

Explaining Cross-Dataset False Positives in Encrypted Traffic Classification for Forensic Interpretation

Madhab Chandra Das

mxd147@shsu.edu

Sam Houston State University <https://orcid.org/0000-0003-2724-4315>

Qingzhong Liu

Sam Houston State University <https://orcid.org/0000-0002-2006-5413>

Research Article

Keywords: Encrypted Traffic Classification, Network Forensics, Explainable Machine Learning, False Positive Analysis, Cross-Dataset Evaluation

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11423701/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Explaining Cross-Dataset False Positives in Encrypted Traffic Classification for Forensic Interpretation

Madhab Chandra Das
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA
mxd147@shsu.edu

Qingzhong Liu
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA
qingzhong.liu@shsu.edu

Abstract—Machine learning models are increasingly used for encrypted traffic analysis because payload inspection is limited. However, high internal detection accuracy does not necessarily indicate that model outputs are reliable for forensic interpretation. This study examines this issue through cross-dataset false positive analysis. Four machine learning models are trained and evaluated internally on the CIRA-CIC-DoHBrw-2020 dataset. Random Forest achieved the highest internal accuracy of 99.62%, while CatBoost achieved 98.92% and was used for detailed cross-dataset and explainable analysis. The trained CatBoost model was then applied without retraining to 947,997 benign flows from the CESNET-TLS-Year22 dataset, where 41,238 flows were classified as malicious, resulting in a false positive rate of 4.35%. This result shows that strong internal performance can still produce misleading alerts when the traffic environment changes. SHAP and permutation importance are used to examine feature dependence in the model predictions. Both methods identify a consistent set of aligned flow-level features, with packet size mean showing the strongest influence in the SHAP analysis. These features support classification but are not unique to malicious traffic and can also occur in benign encrypted communication. Therefore, feature importance alone is insufficient for forensic interpretation. The findings indicate that machine learning alerts should be treated as investigative leads rather than direct forensic evidence, and false positive analysis should be considered when interpreting model explanations across different traffic environments.

Index Terms—Encrypted Traffic Classification, Network Forensics, Explainable Machine Learning, False Positive Analysis, Cross-Dataset Evaluation

I. INTRODUCTION

Today, network communication is largely protected through encryption, which limits the visibility of traditional inspection-based security mechanisms. As a result, machine learning models that rely on flow-level statistical features are widely used to detect malicious activity without accessing packet content [1], [2], [3]. These models often report very high classification accuracy in controlled environments.

However, such performance does not necessarily translate to real-world conditions. Network traffic varies across environments due to differences in applications, user behavior, and network characteristics. A model trained on one dataset may not behave consistently when applied to traffic from a different

source. This raises a critical concern regarding the reliability of machine learning-based encrypted traffic analysis, particularly in operational and forensic contexts [4].

One of the most important challenges in this setting is the presence of false positives, where benign traffic is incorrectly classified as malicious. Even a low false positive rate can produce a large number of incorrect alerts in high-volume network environments. In digital forensic investigations, such errors can mislead analysts, increase investigation workload, and lead to incorrect conclusions if model outputs are interpreted without sufficient validation [3].

Forensic interpretation requires more than a correct class label. Feature importance can show which variables influenced a prediction, but it does not prove that the prediction is reliable under external traffic conditions. This makes false positive analysis necessary for checking whether model explanations remain meaningful when traffic comes from a different environment.

A key limitation in existing research is the reliance on single-dataset evaluation. Many studies train and test models on data with similar characteristics, which can produce optimistic results. This limits understanding of how models behave on external traffic, especially when benign flows are incorrectly classified as malicious [5].

Explainable machine learning provides a way to investigate this problem by offering insight into model decision processes. While prior studies have applied explainable methods such as SHAP to interpret model predictions in encrypted traffic classification [6], [7], these works primarily focus on overall model behavior within a single dataset. They do not examine how explainability can be used to analyze false positive errors across different datasets and traffic environments.

The objective of this study is to examine whether strong machine learning detection performance can support forensic interpretation when the model is tested on external encrypted traffic. The study uses false positive analysis to evaluate whether model explanations remain meaningful under cross-dataset conditions. The focus is not only on which features influence predictions, but also on whether those features

provide useful support for forensic review.

Three research questions guide this study. First, how does false positive behavior change when a model is applied across different encrypted traffic datasets? Second, which features contribute most to false positive predictions in encrypted traffic classification? Third, how can explainable methods support the interpretation of false positives in a forensic context?

To address these questions, four models are trained and evaluated internally on the CIRA-CIC-DoHBrw-2020 dataset, while CatBoost is used for external evaluation on CESNET-TLS-Year22.

This study makes three key contributions. First, it provides a cross-dataset evaluation to quantify false positive behavior across different traffic environments. Second, it applies explainable analysis, including SHAP and permutation importance, to identify the feature-level causes of misclassification. Third, it examines the relationship between false positive behavior and model explanation to assess whether machine learning outputs provide reliable support for forensic interpretation. The findings show that feature importance alone is insufficient for forensic review when model behavior changes across datasets.

II. RELATED WORK

Machine learning has been widely used for encrypted traffic classification because encrypted communication limits payload inspection. These methods usually rely on flow-level statistical features, such as packet size, flow duration, byte count, and timing behavior [3], [8], [9]. Prior studies report strong classification performance using machine learning, deep learning, and hybrid models [7], [5]. However, many studies evaluate performance within a single dataset. This setting can show high detection accuracy, but it does not fully explain how the same model behaves when traffic comes from a different environment.

Cross-dataset generalization is a major concern in encrypted traffic analysis. Traffic characteristics can change across protocols, applications, users, and network settings [10], [11]. A model trained on one dataset may learn patterns that are specific to that dataset. When the model is applied to a new dataset, these learned patterns may not separate benign and malicious traffic in the same way. Existing work often reports performance degradation using accuracy, recall, or other detection metrics [12], [13]. Less attention is given to false positives, although false positives are important when benign traffic is incorrectly classified as malicious.

Explainable machine learning has been used to improve transparency in network traffic classification. SHAP, LIME, and related methods can show which features influence model predictions [14], [15], [16]. Recent studies also apply explainable methods to encrypted traffic analysis and intrusion detection [17], [6], [18]. These studies show that feature-level interpretation can help explain model behavior. However, explainability is often used to validate the model after training. It is less often used to examine whether false positive alerts remain meaningful under cross-dataset conditions.

False positives are a practical problem in network security and forensic analysis. When benign traffic is classified as malicious, analysts may spend time reviewing incorrect alerts. In forensic settings, this can also affect how model outputs are interpreted. Prior work has considered false positive reduction through feature selection, model design, or optimization [16], [19]. However, false positives are often treated as an error rate to reduce, not as a behavior that needs interpretation. This is important for encrypted traffic because benign and malicious flows may share similar statistical patterns.

This study addresses this gap by combining cross-dataset evaluation with explainable analysis. The purpose is not only to report model performance, but also to examine whether high detection performance provides meaningful support for forensic interpretation. The study uses false positive analysis to evaluate whether important features reflect malicious behavior or statistical similarity with benign encrypted traffic. In this way, the work connects model performance, feature explanation, and forensic interpretation within the same experimental setting.

III. METHODOLOGY

A. Framework Overview

This study uses a structured workflow to examine whether machine learning results can support forensic interpretation of encrypted traffic. The workflow includes dataset preparation, feature engineering, label mapping, preprocessing, model training, internal testing, cross-dataset evaluation, and explainable analysis. Fig. 1 shows the overall framework. The purpose of the workflow is not only to classify encrypted traffic, but also to examine whether model explanations remain meaningful when the model is applied to traffic from a different dataset.

B. Dataset Description

This study uses two encrypted traffic datasets: CIRA-CIC-DoHBrw-2020 and CESNET-TLS-Year22. The CIRA dataset is used for training, validation, and internal testing and contains 917,300 benign flows and 249,969 malicious flows. CESNET-TLS-Year22 is used only for external false positive evaluation and contains 947,997 benign TLS flows. No malicious CESNET samples are used in the final experiment because the external evaluation focuses on identifying benign traffic that is incorrectly classified as malicious.

The two datasets were collected from different traffic environments and were processed separately before feature alignment. The aligned data are then used for model development and cross-dataset evaluation, as summarized in Table I. Because the datasets differ in protocol and network conditions, the external evaluation is treated as a cross-environment test rather than a protocol-matched comparison. Therefore, the reported false positive rate reflects how sensitive the trained model is to changes in the external traffic environment rather than protocol-specific generalization.

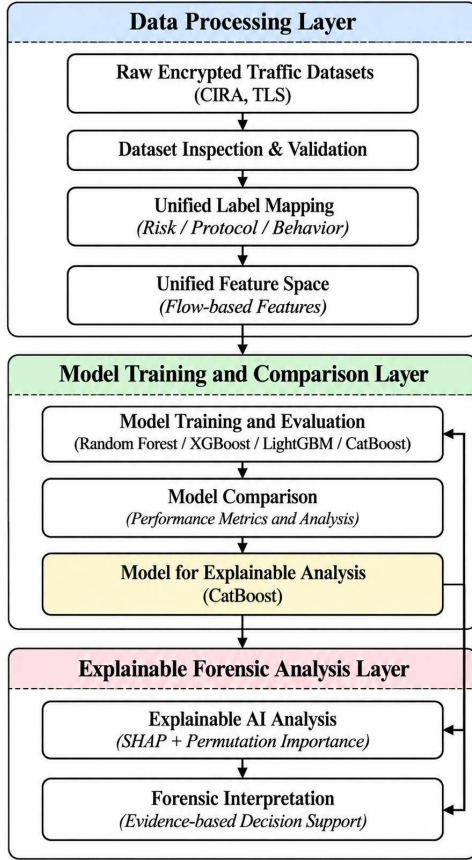


Fig. 1. Framework for encrypted traffic processing, model comparison, and explainable forensic analysis.

TABLE I
DATASETS USED IN THIS STUDY

Dataset	Role	Benign	Malicious
CIRA-CIC-DoHBrw-2020	Training and internal testing	917,300	249,969
CESNET-TLS-Year22	External false positive evaluation	947,997	0

C. Feature Engineering

Encrypted traffic limits payload inspection. Therefore, this study uses flow-level statistical features that describe traffic behavior without using packet content. After feature engineering across the datasets, five aligned features were retained for the experiment: flow duration, byte count, packet count, packet size mean, and inter-arrival time mean. These features were retained because they were consistently available after alignment across both datasets and therefore provided a common feature space for cross-dataset evaluation.

The selected features describe four traffic properties. Flow duration represents connection length. Byte count represents traffic volume. Packet count and packet size mean represent packet-level structure. Inter-arrival time mean represents timing behavior. These features also support forensic interpre-

TABLE II
SELECTED FEATURES AND DESCRIPTIONS

Feature	Description
Flow Duration	Total duration of a flow
Byte Count	Total bytes transferred
Packet Count	Number of packets in a flow
Packet Size Mean	Average packet size in a flow
Inter-Arrival Time Mean	Average time between packets

TABLE III
CIRA-CIC-DoHBrw-2020 INTERNAL DATASET SPLIT DISTRIBUTION

Split	Flows	Percentage
Training	817,071	70.0%
Validation	175,544	15.0%
Test	174,654	15.0%
Total	1,167,269	100.0%

tation because each feature has a direct meaning and can be reviewed without relying on encrypted payload content. Table II describes the selected features.

D. Label Mapping

The datasets use different label structures. To support a consistent binary classification task, all samples are mapped into two classes: benign and malicious. In the CIRA dataset, normal browsing traffic is mapped to the benign class. Tunneling and attack-related traffic are mapped to the malicious class. In the CESNET dataset, all selected flows are treated as benign for external false positive evaluation.

This mapping supports the main objective of the study. The purpose of the CESNET evaluation is not to measure malware recall. Instead, it is used to measure how often benign external encrypted traffic is incorrectly classified as malicious. This design allows the study to evaluate false positive behavior under cross-dataset conditions.

E. Data Preprocessing

Two evaluation settings are used in this study. First, the CIRA-CIC-DoHBrw-2020 dataset is divided into training, validation, and internal test sets, as shown in Table III. These splits are used only for model development and internal evaluation. The CESNET-TLS-Year22 dataset is kept completely separate from the internal splits.

Second, a cross-dataset false positive evaluation is performed by applying the model trained on CIRA-CIC-DoHBrw-2020 to benign CESNET-TLS-Year22 flows without retraining. This setting measures how often benign external traffic is incorrectly classified as malicious when the traffic source changes. The external CESNET data are not used during model training, validation, or internal testing.

F. Model Training

Four machine learning models are trained in this study: Random Forest, XGBoost, LightGBM, and CatBoost. These models are trained using the same aligned feature set. The

model comparison is used to measure internal detection performance and to show how strong internal performance may differ from external false positive behavior. The training data are imbalanced because benign flows are more frequent than malicious flows. To reduce bias toward the majority class, model-level class weighting is applied during training where it is supported by the classifier.

Accuracy measures the overall proportion of correct predictions, while the F1-score provides a balanced measure of precision and recall. Accuracy and F1-score are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$F1 = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}} \quad (2)$$

where TP , TN , FP , and FN represent true positives, true negatives, false positives, and false negatives, respectively.

CatBoost is used for the detailed cross-dataset and explainable analysis because it provides strong internal performance and supports direct SHAP-based interpretation for tree models. SHAP and permutation importance are used to examine the feature patterns associated with CatBoost predictions under the external evaluation setting.

G. Explainable Analysis

SHAP is applied to the CatBoost model to examine feature contributions [14]. Permutation importance is also used to evaluate global feature dependence. The global SHAP analysis identifies which features influence predictions across samples. This supports the main goal of the study by showing whether the features behind false positive alerts have a clear forensic meaning.

Permutation importance is also used to validate feature importance at the global level [16]. This method measures how model performance changes when a feature is randomly shuffled. A larger decrease in performance indicates that the model depends more strongly on that feature. SHAP and permutation importance are applied to CatBoost to examine whether both methods identify a consistent set of influential traffic features.

H. Cross-Dataset False Positive Evaluation

The cross-dataset evaluation is designed to measure false positive behavior on external benign traffic. The model trained on the CIRA dataset is applied to the CESNET-TLS-Year22 dataset without retraining. Since the selected CESNET data contain benign flows only, any malicious prediction is counted as a false positive. Fig. 4 illustrates the cross-dataset evaluation setup used for external false positive analysis.

The false positive rate is calculated as:

$$FPR = \frac{FP}{FP + TN} \quad (3)$$

An approximate 95% confidence interval for the false positive rate was calculated using the normal approximation to the binomial proportion.

TABLE IV
EVALUATION DESIGN

Phase	Dataset	Purpose	Evaluation
Training	CIRA train	Learn traffic patterns	Supervised learning
Validation	CIRA validation	Select model	Model selection
Internal testing	CIRA test	Measure detection performance	Standard metrics
External testing	CESNET TLS	Measure false positives	Cross-dataset test

This evaluation is central to the study because it tests whether a model with strong internal performance can still produce misleading alerts under external traffic conditions. A higher false positive rate indicates that the model may be sensitive to feature patterns that appear in both malicious training samples and benign external traffic. Table IV summarizes the evaluation design.

Algorithm 1 summarizes the workflow used for cross-dataset false positive analysis and explainable interpretation.

Algorithm 1 Cross-dataset false positive analysis

Require: CIRA-CIC-DoHBrw-2020 dataset, CESNET-TLS-Year22 dataset

Ensure: False positive rate and feature-level explanation

- 1: Extract the aligned flow-level features from both datasets
- 2: Map CIRA labels into benign and malicious classes
- 3: Treat selected CESNET flows as benign external traffic
- 4: Train the classifier using the CIRA training data
- 5: Apply the trained classifier to the CESNET dataset
- 6: Count each malicious prediction on CESNET as a false positive
- 7: Compute the false positive rate using $FPR = FP/(FP + TN)$
- 8: Apply SHAP to explain feature contributions for predictions
- 9: Compare SHAP findings with permutation importance
- 10: Interpret false positive alerts as decision support for forensic review

I. Forensic Interpretation Process

The final stage connects model output with forensic interpretation. The process begins with a model alert. If benign external traffic is classified as malicious, the case is treated as a false positive for analysis. SHAP explanations are then used to identify which features contributed to the malicious prediction. These features are reviewed to determine whether they provide meaningful support for forensic interpretation or whether they reflect statistical similarity with benign encrypted traffic.

This process treats machine learning output as decision support rather than direct forensic evidence. It allows the investigator to review the feature-level basis of an alert before drawing any conclusion about malicious activity.

IV. RESULTS

A. Internal Model Performance

The internal evaluation was performed on the CIRA-CIC-DoHBrw-2020 test split. This setting measured model performance when training and testing data came from the

TABLE V
INTERNAL MODEL PERFORMANCE ON THE CIRA TEST DATASET

Model	Acc.	AUC	Prec.	Recall	F1
Random Forest	0.9962	0.9995	0.9929	0.9895	0.9912
XGBoost	0.9935	0.9996	0.9773	0.9930	0.9851
LightGBM	0.9902	0.9993	0.9649	0.9903	0.9774
CatBoost	0.9892	0.9989	0.9621	0.9886	0.9752

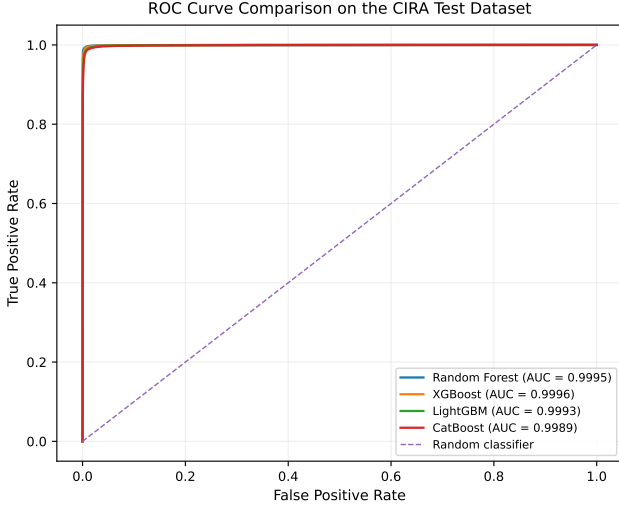


Fig. 2. ROC curve comparison on the CIRA test dataset.

same aligned feature space. Table V summarizes the internal performance of the baseline models.

As shown in Table V, all four models achieved strong internal performance. Random Forest achieved the highest accuracy and precision, while XGBoost achieved the highest AUC and recall. Fig. 2 compares the ROC curves of all four models, while Fig. 3 shows the Random Forest confusion matrix. These results indicate strong internal detection performance, but they do not establish whether the same performance remains reliable when the traffic source changes.

B. Cross-Dataset False Positive Evaluation

The cross-dataset evaluation tested how the trained model behaved when applied to external encrypted traffic. The CatBoost model trained on CIRA-CIC-DoHBrw-2020 was applied to the CESNET-TLS-Year22 dataset without retraining. The selected CESNET data contained benign flows only. Therefore, any malicious prediction on this dataset was counted as a false positive.

Table VI shows the external evaluation result. The model classified 41,238 out of 947,997 benign CESNET flows as malicious. This produced a false positive rate of 4.35%, with an approximate 95% confidence interval of 4.31% to 4.39%. This result shows that strong internal detection performance did not remove false positive behavior under cross-dataset testing. It also shows why internal accuracy alone is not sufficient for forensic interpretation.

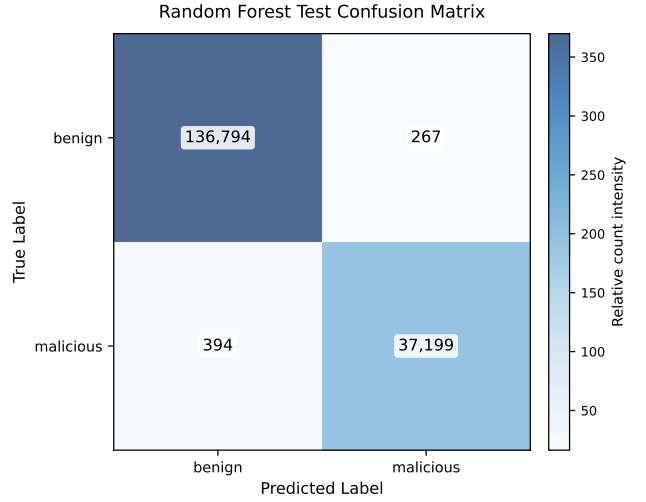


Fig. 3. Confusion matrix of the Random Forest model on the CIRA test dataset.

TABLE VI
CROSS-DATASET EVALUATION ON THE CESNET TLS DATASET

Metric	Value
Total benign flows	947,997
False positives	41,238
False positive rate	4.35%
95% confidence interval	4.31–4.39%

C. SHAP-Based Feature Analysis

Fig. 5 presents the SHAP summary plot. The analysis shows that *packet_size_mean* has the strongest influence on model predictions, while *packet_count*, *inter_arrival_time_mean*, *byte_count*, and *flow_duration* also contribute to the model output.

The SHAP result shows that the model relied mainly on packet structure, flow length, traffic volume, and timing behavior. These features are useful for classification, but they are not unique to malicious traffic. Benign encrypted traffic can also contain large packets, long flows, high byte counts, and regular timing patterns. This helps explain why some benign CESNET flows were classified as malicious.

D. Permutation Importance Analysis

Permutation importance was used to check whether the feature dependence observed in SHAP was also reflected in model performance. Fig. 6 shows the permutation importance result. The result indicates that *flow_duration*, *byte_count*, *inter_arrival_time_mean*, *packet_size_mean*, and *packet_count* contribute to model performance.

The ranking differs from SHAP because the two methods measure different aspects of model behavior. SHAP measures feature contribution to prediction output. Permutation importance measures the effect of feature disruption on model performance. Even with this difference, both methods identify the same set of traffic features as important. This supports the

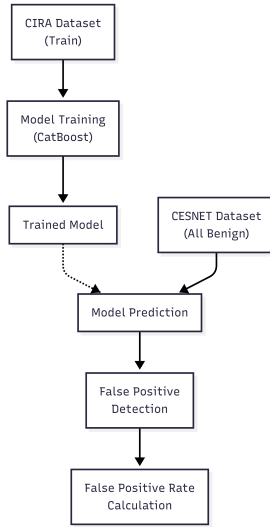


Fig. 4. Cross-dataset evaluation setup for false positive analysis on CESNET TLS traffic.

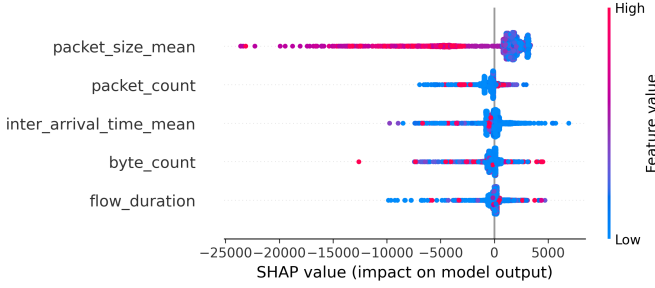


Fig. 5. SHAP summary plot for the CatBoost model. Each point represents one sample. The horizontal position shows the effect on the model output. The color shows the feature value.

finding that the model depends on a small group of aligned flow-level features.

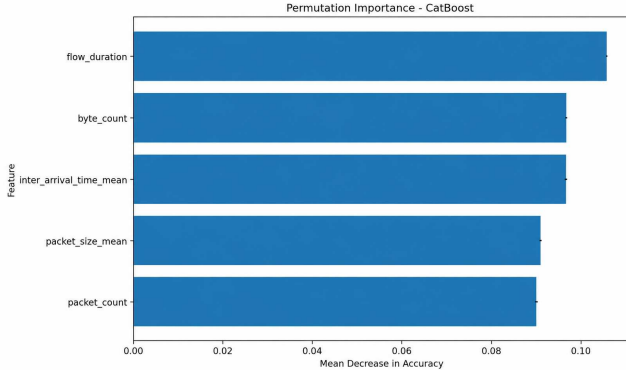


Fig. 6. Permutation importance results for the CatBoost model.

E. Result Summary

The results show a clear difference between internal detection performance and external false positive behavior. The

internal test results show high model performance on the CIRA test data. The external CESNET evaluation shows that benign encrypted traffic can still be classified as malicious when the traffic source changes. SHAP and permutation importance show that the model relies on flow duration, packet size, byte count, packet count, and inter-arrival time. These features support prediction, but they can also appear in benign encrypted communication. Therefore, feature importance must be interpreted together with false positive analysis before model alerts are used in forensic review.

V. DISCUSSION

A. Cross-Dataset Reliability

The results show a clear difference between internal detection performance and external false positive behavior. The models performed well on the internal CIRA test data, but the CatBoost model produced a 4.35% false positive rate on benign CESNET-TLS-Year22 traffic. This shows that strong internal performance does not automatically provide reliable support for forensic interpretation when the traffic source changes.

This issue is important because encrypted traffic behavior can vary across networks, applications, and user activity. A pattern learned as malicious in one dataset may also appear in benign traffic from another dataset. Cross-dataset false positive analysis therefore gives a more direct view of whether model alerts remain useful for forensic review.

B. False Positive Causes

The explainable analysis shows that the model relies on a small group of aligned flow-level features. SHAP identifies *packet_size_mean* as the strongest contributor, while the remaining flow-level features also influence model predictions. Permutation importance identifies the same feature set, although the ranking order differs.

These features are useful for classification, but they are not unique to malicious traffic. Benign encrypted communication can also contain long flows, large byte counts, and structured packet sizes. This overlap helps explain why some benign CESNET flows were classified as malicious. The result shows that feature importance must be interpreted carefully because an important predictive feature is not necessarily a reliable forensic indicator of malicious activity.

C. Forensic Interpretation

The main forensic concern is not only whether the model is accurate but also whether its alerts can be interpreted safely. In this study, the external false positive result shows that a malicious prediction can be produced for benign encrypted traffic. If such an alert is treated as direct evidence, it may lead to incorrect interpretation of normal activity.

SHAP helps reduce this risk by showing which features influenced the model decision. However, SHAP does not prove that an alert is malicious. It only explains the model's reasoning. The investigator still needs to decide whether the contributing features provide meaningful support in the case

context. Therefore, model output should be treated as decision support. It should guide further review, but it should not replace forensic validation. At an FPR of 4.35%, approximately 4,350 benign flows per 100,000 could be flagged as malicious under the evaluated setting. At larger traffic volumes, this can create a substantial review workload for investigators. Therefore, model alerts should be used to prioritize further examination rather than being interpreted as direct evidence of malicious activity.

D. Limitations

This study has several limitations. The analysis uses five aligned flow-level features and one external benign dataset. In addition, CIRA-CIC-DoHBrw-2020 and CESNET-TLS-Year22 are not protocol-matched, so the results represent cross-environment sensitivity rather than protocol-specific generalization. The classification task is binary, and the detailed explanation stage focuses on CatBoost. Computational cost was not evaluated in this study and should be examined in future deployment-oriented work.

E. Future Work

Future work should examine richer feature sets, including bidirectional flow statistics, burst characteristics, and protocol-level contextual features where consistently available across datasets. This would help test whether the observed feature overlap appears across more environments. These features may help distinguish benign high-volume encrypted traffic from malicious traffic more clearly. Another direction is to compare explanation behavior across more model families. This would show whether false positive explanations remain stable across different classifiers.

VI. CONCLUSION

This study examined the reliability of machine learning-based encrypted traffic classification under cross-dataset conditions. Random Forest achieved 99.62% accuracy on the internal CIRA-CIC-DoHBrw-2020 test set, while CatBoost produced 41,238 false positives on 947,997 benign CESNET-TLS-Year22 flows, corresponding to a 4.35% false positive rate. SHAP and permutation importance showed that the model mainly relied on flow-level features that can also occur in benign encrypted traffic.

These results show that strong internal performance and feature importance alone are not sufficient for forensic interpretation. Machine learning alerts should therefore be treated as decision support rather than direct evidence. Future work should evaluate additional external datasets and richer contextual features.

REFERENCES

- [1] M. Shen, K. Ye, X. Liu, L. Zhu, J. Kang, S. Yu, Q. Li, and K. Xu, "Machine learning-powered encrypted network traffic analysis: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 791–824, Jan. 2023. [Online]. Available: <https://doi.org/10.1109/COMST.2022.3208196>
- [2] 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. [Online]. Available: <https://www.mdpi.com/1424-8220/24/11/3509>
- [3] J. Zhang and A. Goscinski, "Flow-based traffic retrieval using statistical features," in *2016 International Workshop on Big Data and Information Security (IWBIS)*, 2016, pp. 25–30.
- [4] S. Ibrahim, C. Catal, and T. Kacem, "The use of multi-task learning in cybersecurity applications: a systematic literature review," *Neural Computing and Applications*, vol. 36, no. 35, pp. 22 053–22 079, 2024. [Online]. Available: <https://doi.org/10.1007/s00521-024-10436-3>
- [5] Z. Wang, Y. Yang, and Y. Wang, "A survey of encrypted traffic classification: Datasets, representation, approaches and future thinking," in *2024 IEEE/ACIS 24th International Conference on Computer and Information Science (ICIS)*, 2024, pp. 113–120.
- [6] 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>
- [7] A. Sharma and A. Habibi Lashkari, "Hybrid attention-enhanced explainable model for encrypted traffic detection and classification," *International Journal of Information Security*, vol. 24, no. 3, p. 144, May 2025. [Online]. Available: <https://doi.org/10.1007/s10207-025-01064-6>
- [8] M. MontazeriShatoori, L. Davidson, G. Kaur, and A. Habibi Lashkari, "Detection of doh tunnels using time-series classification of encrypted traffic," in *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2020, pp. 63–70.
- [9] Y. Lei, J. Wu, X. Sun, L. Zhang, and Q. Wu, "Encrypted traffic classification through deep learning," IAB Workshop on Management Techniques in Encrypted Networks (M-TEN), Nov. 2023, slides, Version 00. Available: <https://datatracker.ietf.org/doc/slides-mtenws-paper-encrypted-traffic-classification-through-deep-learning-lei-wu-sun-zhang-wu-00/>
- [10] B. Gahtan, R. J. Shahla, R. Cohen, and A. M. Bronstein, "Exploring quic dynamics: A large-scale dataset for encrypted traffic analysis," in *2025 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, 2025, pp. 1–6.
- [11] L.-M. Geiginger, "Classification of encrypted quic network traffic," Master's thesis, TU Wien, 2021.
- [12] X. Zhu, J. Wang, W. Ge, and X. Han, "Mt-degcl: Multi-task encrypted traffic classification with dual embedding and graph contrastive learning," *IEEE Transactions on Information Forensics and Security*, vol. 21, pp. 2220–2235, 2026.
- [13] Y. Chen and Y. Wang, "Mpfaf: Encrypted traffic classification with multi-phase attribute fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 7091–7105, 2024.
- [14] L. Garcia, G. Bartlett, S. Ravi, H. Ibrahim, W. Hardaker, and E. Kline, "Explaining deep learning models for per-packet encrypted network traffic classification," in *2022 IEEE International Symposium on Measurements & Networking (M&N)*, 2022, pp. 1–6.
- [15] P. Hermosilla, M. Díaz, S. Berríos, and H. Allende-Cid, "Use of explainable artificial intelligence for analyzing and explaining intrusion detection systems," *Computers*, vol. 14, no. 5, p. 160, 2025. [Online]. Available: <https://www.mdpi.com/2073-431X/14/5/160>
- [16] P. Khani, E. Moenaddini, N. D. Abnavi, and A. Shahraki, "Explainable artificial intelligence for feature selection in network traffic classification: A comparative study," *Transactions on Emerging Telecommunications Technologies*, vol. 35, no. 4, p. e4970, 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.4970>
- [17] A. Sharma, "Xai-driven malicious encrypted traffic detection and characterization to enhance information security," Master's Thesis, York University, Jul. 2025, available at YorkSpace Institutional Repository. [Online]. Available: <https://hdl.handle.net/10315/42974>
- [18] M.-D. Nguyen, A. Bouaziz, V. Valdes, A. Rosa Cavalli, W. Mallouli, and E. Montes De Oca, "A deep learning anomaly detection framework with explainability and robustness," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023. [Online]. Available: <https://doi.org/10.1145/3600160.3605052>
- [19] W. Lin, C. Xia, T. Wang, Y. Zhao, L. Xi, and S. Zhang, "Input and output matter: Malicious traffic detection with explainability," *IEEE Network*, vol. 39, no. 2, pp. 259–267, Mar. 2025. [Online]. Available: <https://doi.org/10.1109/MNET.2024.3481045>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTablesandAlgorithms.pdf](#)