Explainable Network Traffic Classification Using XGBoost and SHAP: Interpreting Feature Contributions Across Traffic Classes

Boubakar Seddik Deghem, Lahcene Aid, and Kadda Mostefaoui

University of Tiaret, Tiaret, Algeria

https://doi.org/10.26636/jtit.2026.3.2724

Submitted: 13 Jul. 2026, Accepted: 31 Aug. 2026

Abstract — Network traffic classification has become an essential component of modern network management, cyber security, and quality-of-service provisioning. The widespread adoption of encryption technologies has reduced the effectiveness of traditional traffic identification techniques, leading to the use of ML approaches based on flow-level characteristics. Although several machine learning and deep learning models were evaluated, the black-box nature complicates the practical deployment in security critical environments. This research proposes a framework for ML-based network traffic classification of explainable artificial intelligence (XAI). The CIC-Darknet2020 dataset is divided into four classes: non-TOR, non-VPN, TOR and VPN. Additionally, machine learning algorithms, including ID3, k-nearest neighbors (KNN), random forest, CatBoost, XGBoost, and LSTM are used as a deep learning approach. The evaluation is carried out through stratified split of trains and 10-fold cross-validation, while XGBoost is classified for explainability analysis. By ensuring model transparency, SHAP (SHapley Additive exPlanations) identifies the most influential features contributing to classification predictions. Furthermore, a novel category SHAP analysis is introduced by grouping higher-level behavioral categories, including temporal, statistical, rate-based, and TCP-related features. The results revealed that temporal traffic characteristics and transport layer behavioral features influence classification outcomes, particularly for encrypted traffic classes. Moreover, the framework demonstrates that high classification performance and model interpretability are achievable simultaneously within a category-level study that enhances the transparency, trustworthiness, and practical applicability of machine learning-based network traffic classification systems.

Keywords — deep learning, explainable AI, machine learning, network traffic classification

1. Introduction

The rapid proliferation of anonymity-preserving technologies such as Tor [1] and virtual private networks (VPN) services [2] has transformed the significance of modern network communication. Although these technologies are being investigated to improve user privacy and protect against surveillance by encrypting data and masking traffic, they also present network monitoring and security enforcement-related challenges.

Recent reports indicate that the overall percentage of web encryption via the HTTPS and TLS protocols will stabilize between 94.6% and 95.7% globally [3], with TLS 1.3 becoming the dominant protocol, accounting for more than 72.8% of encrypted traffic. This shift is accompanied by a stable HTTP/3 (QUIC-based over UDP) adoption rate equaling 19.8% to 21.1% of total web requests due to operational challenges affecting throughput on wired networks, while HTTP/2 maintains a 52.3% share. As far as emerging security trends are concerned, the use of postquantum cryptography (PQC) algorithms such as X25519MLKEM768 on the browser side has increased to 55.7% to protect against "collect now, decompose later" attacks [4]. While encryption enhances user privacy, it has also led to a security vulnerability. Cybersecurity reports [5] show that more than 86% of threats and malware are now hidden inside encrypted tunnels, while one in three organizations is unable to scan encrypted traffic internally due to privacy and performance considerations.

As encrypted traffic increasingly dominates network environments, traffic identification methods based on port numbers and deep packet inspection (DPI) have become less effective. Since payload contents are no longer accessible, modern traffic analysis systems must depend on behavioral and statistical characteristics, such as packet interarrival times, flow durations, packet size distributions, traffic rates, and detonate patterns. Consequently, automatic network traffic classification (ANTC) has emerged as an important component of network management, cybersecurity, quality of service provisioning, and anomaly detection. Precise traffic identification enables network administrators to optimize resource allocation, enforce security policies, detect malicious activities, and improve network performance.

To cope with these limitations, researchers have increasingly adopted machine learning (ML) and deep learning (DL) techniques capable of learning distinctive patterns directly from traffic statistics and flow-level characteristics. In the past, numerous studies have investigated ML-based classifiers,

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License.
For more information, see https://creativecommons.org/licenses/by/4.0/

including random forest, support vector machines, gradient boosting, and XGBoost. More recently, DL architectures such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs), gated recurrent units (GRUs), and hybrid models have demonstrated promising performance across diverse traffic environments. These approaches have mainly advanced the state of high classification accuracy on several benchmark datasets.

Most classification systems operate as black-box models, making it difficult for network administrators and security analysts to understand the factors that influence the predictions of the model. Furthermore, although explainable artificial intelligence (XAI) techniques have recently attracted increasing attention, most studies focus on individual feature importance while skipping the analysis of broader traffic feature categories, such as temporal, statistical, rate-based, and protocol-related characteristics. Consequently, the behavioral patterns on which traffic classification predictions are based remain insufficiently understood, particularly in encrypted traffic environments.

In such a context, this study addresses the problem of classifying encrypted network traffic, focusing on distinguishing between non-TOR, non-VPN, TOR, and VPN classes. The proposed framework is designed to classify multiple traffic categories while providing transparent and interpretable explanations of the model predictions. Consequently, this can be regarded as a significant issue in the field of cybersecurity in light of the increasing use of encrypted applications and the difficulty of analyzing them through traditional methods. Furthermore, this study proposes an explainable network traffic classification framework based on XGBoost and SHAP. The study also addresses the challenges related to ML models limited interpretability by using the category-SHAP method.

By revealing how feature categories influence classification predictions, the framework facilitates more effective feature selection, simultaneously improving reliability, interpretability, and performance of network traffic classification systems.

2. Related Work

In recent years, the classification of network traffic has attracted considerable interest due to its importance in the proves of managing networks, ensuring the required QoS, and guaranteeing cybersecurity. Earlier methods used manual feature engineering and traditional ML techniques to perform network traffic classification. Decision trees were effective in sorting multiple types of network traffic using both categorical and continuous features, thus allowing the identification of complex traffic patterns and malicious behaviors [6]. Similarly, interpretable probabilistic approaches based on Markov chains and hidden Markov models were investigated for traffic characterization and prediction. These methods offered precise information on traffic behavior while achieving competitive performance compared to advanced ML techniques [7].

As modern communication becomes more complex and encryption becomes more widespread, researchers have increasingly moved towards using deep learning for identifying Internet traffic. In [3], the authors used deep learning to find distinctive characteristics of raw Internet traffic data, by taking advantage of the earlier studies. They proposed a traffic identification framework based on deep learning methods that reduced the dependence on manually engineered features by automating extraction relevant to the characteristics of the traffic network flows. The authors of [4] developed an end-to-end encrypted traffic classification framework specifically implementing 1D-CNNs, which resulted in substantial performance improvements on the ISCX VPN non-VPN datasets without dependence on hand-crafted features.

Further advances were achieved with the deep packet approach, proposed in [8], which combines stacked autoencoders and CNNs to automatically learn traffic representations from raw packet data. The framework achieves high classification performance for both application identification and traffic categorization tasks, including encrypted, VPN, and non-VPN.

More recently, lightweight deep learning architectures based on depth-wise separable convolutions (DSCs) have been introduced to reduce computational complexity while maintaining high classification performance. Such models have achieved state-of-the-art accuracy in benchmark datasets and demonstrated suitability for resource-constrained environments, including IoT and edge computing [9].

Despite the remarkable classification performance achieved by ML and deep learning models, most existing approaches operate as black-box systems, providing limited insight into the rationale behind their predictions. Consequently, explainable artificial intelligence (XAI) has appeared as a promising research direction to improve transparency, trustworthiness, and interpretability of network traffic analysis [10]. XAI techniques can be widely classified into model-specific and model-agnostic approaches, connected to local and global scales operating countless popular methods including SHAP [11], LIME [12], integrated gradients [13], layer-wise relevance propagation (LRP) [14], and occlusion analysis [15].

To address the black-box nature of deep learning models, researchers used XAI techniques to interpret multimodal traffic classification systems and generate global explanations of classifier behavior [16]. Similarly, explainable anomaly detection frameworks have combined ML models with SHAP explanations to improve transparency in encrypted traffic analysis, achieving high detection performance on data sets such as CIC-Darknet2020 while identifying the most influential traffic characteristics [17].

More recently, in [18], the effectiveness of SHAP in encrypted traffic analysis was demonstrated by identifying critical traffic characteristics such as packet size, inter-arrival time, and TLS-related attributes, while maintaining detection accuracy above 99%.

Beyond traffic classification, XAI has also been applied to cybersecurity domains such as IoT anomaly detection and log-based anomaly analysis. In IoT environments, SHAP-

Tab. 1. Comparative analysis with existing approaches.

Ref.	Classification	Encrypted traffic	SHAP	Class exp.	Feature category	Behavioral interp.
[9]	✓	✓	✓	✓	×	×
[16]	✓	✓	✓	×	×	×
[17]	×	✓	✓	✓	×	×
[18]	✓	✓	✓	✓	×	×
Proposed	✓	✓	✓	✓	✓	✓

based frameworks have been used to provide explanations for anomaly detection decisions, thus enhancing trust in AI-driven security systems [19]. Similarly, the SXAD framework integrated SHAP into the logarithmic anomaly detection process, providing local and global explanations while also achieving near-perfect detection performance, and enabling root-cause analysis of abnormal events [20].

Recent studies focusing on XAI improved the interpretability of network traffic analysis systems; most existing approaches focus primarily on individual feature importance or model-level explanations. Adequately, understanding of traffic behavior through higher-level semantic feature groups has received little attention. Consequently, there remains a need for explainability frameworks capable of providing behavioral interpretations of network traffic by analyzing the contribution of feature categories rather than isolated features. This research gap is the main focus of this work, which introduces a category-based SHAP analysis framework to explain encrypted traffic classification decisions from a behavioral perspective.

As summarized in Tab. 1, existing studies have integrated SHAP into network traffic analysis to provide explanations and class-level interpretations. However, most of these works focus on the importance of individual features without investigating the contribution of higher-level behavioral feature groups. Adequately, no previous study has introduced a category-based SHAP framework for encrypted traffic classification that explains model predictions through semantic feature categories. Therefore, this work deals with existing explainability approaches by providing both feature-level and category-level interpretations, enabling a more comprehensive understanding of network traffic behavior.

3. Proposed Framework

The framework's structure comprises four main stages: data preprocessing, feature engineering, model training, and explainability analysis using SHAP-based methods (Fig. 1).

3.1. Data Preprocessing and Feature Preparation

This study uses the CIC-Darknet2020 dataset [21], which is widely used for encrypted and darknet traffic analysis, and provides realistic network traffic traces collected from various communication scenarios. To create a representative dataset, we merge our previously generated data sets (ISCXTor2016 [22] and ISCXVPN2016 [23]) and combine the respective VPN and TOR with the corresponding darknet categories.

Each network flow is represented by a set of statistical and behavioral features extracted using CICFlowMeter. These features describe various characteristics of network traffic, including flow duration, packet length statistics, interarrival times, packet rates, TCP flag information, and other flow-level attributes. Since the data set relies on flow-based features rather than payload inspection, it is particularly suitable for analyzing encrypted traffic while preserving user privacy.

For the purposes of this study, the classification of the original data was retained, comprising four main categories of traffic: non-TOR, non-VPN, TOR, and VPN.

The CIC-Darknet2020 dataset was subjected to a preprocessing stage to ensure data quality and suitability for ML-based traffic classification. First, records containing missing values were removed from the dataset. Subsequently, the infinite values resulting from flow-level calculations were replaced with missing values and eliminated to avoid numerical inconsistencies during the model training phase.

To prevent information leakage and improve model generalization, non-informative identifiers such as Src IP, Dst IP, Src Port, Dst Port and Flow ID were excluded from the feature set. Categorical attributes were transformed into numerical representations using label encoding, while numerical flow-based features were retained in their original form (Fig. 2). As a result, each network flow was represented by a set of statistical and behavioral characteristics describing temporal behavior, packet size distributions, traffic rates, and TCP-related dynamics, enabling encrypted traffic classification without relying on packet payload inspection.

3.2. Traffic Modeling

To classify encrypted network traffic, six algorithms were evaluated: ID3, k-nearest neighbors (KNN), random forest, CatBoost, XGBoost, and LSTM as a deep learning approach. Due to their capabilities, the first five models are conventional ML algorithms designed for tabular flow level data, while LSTM was selected to capture long-term temporal dependencies and sequential patterns in network traffic flows. Such an approach renders the model well prepared in terms of temporal characteristics, encrypted traffic, and distinguishing between different traffic classes by identifying the most effective model, whereas random forest was included as an ensemble tree-based baseline to provide a comparison with

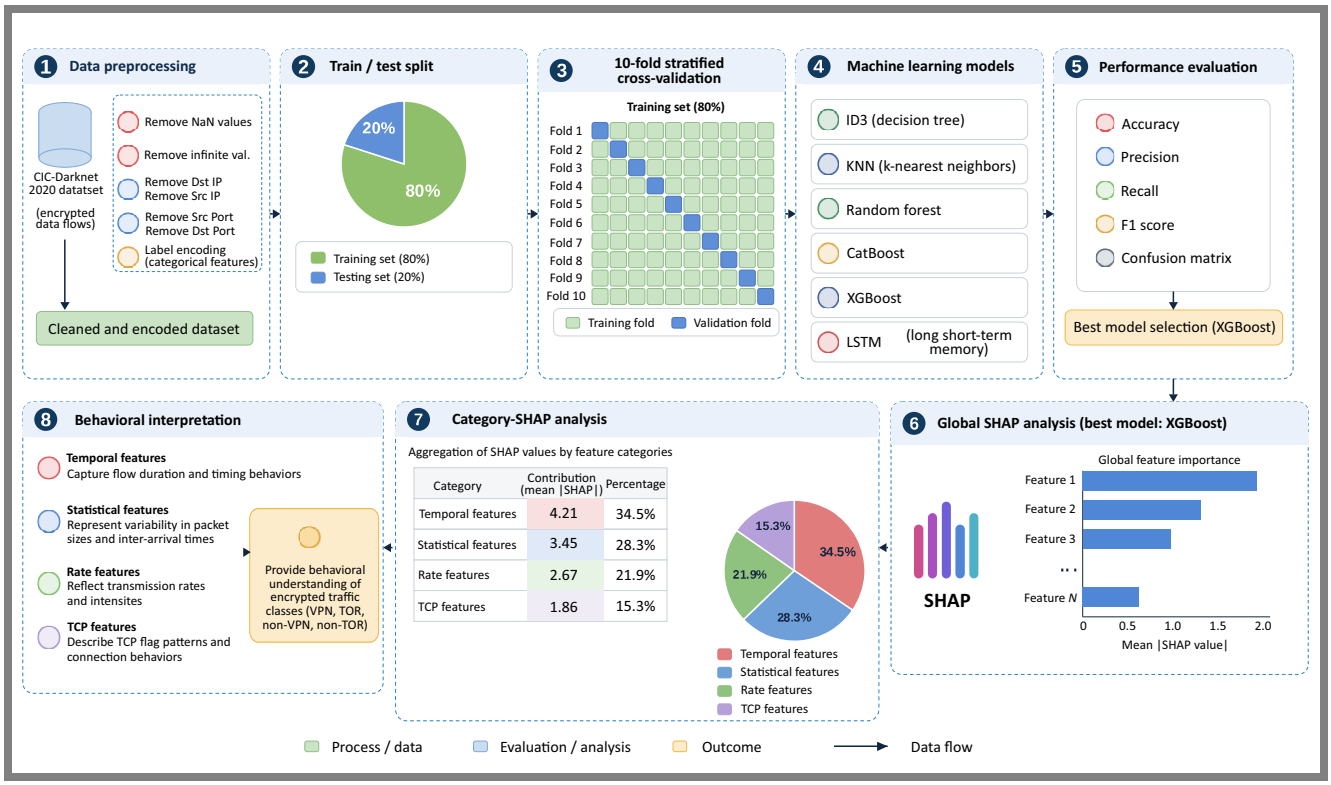


Fig. 1. Overview of the proposed framework.

the gradient-boosting approaches, namely CatBoost and XGBoost.

The extreme gradient boosting algorithm is widely used as a model in multiclass classification, due to its high proficiency in dealing with tabular data and its ability to model nonlinear feature relationships. The model is developed using 300 decision trees to guarantee incremental learning and improve performance. To avoid overfitting, the maximum depth of each tree is limited to 8 levels, achieving balance between learning ability and avoiding overfitting. In addition, a 0.05 learning rate is adopted to ensure gradual and stable updates during the training process.

To increase the generalizability of the model, subsampling is used at 80%, so that each tree is trained on a random

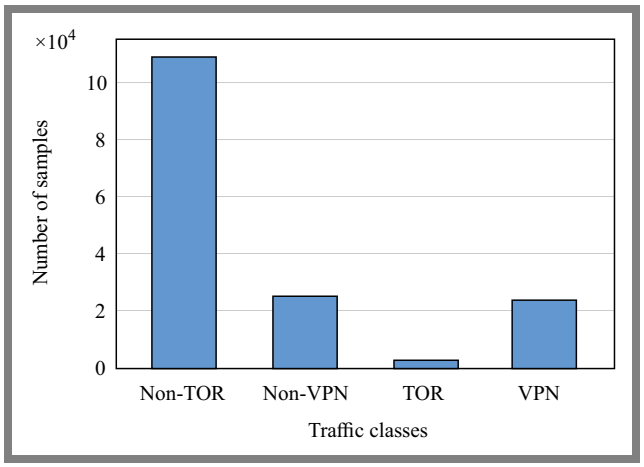


Fig. 2. Distribution of traffic categories in the Darknet 2020 dataset.

sample of training data to select a subset of features during the construction of each tree, reducing correlation between the trees and improving performance. Moreover, the softmax function is adopted to directly retrieve the predicted class using the multiclass loss function to monitor the learning quality during the training phase, considering that the issue is mainly about the classification of four different categories. To guarantee reproducibility of the results, the random seed is set at a parameter value of 42.

To evaluate the generalization capacity of the model's generalization ability and its capability to reduce the impact of data splitting, a k-fold stratified cross-validation strategy was applied to the training set in such a way that the training data is divided to ten equal components while preserving the exact same relative class distribution in each component [24].

In each training course, 9 components are used to train the model and 1 component is relied upon for the verification of results. The choice of $k = 10$ provides a balance between computational efficiency and reliable performance estimation [25]. The process repeats itself ten times in such a way that each component is used only once for verification purposes. In the end, the average classification accuracy is calculated across all folds, providing a more reliable evaluation of the model's performance.

Each model was trained to classify network flows into traffic categories: non-TOR, non-VPN, TOR, and VPN, using flow-level statistical and behavioral features extracted from network traffic. After comparing the classification performance of all candidate models using standard evaluation metrics, the best-performing model was selected and then used for the SHAP-

based explainability analysis to investigate the contribution of individual features and feature categories to the classification process.

3.3. Feature Learning and Selection via SHAP

To move beyond black-box prediction, the Shapley additive explanations (SHAP) approach is integrated to quantify feature contributions to model predictions. First, a global SHAP analysis is performed to classify all features according to their importance. Instead of selecting an arbitrary number of characteristics, a cumulative importance strategy is applied, retaining only those that contribute to 95% of the total explanatory power.

This results in a reduced and meaningful subset of discriminative features, ensuring both efficiency and interpretability.

3.4. Category-based Explainability Analysis

To provide a higher-level semantic interpretation, selected features are grouped into four behavioral categories: temporal, statistical, rate, and window-based features. Then, category-SHAP is applied to quantify the contribution of each category to the prediction of each traffic class. This step enables the transition from feature-level explanation to behavior level understanding, allowing to interpret the influence of different traffic dynamics on classification decisions in VPN and TOR environments.

To obtain the category-SHAP values, the individual SHAP contributions of features belonging to the same category are aggregated.

Let ϕ_i denote the SHAP value of feature i, and C_k represent the k-th feature category containing a set of features $\{i_1, i_2, \ldots, i_m\}$. The category-SHAP value for category C_k is computed as the sum of the absolute SHAP values of all features belonging to that category as:

$$\phi(C_k) = \sum_{i \in C_k} \left|\phi_i\right|, \tag{1}$$

where $\phi(C_k)$ represents the magnitude of the overall contribution of category C_k to the model prediction, and $|\phi_i|$ measures the magnitude of the contribution of each feature regardless of its direction (positive or negative impact).

To enable a comparison between different categories, we aggregated the values from Eq. (1) which are normalized according to:

$$\Phi_{\text{Norm}}(C_k) = \frac{\phi(C_k)}{\sum_{j=1}^{K} \phi(C_j)}, \tag{2}$$

where K is the total number of feature categories.

The resulting normalized category-SHAP values represent the relative importance of each category in the final prediction. Algorithm 1 illustrates the proposed prediction method.

Algorithm 1 Category-SHAP aggregation and normalization

Require: SHAP values $\boldsymbol{\phi} = \phi_1, \phi_2, \ldots, \phi_N$ and feature categories $\mathcal{C} = C_1, C_2, \ldots, C_K$
Ensure: Normalized category-SHAP values $\Phi_{\text{Norm}}(C_1), \ldots, \Phi_{\text{Norm}}(C_K)$

for $k = 1$ to K **do**
  $\phi(C_k) \leftarrow 0$
  for each feature $i \in C_k$ **do**
    $\phi(C_k) \leftarrow \phi(C_k) + |\phi_i|$
  end for ▷ Eq. (1)
end for
$S \leftarrow \sum_{j=1}^{K} \phi(C_j)$
for $k = 1$ to K **do**
  if $S > 0$ **then**
    $\Phi_{\text{Norm}}(C_k) \leftarrow \frac{\phi(C_k)}{\sum_{j=1}^{K} \phi(C_j)}$ ▷ Eq. (2)
  else
    $\Phi_{\text{Norm}}(C_k) \leftarrow 0$
  end if
end for
return $\Phi_{\text{Norm}}(C_1), \ldots, \Phi_{\text{Norm}}(C_K)$

4. Results and Discussion

This section presents the experimental findings of the proposed framework and discusses both classification performance and explainability insights derived from the SHAP-based analysis.

4.1. Classification Performance

The XGBoost classifier achieved strong performance in distinguishing between the four traffic categories (non-TOR, non-VPN, TOR, and VPN). Using an 80/20 stratified split and 10-fold cross-validation, the model demonstrated stable accuracy across folds, confirming its robustness on imbalanced and high-dimensional network traffic data.

The confusion matrix further indicates that most misclassifications occur between closely related encrypted traffic types (e.g., VPN vs. non-VPN), which is expected due to similar flow-level statistical behavior. Overall, the model achieves high predictive capacity, confirming that flow-based features extracted from CICFlowMeter are sufficient to model encrypted traffic patterns without relying on payload inspection.

Figure 3 presents the confusion matrix of the XGBoost classifier in the CIC-Darknet2020 dataset. Most samples are correctly classified, as indicated by the strong concentration of values along the main diagonal. The model achieved excellent recognition of non-TOR (22 068 samples), non-VPN (4 629 samples), TOR (257 samples) and VPN (4 318 samples).

The majority of misclassifications occurred between VPN and non-VPN, suggesting that these classes share similar statistical characteristics despite the presence of VPN tunneling. On the contrary, very limited confusion was observed between TOR and VPN, indicating that the model successfully

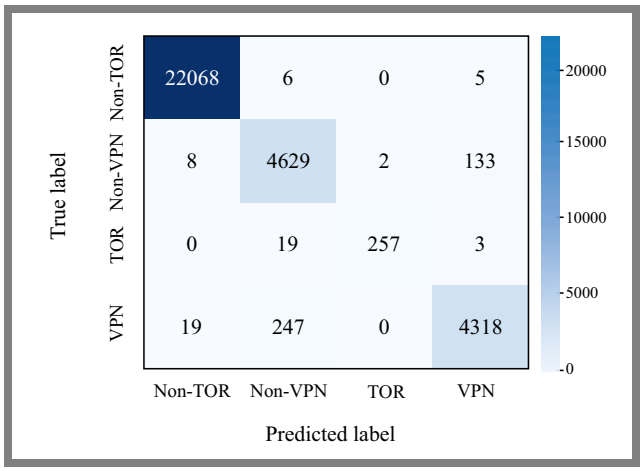


Fig. 3. Confusion matrix of the XGBoost model.

Tab. 2. Performance comparison of different models for encryption traffic classification.

Model	Accuracy	Precision	Recall	F1 score
ID3	0.9824	0.9508	0.9546	0.9526
KNN	0.9690	0.9687	0.9690	0.9688
Random forest	0.9814	0.9633	0.9471	0.9549
CatBoost	0.9837	0.9731	0.9493	0.9606
LSTM	0.9638	0.9348	0.9100	0.9217
XGBoost	0.9861	0.9862	0.9861	0.9860

captures the distinctive behavioral patterns of TOR communications. Overall, the confusion matrix confirms the effectiveness of XGBoost in discriminating between encrypted and unencrypted traffic categories using flow-level features.

Table 2 summarizes the performance of the evaluated models. XGBoost achieved the highest overall performance, with an accuracy of 0.9861 and an F1 score of 0.9860. The random forest achieved an F1 score of 0.9549, providing a competitive ensemble tree baseline but remaining below the performance benchmark of the gradient boosting models. LSTM achieved an F1 score of 0.9217, which was lower than the performance of the evaluated tree-based models for the considered flow level representation. The strong performance of XGBoost demonstrates its ability to effectively capture complex traffic patterns and distinguish between classes. Consequently, XGBoost was selected as the reference model for subsequent SHAP and category-SHAP explainability analyzes.

The receiver operating characteristic (ROC) curves of the four proposed models for the selected traffic types are shown in Fig. 4. All of the ROC curves are situated preferentially at the top left corner of the diagram, exhibiting an impressive balance between true positive rate (TPR) and false positive rate (FPR) throughout different threshold levels. The AUC indexes of the ROC curves are quite impressive with 1.0000 for the non-TOR, 0.9984 for the non-VPN, 0.9998 for the TOR, and 0.9985 for the VPN.

These values indicate the great discriminative power of the model, which allows it to effectively classify each traffic type. Moreover, the considerable distance of the ROC curves from the diagonal line of the random classifier would be an additional proof of the model's superiority to random guessing. Lastly, the high values of AUC mean that the model is both reliable, robust, and offers a high level of generalization in terms of its classification capability.

4.2. Global SHAP Interpretation

To improve interpretability, SHAP analysis was applied to quantify the contribution of each feature to the model's predictions.

The global SHAP results revealed that the model is primarily driven by a small subset of dominant features. The cumulative SHAP analysis showed that only 35 features are required to explain 95% of the model's predictions, significantly reducing the feature space while preserving interpretability, As shown in Fig. 5, of all features, *Idle Max* was identified as the most influential variable, followed by *FWD Init Win Bytes*, *Bwd Init Win Bytes* and *Idle Mean*. These results indicate that temporal behavior and TCP window-related characteristics are the most discriminative factors in encrypted traffic classification, while traditional packet-level features contribute less significantly.

4.3. Category-SHAP Analysis

For a better understanding of the impact of the features, the extracted SHAP features were classified into four semantic groups. The classification success shows that the temporal characteristics always exerted the greatest impact on the entire traffic class. It is evident from Fig. 6 that the rate features were the second most effective group after the temporal features for TCP and non-TOR. Statistical features have shown moderate and steady results when it comes to their contribution. Meanwhile, the window features were able to exert the least substantial impact in general, although they still played a role in distinguishing the classes.

The class-level analysis showed that while VPN relied on temporal features the most, non-TOR was influenced the most by all feature groups. Overall, the suggested model does not rely on payloads much when it comes to classification.

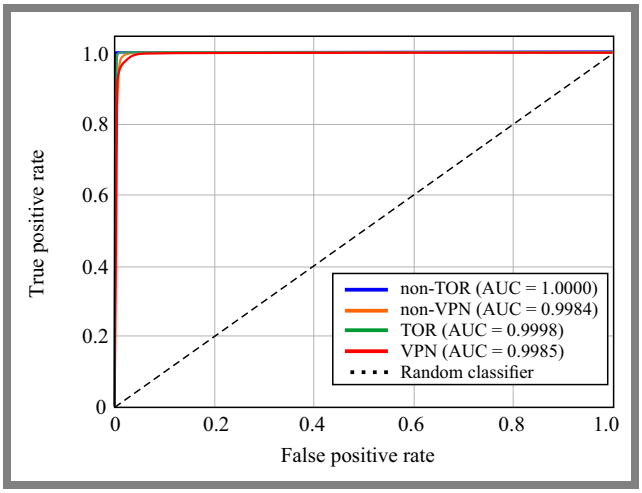


Fig. 4. ROC curves and corresponding AUC values for four types of traffic.

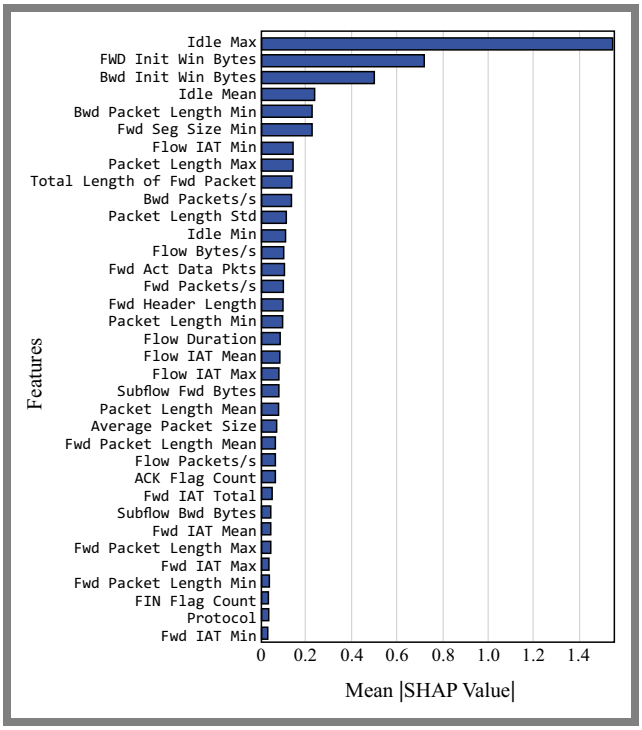


Fig. 5. 35 selected features representing 95% cumulative SHAP importance.

Figure 6 presents the category-SHAP analysis for the four traffic classes. Temporal features contribute to the classification process the most, particularly for non-VPN and VPN, indicating that timing-related behavior is the primary factor used by the model. TCP window features also play a significant role, especially for TOR and non-TOR. On the contrary, rate features show the lowest contribution across all classes.

These results suggest that the model relies mainly on traffic behavior and TCP communication dynamics to distinguish encrypted traffic categories.

4.4. Class-level Behavioral Insights

The results of the category-level SHAP analysis are shown in Tab. 3 and reveal different behavioral patterns across the various traffic classes. Temporal characteristics provide the dominant contribution for non-TOR (38.99%), non-VPN

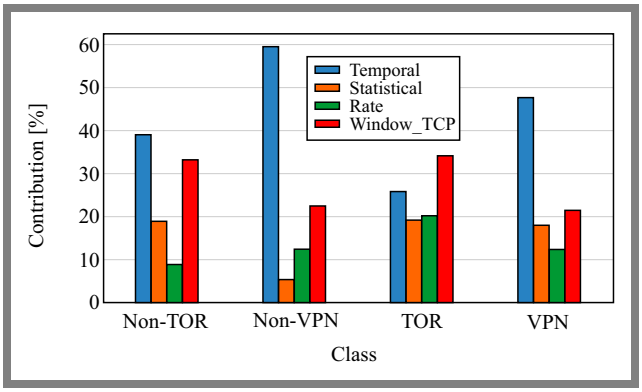


Fig. 6. Category-SHAP contribution analysis across traffic classes.

Tab. 3. Category-SHAP contribution analysis across traffic classes.

Class	Temporal	Statistical	Rate	Window_TCP
Non-TOR	38.99%	18.86%	8.90%	33.25%
Non-VPN	59.62%	5.35%	12.52%	22.52%
TOR	25.86%	19.32%	20.42%	34.40%
VPN	47.85%	18.12%	12.51%	21.52%

(59.62%), and VPN (47.85%), indicating that temporal flow characteristics are important for their classification.

In contrast, Window_TCP is the most influential feature category for TOR (34.40%), followed by rate (20.42%) and statistical features (19.32%). Also, non-TOR shows a substantial contribution from Window TCP (33.25%), meanwhile rate features are characterized by a comparatively lower contribution (8.90%).

Broadly speaking, the results indicate that the model exploits observable temporal and transport-layer behavioral characteristics, rather than payload information distinguishing encrypted and anonymized traffic.

4.5. Key Findings and Implications

Combining XGBoost with SHAP-based interpretability delivers robust classification performance and meaningful interpretability. The cumulative SHAP approach reduces the mean of the feature dimensions, while maintaining interpretability. Additionally, it makes the model more efficient and applicable. Furthermore, category-SHAP analysis provides a deeper understanding of traffic behavior. In particular, valuable cybersecurity applications are as interpretable and confident as model predictions.

To evaluate the effectiveness of the main framework, it was also compared to the LSTM approach, given its ability to model dependencies and temporal relationships in sequencing network traffic. The performance differences among XGBoost, random forest, and LSTM can be attributed primarily to the representation of the input data. Although network traffic contains some temporal characteristics, the data used in this study are represented as aggregated flow-level features rather than fine-grained sequences.

This representation favors tree-based models which can effectively capture non-linear relationships and interactions among structured features, while also limiting the ability of LSTM to exploit temporal dependencies.

Nevertheless, the superior performance of the proposed model indicates that effective traffic classification relies not only on temporal information or transport-layer characteristics, but also on complementary statistical, behavioral, and protocol-related patterns. By capturing these multidimensional characteristics, the proposed approach provides a more comprehensive representation of the behavior of network traffic, leading to better performance.

5. Conclusions

This study presented an explainable machine learning framework for encrypted network traffic classification using flow-based statistical and behavioral features extracted from the CIC-Darknet2020 dataset. Several algorithms, including ID3, k-nearest neighbors (KNN), random forest, CatBoost, XGBoost, and LSTM were employed to evaluate the classification of network traffic into four categories: non-TOR, non-VPN, TOR, and VPN. The experimental results demonstrated that the selected model offers good classification performance, confirming the effectiveness of flow-level features for encrypted traffic analysis without relying on packet payload inspection.

To improve model transparency, SHAP was used to explain the contribution of individual characteristics to classification decisions. The global SHAP analysis revealed that a limited subset of characteristics accounted for most of the model's predictive behavior, with temporal and TCP-related attributes emerging as the most influential factors. Furthermore, a cumulative SHAP-based feature selection strategy showed that a reduced set of features was sufficient to explain most of the model's predictions.

The main contribution of this work is dependent on the introduction of a category-SHAP analysis framework, which extends traditional feature-level explanations by grouping features into higher-level behavioral categories. This approach provides a more intuitive understanding of encrypted traffic behavior and enables cybersecurity analysts to interpret classification decisions from a behavioral perspective rather than through isolated features alone.

In general, the proposed framework demonstrates that explainability and high predictive performance can be achieved simultaneously in encrypted traffic classification. The obtained results contribute to the development of more transparent, trustworthy, and operationally useful AI-based network security systems.

References

[1] D. Sarkar, P. Vinod and S.Y. Yerima, "Detection of Tor Traffic using Deep Learning", *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, Antalya, Turkey, 2020 (https://doi.org/10.1109/AICCSA50499.2020.9316533).

[2] R. Venkateswaran, "Virtual Private Networks", *IEEE Potentials*, vol. 20, pp. 11–15, 2001 (https://doi.org/10.1109/45.913204).

[3] Cloudflare, "Worldwide Overview: Cloudflare Radar". [Online] Available: https://radar.cloudflare.com [Accessed: 10.07.2026].

[4] V. Gautam and S. Ganguli, *Securing Digital India Against AI and Cyber Threats*, Zscaler, 2026 (ISBN: 9798903339068).

[5] Y. Chandiramani, "Attention aux Fausses Versions de Super Mario Run sur Android!", *Global Security Mag Online*, 2017 https://www.globalsecuritymag.fr/Yogi-Chandiramani-Zscaler,20170116,68299.html (in French).

[6] S. Mali, M. Gujral, and A.K. Cherukuri, "Encrypted Network Traffic Classification Using Intelligent Techniques", *Cureus Journal of Computer Science*, 2025 (https://doi.org/10.7759/s44389-024-02701-2).

[7] G. Aceto *et al.*, "Characterization and Prediction of Mobile-app Traffic Using Markov Modeling", *IEEE Transactions on Network and Service Management*, vol. 18, pp. 907–925, 2021 (https://doi.org/10.1109/TNSM.2021.3051381).

[8] M. Lotfollahi, R.S.H. Zade, M.J. Siavoshani, and M. Saberian, "Deep Packet: A Novel Approach for Encrypted Traffic Classification Using Deep Learning", *ArXiv*, 2018 (https://doi.org/10.48550/arXiv.1709.02656).

[9] M. Ghaleb *et al.*, "Explainable AI for Lightweight Network Traffic Classification Using Depthwise Separable Convolutions", *IEEE Open Journal of the Computer Society*, vol. 6, pp. 908–920, 2025, (https://doi.org/10.1109/OJCS.2025.3576495).

[10] H. Hagras, "Towards Human Understandable Explainable AI", *Computer*, vol. 51, pp. 28–36, 2018 (https://doi.org/10.1109/MC.2018.3620965).

[11] S.M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", ArXiv, 2017 (https://doi.org/10.48550/arXiv.1705.07874).

[12] M.T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier", *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016 (https://doi.org/10.1145/2939672.2939778).

[13] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks", *ArXiv*, 2017 (https://doi.org/10.48550/arXiv.1703.01365).

[14] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences", *ArXiv*, 2019 (https://doi.org/10.48550/arXiv.1704.02685).

[15] M.D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks", *Computer Vision – ECCV 2014*, pp. 818–833, 2014 (https://doi.org/10.1007/978-3-319-10590-1_53).

[16] A. Nascita *et al.*, "XAI Meets Mobile Traffic Classification: Understanding and Improving Multimodal Deep Learning Architectures", *IEEE Transactions on Network and Service Management*, vol. 18, pp. 4225–4246, 2021 (https://doi.org/10.1109/TNSM.2021.3098157).

[17] K. Singh, A. Kashyap, and A.K. Cherukuri, "Interpretable Anomaly Detection in Encrypted Traffic Using SHAP with Machine Learning Models", *ArXiv*, 2025 (https://doi.org/10.48550/arXiv.2505.16261).

[18] S.N. Zeleke, A.F. Jember, and M. Bochicchio, "Integrating Explainable AI for Effective Malware Detection in Encrypted Network Traffic", *ArXiv*, 2025 (https://doi.org/10.48550/arXiv.2501.05387).

[19] A.N. Gummadi, J.C. Napier, and M. Abdallah, "XAI-IoT: An Explainable AI Framework for Enhancing Anomaly Detection in IoT Systems", *IEEE Access*, vol. 12, pp. 71024–71054, 2024 (https://doi.org/10.1109/ACCESS.2024.3402446).

[20] K. Alam *et al.*, "SXAD: Shapely eXplainable AI-Based Anomaly Detection Using Log Data", *IEEE Access*, vol. 12, pp. 95659–95672, 2024 (https://doi.org/10.1109/ACCESS.2024.3425472).

[21] Canadian Institute for Cybersecurity (CIC), "CIC-Darknet2020", *University of New Brunswick*. [Online] Available: https://www.unb.ca/cic/datasets/darknet2020.html [Accessed: 10.07.2026].

[22] Canadian Institute for Cybersecurity (CIC), "Tor-nonTor dataset (ISCXTor2016)", *University of New Brunswick*. [Online] Available: https://www.unb.ca/cic/datasets/tor.html [Accessed: 10.07.2026].

[23] Canadian Institute for Cybersecurity (CIC), "VPN-nonVPN dataset (ISCXVPN2016)", *University of New Brunswick*. [Online] Available: https://www.unb.ca/cic/datasets/vpn.html [Accessed: 10.07.2026].

[24] J.R. Almonteros and J.B. Matias, "Integration of Stratified KFold Cross Validation to Enhance Prediction Accuracy: A Comparison

Study", *2024 5th International Conference on Data Analytics for Business and Industry (ICDABI)*, Zallaq, Bahrain, 2024 (https://doi.org/10.1109/ICDABI63787.2024.10800425).

[25] O. Oyedele, "Determining the Optimal Number of Folds to Use in a K-fold Cross-validation: A Neural Network Classification Experiment", *Research in Mathematics*, vol. 10, art. no. 2201015, 2023 (https://doi.org/10.1080/27684830.2023.2201015).

Boubakar Seddik Deghem, Ph.D. Student
Artificial Intelligence and Systems Research Laboratory (LRIAS)
https://orcid.org/0009-0001-9828-4544
E-mail: boubakarseddik.deghem@univ-tiaret.dz
University of Tiaret, Tiaret, Algeria
https://www.univ-tiaret.dz

Lahcene Aid, Ph.D., Professor
Artificial Intelligence and Systems Research Laboratory (LRIAS)
https://orcid.org/0000-0001-9812-2871
E-mail: a_lahcen@esi.dz
University of Tiaret, Tiaret, Algeria
https://www.univ-tiaret.dz

Kadda Mostefaoui, Ph.D., Professor
Artificial Intelligence and Systems Research Laboratory (LRIAS)
https://orcid.org/0000-0002-6367-1245
E-mail: kadda.mostefaoui@univ-tiaret.dz
University of Tiaret, Tiaret, Algeria
https://www.univ-tiaret.dz