.

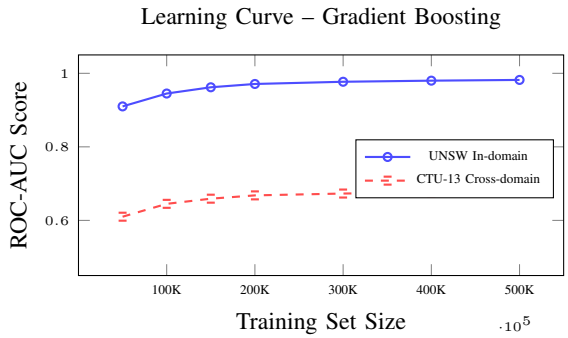

Fig. 8. Learning curve of best-performing model (Gradient Boosting). Good in-domain learning plateaus early while cross-domain performance remains substantially lower throughout.

V. THREATS TO VALIDITY

There are a number of limitations in this study. The first is the experimental evaluation was limited to two publicly available datasets (UNSW-NB15 and CTU-13). These databases are commonly used for intrusion detection studies, but conclusions drawn from the results might be applicable to different network environments or attacks. Second, only five flow features were used for cross-dataset evaluation due to semantic similarity between the datasets. This allowed for a fair comparison between the datasets, but other potentially informative features were not included because of differences in the datasets structure. Finally, SHAP analysis was performed for a subset of samples to perform explainability analysis. While SHAP offers interpretable explanations of the features, the stability of the explanations might differ across the various samples used and model settings.The study considered only one transfer direction (UNSW-NB15→CTU-13). Reverse evaluation (CTU-13→UNSW-NB15) was not performed and may reveal additional insights regarding asymmetry and domain-specific bias. The study excluded deep learning baselines including MLPs, CNNs, Transformers, and TabNet in favour of tree-based machine learning models. As a result, not all model families may be properly represented by the observed generalisation behaviour.

VI. DISCUSSION

This study reveals that the domain shift has a huge effect on the cross-dataset generalization in network intrusion detection

systems. All models performed well on the UNSW-NB15 validation set (AUC $\approx$ 0.98) but their performance on the test set dropped significantly when they were evaluated outside of the settings in which they were trained. The CTU-13 dataset was used for training. This validates models trained with one dataset are very sensitive to variations in data distribution.

The Kolmogorov-Smirnov analysis revealed significant variations in distribution of all features with the highest differences being observed for flow duration and features related to the bytes. These features also stand out as most significant when training the model, and is why performance suffers when taken to a new set of data. This is a direct cause of the observed generalization gap because of this feature mismatch.

The KS test confirms a significant distribution mismatch between UNSW-NB15 and CTU-13 but does not specify mechanisms to close this gap. How to do it Potential approaches : Feature alignment techniques , distribution matching methods based on CORAL , adversarial domain adaptation , learning invariant features . These methods are left for future investigation.

The generalization gap analysis has also been done for the model, which shows a degradation in performance for all models from 0.2978 to 0.4393 in AUC. In this case, the models evaluated are more robust than Random Forest and XGBoost as seen in the Gradient Boosting model with relatively better cross-domain performance.

The performance of the reduced feature model (using packet-count features) was significantly decreased. This means that, the distribution similarity is not enough for feature selection to assure predictive power. SHAP Analysis was used to gain further understanding of model behavior.

The distribution of feature importance values for each data set shows that there are different decision patterns between the different data sets. This inconsistency in feature contributions is another clue to the fall in cross-domain performance and validates that learned representations are not transferable over datasets. Statistical tests revealed significant differences among the models evaluated in the cross-domain predictions that emphasise the importance of model selection under the presence of domain shift.

In general, the results indicate that high accuracy in-domain does not necessarily translate to generalization in new domains. The findings highlight the importance of using more robust methods that directly combat the domain shift, such as the use of domain adaptation methods or learning of invariant features.

VII. Conclusion

The cross-dataset generalization was analyzed in this study with UNSW-NB15 and CTU-13. The models performed very well on in-domain tests (AUC $\approx$ 0.98), but not so well on CTU-13 (due to the domain shift). KS and SHAP analysis found that the data were different in terms of distribution and importance. The results demonstrate that high accuracy on one data set does not necessarily imply high accuracy on real data. In future, the aim is to develop more generalization with the help of domain adaptation techniques.

References

[1] R. Gao, H. Lu, and X. Du, "Active and passive attack detection methods for malicious encrypted traffic," in *2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2024.

[2] H. Guidry, "Limitations of signature-based network intrusion detection under modern traffic conditions," 2026, paper presented at the 2026 Spring Cybersecurity Undergraduate Research Showcase, Old Dominion University. [Online]. Available: https://digitalcommons.odu.edu/covacci-undergraduateresearch/2026spring/projects/4/.

[3] ipoque GmbH and The Fast Mode, "Deep packet inspection and encrypted traffic visibility for IP networks," 2023, research report based on a survey of 34 networking vendors. [Online]. Available: https://www.ipoque.com/news-media/resources/ebooks/dpi-encrypted-traffic-visbility.

[4] I. A. Alwhbi, C. C. Zou, and R. N. Alharbi, "Encrypted network traffic analysis and classification utilizing machine learning," *Sensors*, vol. 24, no. 11, p. 3509, 2024.

[5] S. N. Zeleke, A. F. Jember, and M. Bochicchio, "Integrating explainable AI for effective malware detection in encrypted network traffic," *arXiv preprint arXiv:2501.05387*, 2025, [Online]. Available: https://arxiv.org/abs/2501.05387.

[6] S. Saidane, F. Telch, K. Shahin, and F. Granelli, "Optimizing intrusion detection system performance through synergistic hyperparameter tuning and advanced data processing," *CoRR*, vol. abs/2408.01792, 2024, [Online]. Available: https://arxiv.org/abs/2408.01792.

[7] M. Bondarev and O. Guzev, "Capabilities and limitations of attack classification in encrypted traffic by machine learning methods," *International Journal of Open Information Technologies*, vol. 13, no. 7, 2025, [Online]. Available: http://injoit.ru/index.php/j1/article/view/2163.

[8] M. I. Amin, M. Shen, S. U. A. Laghari, M. Parthipan, and S. Karuppayah, "Unveiling the generalizability gap: A cross-domain evaluation of machine learning algorithms for network intrusion detection," in *2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2024, pp. 1–7.

[9] M. Cantone, C. Marrocco, and A. Bria, "Machine learning in network intrusion detection: A cross-dataset generalization study," *IEEE Access*, vol. 12, pp. 144 489–144 508, 2024.

[10] D. Simbana *et al.*, "Explainable AI for forensic analysis: A comparative study of SHAP and LIME in intrusion detection models," *Applied Sciences*, vol. 15, no. 13, p. 7329, 2025.

[11] K. S. Alketbi and A. Mehmood, "A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection," *IEEE Access*, vol. 13, pp. 121 772–121 798, 2025.

[12] C. C. Käsbohrer, S. Mair, and L. Jiang, "Assessing explanation fragility of SHAP using counterfactuals," in *Proceedings of the 7th Northern Lights Deep Learning Conference (NLDL)*, ser. Proceedings of Machine Learning Research, vol. 307. PMLR, 2026, pp. 211–234.

[13] X. Zang, Z. Zheng, H. Zheng, X. Liu, M. K. Khan, and W. Jiang, "HyperEye: A lightweight features fusion model for unknown encrypted malware traffic detection," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 5079–5089, 2025.

[14] W. Lin, C. Xia, T. Wang, M. Liu, and Y. Li, "Enhancing encrypted traffic analysis via source APIs: A robust approach for malicious traffic detection," *Computers and Security*, vol. 156, p. 104529, 2025.

[15] G. Apruzzese, L. Pajola, and M. Conti, "The cross-evaluation of machine learning-based network intrusion detection systems," *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 5152–5169, 2022.

[16] P. Hermosilla, S. Berrios, and H. Allende-Cid, "Explainable AI for forensic analysis: A comparative study of SHAP and LIME in intrusion detection models," *Applied Sciences*, vol. 15, no. 13, p. 7329, 2025, [Online]. Available: https://www.mdpi.com/2076-3417/15/13/7329.

[17] K. Singh, A. Kashyap, and A. K. Cherukuri, "Interpretable anomaly detection in encrypted traffic using SHAP with machine learning models," 2025, [Online]. Available: https://arxiv.org/abs/2505.16261.