Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 31−39
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.163
https://ejurnal.pdsi.or.id/index.php/bids/index

Explainable Machine Learning for Multi-Class Classification of Internet Firewall Traffic

Titik Misriati, Riska Aryanti*

Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jakarta, Indonesia
Email: [1]titik.tmi@bsi.ac.id, [2,*]riska.rts@bsi.ac.id

ARTICLE INFORMATION

ARTICLE HISTORY:
Submited : May 10, 2026
Revised : May 26, 2026
Accept : May 30, 2026
Publish : May 30, 2026

KEYWORD

Ensemble Machine Learning;
Firewall Log Analysis;
Network Traffic Classification;
SHAP Explainability;
XAI

CORRESPONDENCE AUTHOR

Email: riska.rts@bsi.ac.id

A B S T R A C T

The increasing diversity and scale of network traffic introduce significant challenges in performing accurate and interpretable firewall analysis. This research aims to bridge the gap between predictive performance and model transparency by developing an explainable machine learning framework for multi-class firewall traffic classification. The study utilizes the Internet Firewall Data dataset consisting of 65,532 network traffic instances distributed across four firewall action classes and evaluates seven classification algorithms, including Decision Tree, Random Forest, XGBoost, Support Vector Machine, k-Nearest Neighbors, Naïve Bayes, and Logistic Regression. The dataset was partitioned using a stratified 80:20 hold-out approach to preserve the original class distribution and the experimental process involves data preprocessing, normalization, and validation on an independent test set using accuracy, precision, recall, and F1-score metrics. The findings reveal that XGBoost achieves the highest performance, reaching an accuracy of 99.81%, followed by Decision Tree and Random Forest. This indicates that ensemble and tree-based approaches are highly effective in modeling complex and non-linear traffic patterns. To improve interpretability, this study incorporates explainable artificial intelligence techniques, including feature importance and SHAP analysis. The results show that traffic-related attributes significantly influence classification outcomes, providing meaningful insights into firewall decision behavior

1. INTRODUCTION

The fast expansion of networked systems and internet-based services has dramatically increased the volume and complexity of network traffic, posing critical challenges for effective firewall management and security monitoring [1]. Firewall systems generate large-scale log data that contain valuable information for identifying normal and anomalous traffic patterns. Unlike traditional binary intrusion detection that merely distinguishes between benign and malicious traffic, modern firewalls execute granular, multi-class actions such as allow, deny, drop, or reset-both which require more nuanced classification approaches to capture complex policy enforcements [2]. Traditional rule-based approaches, however, are often insufficient to handle dynamic and high-dimensional network environments, leading to the growing adoption of machine learning techniques for automated traffic classification [3], [4], [5].

Recent research has proved the usefulness of machine learning models, including ensemble and tree-based methods, in achieving high classification accuracy for firewall and network traffic data [5], [6], [7], [8]. Models such as Random Forest [9] and XGBoost have demonstrated superior performance in capturing complex and non-linear relationships within network features [10], [11], [12]. Despite these advances, existing research predominantly focuses on maximizing predictive performance, often neglecting the interpretability of the models. This constraint is especially important in cybersecurity, where understanding the reasoning behind model decisions is key for confidence, accountability, and operational deployment. In practical network operations, when a high-accuracy black-box model misclassifies legitimate traffic as a threat (false positive), security analysts cannot merely trust the output; they require feature-level explanations to reverse-engineer the decision, update firewall rules, and prevent critical service disruptions [13].

Various machine learning algorithms have been employed for firewall traffic classification, each offering different strengths and limitations. Tree-based models, such as Decision Tree and Random Forest, are recognized for their capability to capture non-linear decision boundaries and provide interpretable classification rules. Ensemble learning methods, particularly XGBoost, generally achieve higher predictive performance by combining multiple weak learners through gradient boosting. Meanwhile, Support Vector Machine, k-Nearest Neighbors, Naïve Bayes, and Logistic Regression represent widely adopted baseline classifiers with distinct learning mechanisms. Therefore, comparing these algorithms under the same experimental setting is essential to identify the most suitable model for multi-class firewall traffic classification. Furthermore, prior works typically evaluate a limited number of models or lack comprehensive benchmarking across diverse algorithms, resulting in insufficient comparative insights. In addition, the integration of explainability techniques, such as SHAP, is still underexplored in firewall traffic classification, particularly in providing both global and local interpretability within a unified framework. To systematically address these critical gaps, This research focuses on developing and assessing a robust, explainable machine learning framework specifically tailored for multi-class firewall action prediction [14].

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 31−39
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.163
https://ejurnal.pdsi.or.id/index.php/bids/index

To fill these deficiencies, this research provides an explainable machine learning framework for multi-class classification of internet firewall traffic. The novelty of this research lies in three main contributions: a comprehensive comparative evaluation of seven widely used machine learning models under a unified experimental setting; the integration of Explainable Artificial Intelligence through feature importance and SHAP to provide both global and instance-level interpretability; and an in-depth analysis of feature contributions to uncover meaningful patterns in firewall behavior. The proposed approach aims to deliver not only high predictive performance but also transparent and interpretable insights, thereby supporting more reliable and intelligent network security systems. Therefore, there remains a research gap in providing a comprehensive comparison of multiple machine learning algorithms integrated with explainable artificial intelligence under a unified experimental framework for multi-class firewall traffic classification.

2. RESEARCH METHOD

The research methodology begins with dataset and data understanding to identify relevant features of firewall traffic. The preprocessing stage includes label encoding, stratified data splitting, and feature normalization to ensure data consistency and prevent bias. Multiple machine learning models are then developed and evaluated using standard classification metrics. The best-performing model is selected for further analysis using explainable artificial intelligence techniques, including feature importance and SHAP, to enhance model interpretability. Finally, the results are analyzed to derive insights and conclusions for intelligent firewall traffic classification. The research workflow used in this study is illustrated in Figure 1.

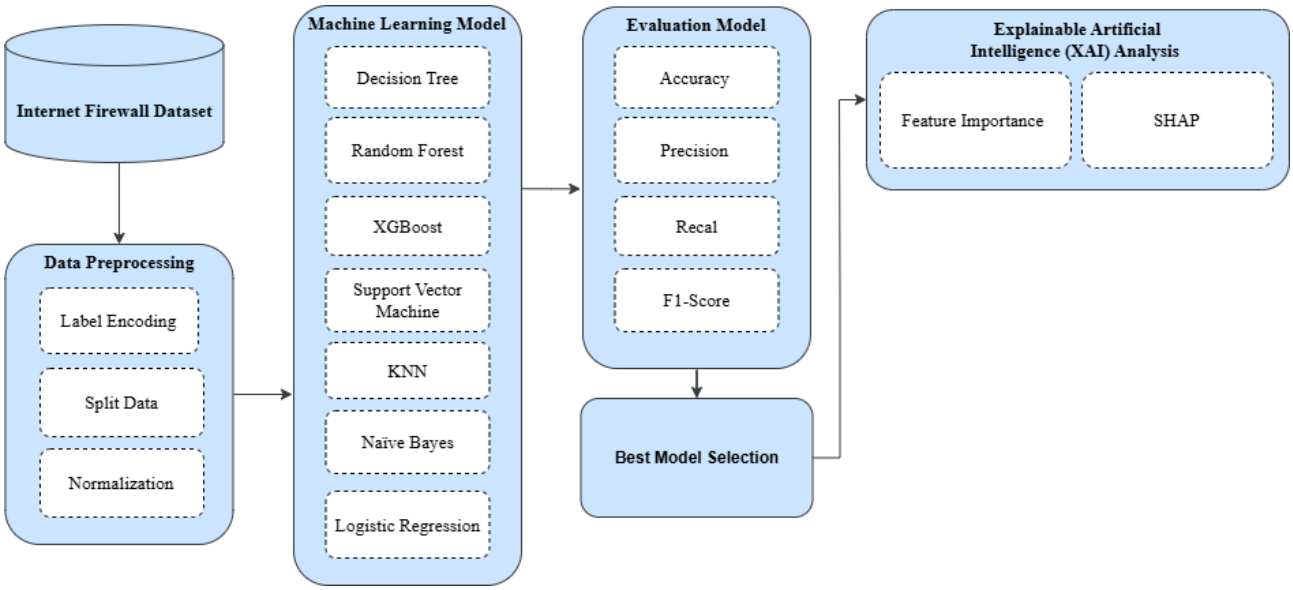

Figure 1. The Research Workflow

2.1 Dataset

The dataset used in this study is the Internet Firewall Data, which consists of 65,532 instances representing real-world network traffic captured from firewall logs [15]. The dataset contains 12 numerical features describing traffic characteristics, including Source Port, Destination Port, NAT Source Port, NAT Destination Port, Bytes, Bytes Sent, Bytes Received, Packets, Elapsed Time (sec), pkts_sent, and pkts_received. The target variable is Action, which is categorized into four classes representing different firewall decisions [16], [17]. A detailed explanation of the distribution of that class can be found in Table 1.

Table 1. Class

	class	count
0	allow	37640
1	deny	14987
2	drop	12851
3	reset-both	54

All features are quantitative and reflect key aspects of network communication, such as traffic volume, packet transmission, and connection duration. The dataset does not contain missing values, ensuring data completeness and eliminating the need for imputation techniques during preprocessing. This characteristic enhances data reliability and reduces potential bias in model training.

The diversity of features enables comprehensive representation of network behavior, making the dataset suitable for multi-class classification tasks and machine learning-based firewall analysis. Furthermore, the presence of NAT-

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 31−39
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.163
https://ejurnal.pdsi.or.id/index.php/bids/index

related attributes provides additional contextual information, allowing models to capture complex patterns in network traffic more effectively.

2.2 Data Preprocessing

Data preprocessing was conducted to ensure data consistency and optimal model performance. The categorical target variable (Action) was transformed into numerical labels using Label Encoding to facilitate compatibility with machine learning algorithms. To ensure a fair and reliable evaluation, the dataset was divided into training and testing sets using a stratified split with a ratio of 80:20, preserving the original class distribution. The training set was used to build the models, while the testing set was used exclusively for performance evaluation to avoid data leakage. To address potential scale variability among numerical features, normalization was applied using the StandardScaler technique, which standardizes features to have zero mean and unit variance. The scaler was fitted on the training data and then applied to the test data to prevent data leakage. This preprocessing pipeline ensures that the input features are well-conditioned, enabling robust and unbiased model training and evaluation.

2.3 Model

This study employs a diverse set of machine learning algorithms to comprehensively evaluate classification performance on firewall traffic data, encompassing tree-based, ensemble, instance-based, probabilistic, and linear models, as outlined in Table 1.

Table 1. Machine Learning Models Description

Model	Description	Strengths	Limitations
Decision Tree [18], [19]	A non-parametric supervised learning algorithm that partitions data into hierarchical decision rules based on feature thresholds.	Interpretable, fast training, handles non-linear relationships	Prone to overfitting, sensitive to small data variations
Random Forest [20], [21]	An ensemble method that constructs multiple decision trees using bootstrap sampling and aggregates predictions via majority voting.	High accuracy, robust to overfitting, handles high-dimensional data	Less interpretable, increased computational cost
XGBoost [22], [23]	An optimized gradient boosting algorithm that builds sequential trees with regularization to minimize loss and prevent overfitting.	Superior performance, efficient, handles complex patterns	Requires hyperparameter tuning, less interpretable
Support Vector Machine [24], [25]	A classifier that constructs optimal hyperplanes in high-dimensional space to separate classes with maximum margin.	Effective in high-dimensional spaces, robust to overfitting	Computationally expensive, sensitive to kernel choice
KNN [26], [27], [28], [29]	A non-parametric method that classifies instances based on the majority label of the k closest neighbors in feature space.	Simple, no training phase, adaptable to complex boundaries	High memory usage, sensitive to feature scaling
Naïve Bayes [26], [30]	A probabilistic classifier based on Bayes' theorem with the assumption of feature independence.	Fast, efficient, performs well on large datasets	Assumption of independence often unrealistic
Logistic Regression [26], [31]	A linear model that estimates class probabilities using a logistic function and decision boundary.	Interpretable, efficient, well-suited for linear separability	Limited in capturing non-linear relationships

2.4 Evaluation Model

To comprehensively assess the performance of the classification models, four standard evaluation metrics were employed, namely accuracy, precision, recall, and F1-score. Accuracy measures the overall proportion of correctly classified instances among all predictions, providing a general indication of model performance. However, accuracy alone may be insufficient in multi-class or imbalanced scenarios. Precision quantifies the proportion of correctly predicted positive instances among all predicted positives, reflecting the model's ability to minimize false positive errors. In contrast, recall (also known as sensitivity) measures the proportion of correctly identified positive instances among all actual positives, indicating the model's capability to detect relevant instances and reduce false negatives.

The F1-score, defined as the harmonic mean of precision and recall, provides a balanced evaluation by simultaneously considering both false positives and false negatives. This metric is particularly important in scenarios where a trade-off between precision and recall is required.

2.5 Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) was employed to enhance the transparency and interpretability of the proposed classification models, particularly in the context of complex and high-performing algorithms. XAI aims to provide

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 31−39
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.163
https://ejurnal.pdsi.or.id/index.php/bids/index

human-understandable explanations of model predictions, thereby addressing the limitations of black-box models in critical applications such as network security. In this study, interpretability was achieved through two complementary approaches: model-based feature importance and post-hoc explanation using SHAP (Shapley Additive Explanations).

Feature importance analysis quantifies the contribution of each input variable to the predictive performance of the model by measuring the reduction in impurity across decision trees. This approach provides a global interpretation of model behavior by identifying the most influential features in classification decisions. However, it does not capture the directionality or instance-level impact of features.

To overcome this limitation, SHAP was utilized as a unified framework grounded in cooperative game theory, where each feature is assigned a Shapley value representing its marginal contribution to a specific prediction. SHAP enables both global and local interpretability by explaining overall feature influence as well as individual predictions. This dual-level explanation facilitates a deeper understanding of model decision-making processes, supporting more transparent, reliable, and trustworthy deployment of machine learning models in firewall traffic analysis.

3. RESULTS AND DISCUSSION

3.1 Model and Performance Evaluation

The experimental results indicate that all evaluated models achieve high classification performance, demonstrating the effectiveness of the selected features in representing firewall traffic behavior. The classification results presented in Table 2 demonstrate that all evaluated models achieve high performance, with accuracy values exceeding 98%, indicating the effectiveness of the selected feature set in capturing firewall traffic characteristics. Among the models, XGBoost achieves the best performance, with an accuracy of 99.81%, precision of 99.79%, recall of 99.81%, and F1-score of 99.80%, confirming its superior capability in handling multi-class classification. Decision Tree and Random Forest follow closely, both achieving accuracy above 99.77%, highlighting the strong effectiveness of tree-based methods in modeling non-linear relationships. The k-Nearest Neighbors model also demonstrates competitive results, while Naive Bayes maintains high precision but slightly lower recall, indicating a tendency toward conservative predictions. In contrast, Support Vector Machine and Logistic Regression show relatively lower performance, suggesting limitations in capturing the complex and non-linear patterns inherent in firewall traffic data.

Table 2. Classification results of ML algorithms

Model	Accuracy	Precision	Recall	F1
XGBoost	0.998093	0.997912	0.998093	0.997950
Decision Tree	0.997787	0.997705	0.997787	0.997734
Random Forest	0.997787	0.997792	0.997787	0.997664
KNN	0.997177	0.996345	0.997177	0.996760
Naive Bayes	0.990845	0.997737	0.990845	0.994184
SVM	0.988556	0.987972	0.988556	0.988151
Logistic Regression	0.985504	0.985163	0.985504	0.985137

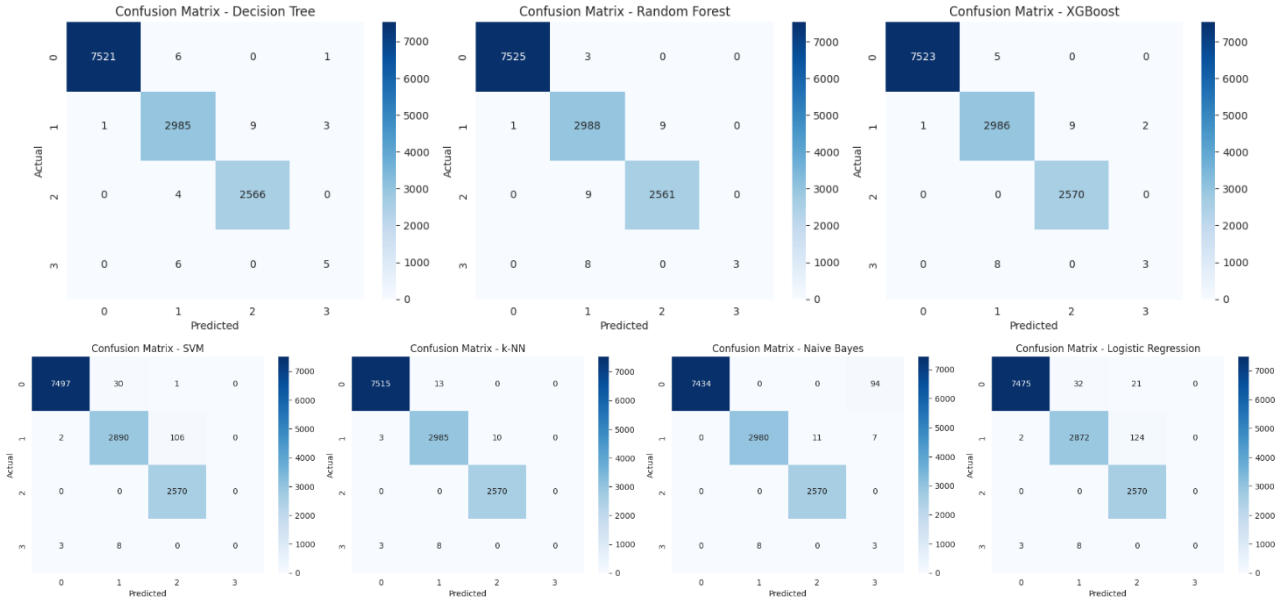

Figure 2. Confusion Matrix

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 31−39
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.163
https://ejurnal.pdsi.or.id/index.php/bids/index

The confusion matrices, presented in Figure 2, provide valuable insights into the performance of each model across the four categories: allow (class 0), deny (class 1), drop (class 2), and reset-both (class 3). Overall, the results show a strong dominance of diagonal values, indicating that most instances are correctly classified across all models. The XGBoost, Decision Tree, and Random Forest models exhibit near-perfect classification, with minimal misclassification across all categories, particularly for class 0 and class 2, which appear to be highly separable. However, slight confusion is observed in class 1 and class 3, where some instances are misclassified into neighboring classes, suggesting overlapping feature characteristics. The Support Vector Machine and Logistic Regression models demonstrate relatively higher misclassification rates, especially between class 1 and class 2, indicating limitations in capturing non-linear decision boundaries. Additionally, Naive Bayes shows notable misclassification in class 0, likely due to its independence assumption, while k-Nearest Neighbors maintains strong performance with minor confusion in adjacent classes. These findings confirm that ensemble and tree-based models provide superior class discrimination, while other models exhibit minor ambiguities in complex class boundaries.

3.2 Explainability and Feature Analysis

The SHAP-based feature importance analysis provides a comprehensive understanding of the contribution of each feature to the predictions generated by the XGBoost model across all classes shown in Figure 3. The results indicate that Elapsed Time (sec) is the most influential feature, exhibiting the highest mean absolute SHAP value, which signifies its dominant role in distinguishing firewall traffic behavior. This suggests that connection duration is a critical indicator in determining the classification of network actions.

Following this, Destination Port and Bytes also demonstrate substantial contributions, highlighting the importance of traffic endpoints and data volume in characterizing network flows. The variation of SHAP values across different classes further reveals that these features contribute differently depending on the traffic category, indicating class-specific decision boundaries learned by the model.

In contrast, features such as pkts_sent and pkts_received show minimal impact, suggesting limited discriminative power in the presence of more informative attributes. Similarly, NAT-related features contribute moderately, implying that while they provide contextual information, they are not primary drivers of classification decisions.

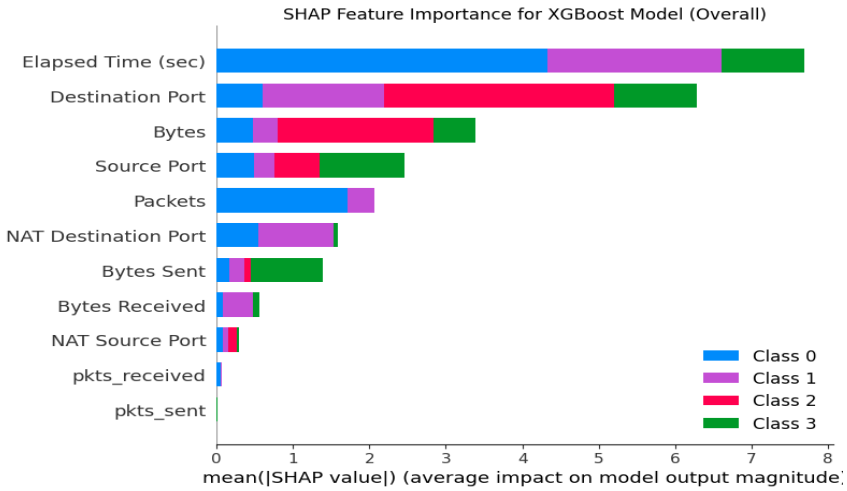

Figure 3. SHAP Feature Importance for Best Model

SHAP Forces Plots for Local Interpretability, as shown in Figure 4, with a baseline value of -8.511 representing the average log-odds prediction. Among the evaluated features, Source Port showed the strongest positive impact (SHAP = 0.5356), indicating that higher source port numbers substantially increase the likelihood of the predicted class (allow). Conversely, NAT Destination Port (−0.2297), Elapsed Time (−0.1211), Bytes (−0.01502), and Packets (−0.01499) all showed negative contributions, indicating that larger values of these features push the prediction toward the opposite class (deny).

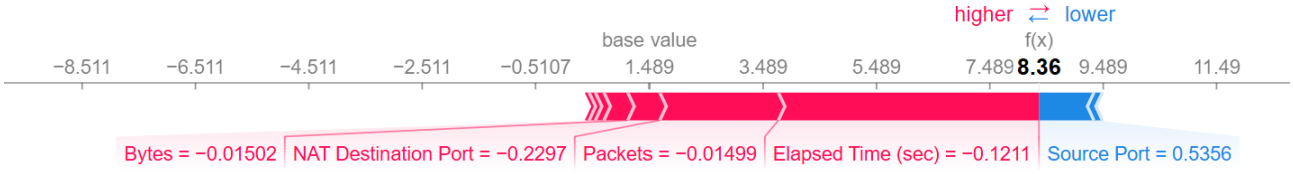

Figure 4. SHAP Forces Plots for Local Interpretability

The SHAP interaction analysis provides deeper insight into the pairwise relationships between features and their combined influence on the predictions generated by the XGBoost model. Unlike standard SHAP values, which quantify

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 31−39
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.163
https://ejurnal.pdsi.or.id/index.php/bids/index

individual feature contributions, interaction values capture how the effect of one feature depends on the value of another, enabling a more comprehensive understanding of non-linear dependencies within the data.

The results reveal that Source Port and Destination Port exhibit strong interaction effects, as indicated by the wide dispersion of SHAP interaction values across both positive and negative ranges. This suggests that the classification outcome is not solely determined by individual port values, but rather by their joint configuration, reflecting the contextual nature of network communication patterns. Similarly, interactions between Destination Port and NAT Destination Port demonstrate significant variability, indicating that NAT transformations play a critical role in shaping traffic behavior and influencing model decisions.

The global SHAP feature importance illustrated in Figure 5 provides a quantitative overview of the relative contribution of each feature to the predictions generated by the XGBoost model. The results clearly indicate that Elapsed Time (sec) is the most dominant feature, followed by Destination Port and Bytes, demonstrating that temporal characteristics and traffic volume play a crucial role in distinguishing firewall traffic behavior. Moderate contributions are observed from Source Port and Packets, while NAT-related attributes and packet-level features such as pkts_sent and pkts_received exhibit relatively low importance. This distribution suggests that the model relies primarily on a subset of highly informative features, enabling efficient and interpretable decision-making.

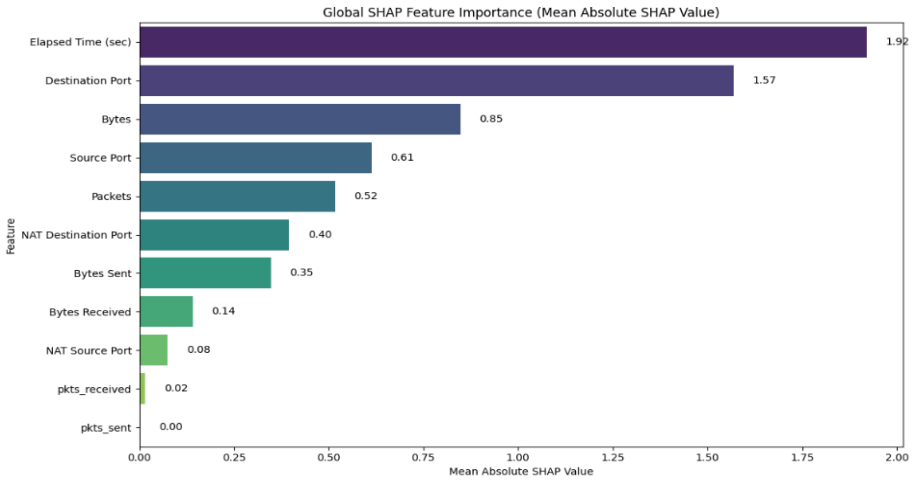

Figure 5. Global SHAP Feature Importance

In contrast, interactions involving NAT Source Port appear more concentrated around zero, suggesting a relatively weaker joint effect compared to other feature pairs. This indicates that while NAT-related features contribute to the model, their interaction strength is less dominant in determining classification outcomes.

The SHAP interaction values presented in Figure 6 further reveal the presence of non-linear dependencies between features, providing deeper insight into how feature combinations influence model predictions. Strong interaction effects are observed between port-related features, particularly Source Port and Destination Port, as well as between Destination Port and NAT Destination Port. These interactions indicate that the classification process is influenced not only by individual feature values but also by their joint behavior, reflecting the contextual nature of network communication. In contrast, interactions involving NAT Source Port appear less significant, suggesting a comparatively weaker role in shaping classification outcomes.

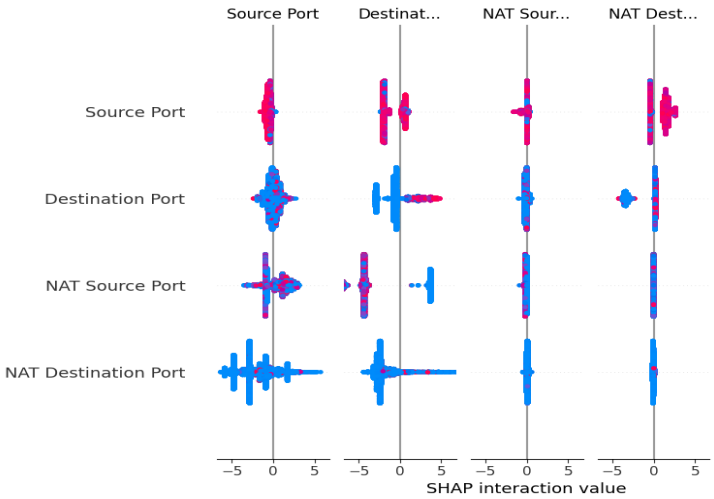

Figure 6. SHAP Interaction Value

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 31−39
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.163
https://ejurnal.pdsi.or.id/index.php/bids/index

The SHAP summary plot (beeswarm) in Figure 7 offers both global and local interpretability by visualizing the distribution and direction of feature impacts across all instances. Features such as Elapsed Time (sec) and Destination Port exhibit wide SHAP value distributions, confirming their strong and consistent influence on model predictions. The color gradient further indicates how high and low feature values contribute differently to the output, revealing complex, instance-specific behaviors. Additionally, features like Bytes and Source Port show varying directional impacts, reinforcing the presence of non-linear relationships. Overall, the combined analysis of global importance, interaction effects, and instance-level contributions demonstrates that the model effectively captures complex traffic patterns while maintaining a high level of interpretability, which is essential for reliable deployment in real-world firewall systems.

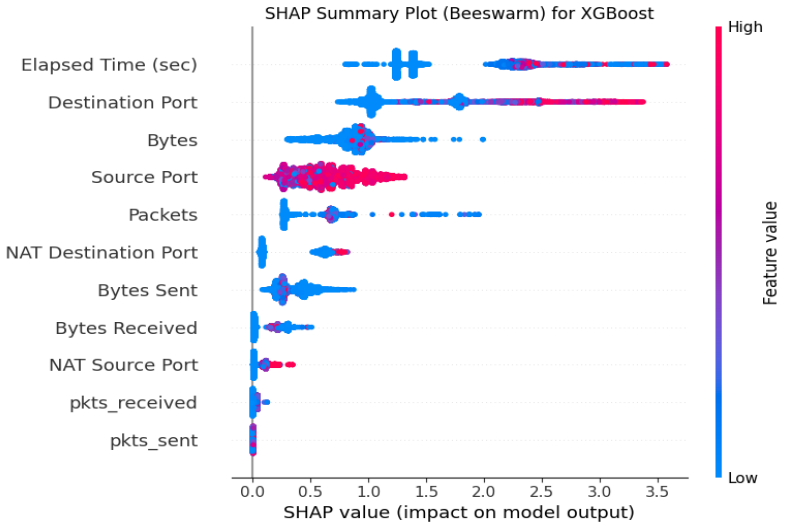

Figure 7. SHAP Summary Plot

Overall, the interaction analysis confirms the presence of complex, non-linear relationships among port-based features, highlighting the importance of feature interplay in firewall traffic classification. These findings reinforce the capability of the model to capture intricate dependencies in network data, thereby enhancing both predictive performance and interpretability in real-world cybersecurity applications.

Table 3. Comparative Performance of Internet Firewall Trafic Classification Studies

Ref	Dataset	Attribute	Split Data	Model	Imbalance Data	XAI	Accuracy Score
[2]	Internet Firewall Data	11 features	80%:20%	ANN ANN+SMOTE ANN+ADASYN ANN+BorderlineSMOTE	Yes	No	99.95%
[15]	Private real-world dataset	25 features	10-Fold Cross-Validation	KNN Naïve Bayes J48 Random Forest ANN	Yes	No	99.27% 98.50% 99.14% 99.64% 99.53%
[16]	Firat University Internet Firewall	11 features	10-Fold Cross-Validation	DT	No	No	99.80%
[32]	Firewall Log Dataset	11 features	10-Fold Cross-Validation	SimpleCart, NBTree, J48, BFTree, FT, Random Forest, Random Tree, REPTree, Decision Stump, ADTree	Yes	No	99.84%
This Study	Internet Firewall Data	11 features	80%:20%	XGBoost Decision Tree Random Forest KNN Naive Bayes SVM Logistic Regression	Yes	Yes	99.81%

Table 3 presents a comparative summary of previous firewall classification studies and the proposed study in terms of dataset characteristics, number of attributes, data splitting strategy, classification models, imbalance handling, explainability, and accuracy performance. Prior studies generally reported very high accuracy scores, ranging from 98.50% to 99.95%, using various machine learning algorithms such as KNN, Naïve Bayes, Decision Tree, Random Forest, ANN, and several tree-based classifiers. However, most of these studies focused primarily on predictive performance and did not incorporate explainable artificial intelligence to support model transparency. In contrast, this study achieved a competitive accuracy of 99.81% using the Internet Firewall Data with 11 features and an 80:20 train-test split, while also addressing class imbalance and integrating XAI analysis. Although several previous studies obtained slightly higher accuracy, the contribution of this study lies in providing a more interpretable firewall action classification framework, enabling not only accurate prediction but also a clearer understanding of the model's decision-making behavior.

4. CONCLUSION

This study demonstrates that explainable machine learning provides an effective and reliable approach for multi-class classification of internet firewall traffic, achieving both high predictive performance and strong interpretability. The experimental results indicate that XGBoost consistently delivers the highest performance across all evaluation models, achieving an accuracy of 99.81%, precision of 99.79%, recall of 99.81%, and F1-score of 99.80%, outperforming Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbors, Naïve Bayes, and Logistic Regression under the same experimental setting. Tree-based methods such as Decision Tree and Random Forest also demonstrate competitive performance, confirming their effectiveness in capturing complex, non-linear relationships inherent in firewall traffic data. In contrast, models based on linear assumptions or feature independence exhibit relatively lower performance, highlighting the importance of selecting algorithms capable of modeling intricate data patterns. Beyond predictive accuracy, the integration of Explainable Artificial Intelligence plays a crucial role in enhancing model transparency and trustworthiness. Feature importance analysis and SHAP (Shapley Additive Explanations) provide both global and local interpretability, enabling a comprehensive understanding of how individual features and their interactions influence classification outcomes. This dual-level interpretability not only validates the reliability of the model but also offers actionable insights into firewall behavior, which are essential for practical cybersecurity applications. Overall, the proposed framework contributes to the development of accurate, interpretable, and scalable machine learning solutions for intelligent firewall analysis, with strong potential for real-world deployment in adaptive and data-driven network security systems.

REFERENCES

[1] V. N. Gangineni, S. Pabbineedi, M. Penmetsa, J. R. Bhumireddy, R. Chalasani, and M. S. V. Tyagadurgam, "Strengthening Cybersecurity Governance: The Impact of Firewalls on Risk Management," *International Journal of AI, BigData, Computational and Management Studies*, vol. 2, pp. 60–68, 2021, doi: 10.63282/3050-9416.IJAIBDCMS-V2I4P106.

[2] A. Korkmaz, S. Bulut, T. Talan, S. Kosunalp, and T. Iliev, "Enhancing Firewall Packet Classification through Artificial Neural Networks and Synthetic Minority Over-Sampling Technique: An Innovative Approach with Evaluative Comparison," *Applied Sciences*, vol. 14, no. 16, p. 7426, Aug. 2024, doi: 10.3390/app14167426.

[3] H. Dhrir, M. Charfeddine, N. Tarhouni, and H. M. Kammoun, "Machine learning- and deep learning-based anomaly detection in firewalls: a survey," *J. Supercomput.*, vol. 81, no. 6, p. 761, Apr. 2025, doi: 10.1007/s11227-025-07212-y.

[4] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, *Network Traffic Anomaly Detection and Prevention*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-65188-0.

[5] M. Mingze, "Research and Application of Firewall Log and Intrusion Detection Log Data Visualization System," *IET Software*, vol. 2024, no. 1, Jan. 2024, doi: 10.1049/2024/7060298.

[6] Z. Azam, Md. M. Islam, and M. N. Huda, "Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree," *IEEE Access*, vol. 11, pp. 80348–80391, 2023, doi: 10.1109/ACCESS.2023.3296444.

[7] A. A. Afuwape, Y. Xu, J. H. Anajemba, and G. Srivastava, "Performance evaluation of secured network traffic classification using a machine learning approach," *Comput. Stand. Interfaces*, vol. 78, p. 103545, Oct. 2021, doi: 10.1016/j.csi.2021.103545.

[8] M. Rashid, J. Kamruzzaman, T. Imam, S. Wibowo, and S. Gordon, "A tree-based stacking ensemble technique with feature selection for network intrusion detection," *Applied Intelligence*, vol. 52, no. 9, pp. 9768–9781, Jul. 2022, doi: 10.1007/s10489-021-02968-1.

[9] "Hybrid AI Models in Network Security: Combining ML, DL, and Rule-Based Systems," *International Journal of Emerging Research in Engineering and Technology*, vol. 5, 2024, doi: 10.63282/3050-922X.IJERET-V5I4P111.

[10] Z. Fan and Z. You, "Research on network intrusion detection based on XGBoost algorithm and multiple machine learning algorithms," *Theoretical and Natural Science*, vol. 31, no. 1, pp. 161–166, Mar. 2024, doi: 10.54254/2753-8818/31/20241171.

[11] P. Ferreira, E. Martins, J. Silva, and P. Teixeira, "Feature Selection and XGBoost for Enhanced Intrusion Detection: A Comparative Study Across Benchmark Datasets," in *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, Apr. 2025, pp. 1–6. doi: 10.1109/ISDFS65363.2025.11012060.

[12] M. Janati and F. Messaoudi, "Network Behavior–Driven Intrusion Detection: A Hybrid Deep Learning and XGBoost Approach," in *AI-Driven Security for Next-Generation IoT Systems*, Cham: Springer Nature Switzerland, 2026, pp. 107–118. doi: 10.1007/978-3-032-08784-3_8.

[13] M. Hasan, Md. M. Hassan, S. Akter, P. Hajek, and M. Z. Abedin, "Advances in Explainable Big Data Analytics for Enhanced Cybersecurity," *Information Systems Frontiers*, Feb. 2026, doi: 10.1007/s10796-026-10701-x.
[14] Md. A. Hossain, W. Ishtiaq, and Md. S. Islam, "A Comparative Analysis of Ensemble-Based Machine Learning Approaches With Explainable <scp>AI</scp> for Multi-Class Intrusion Detection in Drone Networks," *SECURITY AND PRIVACY*, vol. 9, no. 1, Jan. 2026, doi: 10.1002/spy2.70164.
[15] M. Aljabri, A. A. Alahmadi, R. M. A. Mohammad, M. Aboulnour, D. M. Alomari, and S. H. Almotiri, "Classification of Firewall Log Data Using Multiclass Machine Learning Models," *Electronics (Basel).*, vol. 11, no. 12, p. 1851, Jun. 2022, doi: 10.3390/electronics11121851.
[16] H. AL-Behadili, "Decision Tree for Multiclass Classification of Firewall Access," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 3, pp. 294–302, Jun. 2021, doi: 10.22266/ijies2021.0630.25.
[17] T. C. Hung, D. M. Linh, H. M. Chau, N. X. Thoai, T. D. Phuong, and H. De Thu, "Anomaly-based intrusion detection leveraging optimized firewall log analysis: a real-time machine learning solution," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 15, no. 5, p. 4785, Oct. 2025, doi: 10.11591/ijece.v15i5.pp4785-4802.
[18] M. Arunika, S. Saranya, S. Charulekha, S. Kabilarajan, and G. Kesavan, "A Survey on Explainable AI Using Machine Learning Algorithms Shap and Lime," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10725120.
[19] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.
[20] S. Wali, Y. A. Farrukh, and I. Khan, "Explainable AI and Random Forest based reliable intrusion detection system," *Comput. Secur.*, vol. 157, p. 104542, Oct. 2025, doi: 10.1016/j.cose.2025.104542.
[21] J. K. Mutinda *et al.*, "Explainable AI for Breast Cancer Diagnosis: Comparative Analysis of ML Models Using Random Forest Feature Selection and SHAP Interpretability," *Asian Journal of Research in Computer Science*, vol. 18, no. 10, pp. 30–46, Oct. 2025, doi: 10.9734/ajrcos/2025/v18i10762.
[22] B. Devanathan, K. Jnana Varshitha, L. Pavan Kumar, S. A. Lakshmanan, and N. Krishna Prakash, "Explainable AI Framework Using XGBoost With SHAP and LIME for Multi-Scale Household Energy Forecasting," *IEEE Access*, vol. 13, pp. 149750–149764, 2025, doi: 10.1109/ACCESS.2025.3602673.
[23] X. Yang *et al.*, "Explainable Artificial Intelligence (XAI) framework using XGBoost and SHAP for assessing urban fire risk based on spatial distribution features," *International Journal of Disaster Risk Reduction*, vol. 129, p. 105798, Oct. 2025, doi: 10.1016/j.ijdrr.2025.105798.
[24] Z. Shuai, T. J. Kwon, and Q. Xie, "Using explainable AI for enhanced understanding of winter road safety: insights with support vector machines and SHAP," *Canadian Journal of Civil Engineering*, vol. 51, no. 9, pp. 943–953, Sep. 2024, doi: 10.1139/cjce-2023-0446.
[25] M. K. S. Gowda, Y. I. Murthy, and A. Gupta, "Explainable AI Based Support Vector Machine Models for Soaked CBR Prediction," *International Journal of Pavement Research and Technology*, May 2025, doi: 10.1007/s42947-025-00558-9.
[26] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie, "Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP," *Sci. Rep.*, vol. 13, no. 1, p. 8984, Jun. 2023, doi: 10.1038/s41598-023-35795-0.
[27] A. M. Khairuddin, S. A. M. Aris, K. N. F. K. Azir, and A. Azizan, "Using the K-Nearest Neighbor and Explainable Artificial Intelligence to Classify Arrhythmias," in *2025 International Conference on Artificial Intelligence for Sustainable Innovation (AI-SI)*, IEEE, Aug. 2025, pp. 1–6. doi: 10.1109/AI-SI66213.2025.11341336.
[28] F. Haseeb, M. Peng, F. Arshad, and W. Ali, "Uncertainty aware unsupervised fault diagnosis of PWR nuclear power plant using KNN and SHAP method," *Progress in Nuclear Energy*, vol. 193, p. 106191, Mar. 2026, doi: 10.1016/j.pnucene.2025.106191.
[29] I. Boukrouh, F. Tayalati, and A. Azmani, "Comparative SHAP Analysis on SVM and K-NN: Impacts of Hyperparameter Tuning on Model Explainability," in *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, IEEE, Aug. 2024, pp. 194–198. doi: 10.1109/IICAIET62352.2024.10729995.
[30] O. Islam, Md. Assaduzzaman, S. Akter, N. Fahad, and Md. J. Hossen, "Enhanced cervical cancer diagnosis using a novel Bayesian fusion ensemble method with explainable AI," *Sci. Rep.*, vol. 16, no. 1, p. 12306, Mar. 2026, doi: 10.1038/s41598-026-35334-7.
[31] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/htl2.12039.
[32] E. Efeoğlu and G. Tuna, "Classification of Firewall Log Files with Different Algorithms and Performance Analysis of These Algorithms," *Journal of Web Engineering*, pp. 561–594, Aug. 2024, doi: 10.13052/jwe1540-9589.2344.