CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

Explainable Machine Learning for Network Intrusion Detection Using SHAP-Based Feature Interpretation

Eka Wahyu Sholeha[1], Dery Yuswanto Jaya[2*], Qorry Aina Fitroh[3]
[1,2]Network Computer Engineering Technology, Politeknik Negeri Tanah Laut, Indonesia
[3]Industrial Engineering, Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan, Indonesia
[1]ekawahyus@politala.ac.id, [2*]deryyuswantojaya@politala.ac.id,
[3]qorry.aina.fitroh@uingusdur.ac.id

Abstract: Network Intrusion Detection Systems (NIDS) play a crucial role in protecting computer networks from increasingly sophisticated cyberattacks. Although machine learning techniques have demonstrated high detection performance, many models operate as black-box systems, making it difficult for security analysts to understand the reasoning behind prediction outcomes. This study proposes an explainable machine learning framework for network intrusion detection using the Random Forest algorithm and SHAP (SHapley Additive exPlanations)-based feature interpretation. The CICIDS2017 Friday-WorkingHours-Afternoon-DDos dataset was utilized to evaluate the effectiveness of the proposed approach. Data preprocessing included data cleaning, handling missing values, label encoding, and dataset partitioning. The Random Forest classifier was trained and evaluated using Accuracy, Precision, Recall, and F1-Score metrics. Experimental results demonstrated excellent classification performance, achieving an accuracy of 99.9889%, precision of 99.9922%, recall of 99.9883%, and F1-score of 99.9902%. Furthermore, SHAP analysis was employed to improve model interpretability by identifying the contribution of individual features to intrusion detection decisions. The results revealed that Fwd Packet Length Max, Destination Port, Avg Fwd Segment Size, and Fwd Packet Length Mean were among the most influential features affecting classification outcomes. The integration of Random Forest and SHAP not only achieved highly accurate intrusion detection but also enhanced transparency and trustworthiness by providing meaningful explanations for model predictions. Therefore, the proposed framework offers an effective and interpretable solution for network intrusion detection in modern cybersecurity environments.
Keywords: Network Intrusion Detection System; Machine Learning; Random Forest; Explainable Artificial Intelligence; Cybersecurity.

1. INTRODUCING

The rapid growth of information technology and internet-based services has significantly increased the complexity of modern computer networks. Along with the expansion of network infrastructures, organizations are facing a growing number of cybersecurity threats, including Distributed Denial of Service (DDoS), brute-force attacks, malware infections, botnet activities, and unauthorized access attempts [1], [2]. These

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

attacks can compromise the confidentiality, integrity, and availability of information systems, resulting in operational disruptions and substantial financial losses [3]. Consequently, effective cybersecurity mechanisms have become increasingly important to protect critical network infrastructures.

Intrusion Detection Systems (IDS) have emerged as essential security components for monitoring network traffic and identifying malicious activities within computer networks. Traditional IDS approaches primarily rely on signature-based detection mechanisms, which are effective for detecting previously known attack patterns. However, these approaches often fail to identify new or evolving cyber threats that do not match predefined signatures [4]. As cyberattacks continue to evolve, researchers have increasingly explored data-driven approaches to improve intrusion detection performance.

Machine Learning (ML) techniques have gained significant attention in intrusion detection research due to their ability to automatically learn complex patterns from network traffic data and classify malicious activities with high accuracy [5]. Various machine learning algorithms, including Decision Tree, Support Vector Machine (SVM), Random Forest, and Deep Learning models, have demonstrated promising performance in detecting network intrusions using benchmark datasets such as CICIDS2017 and UNSW-NB15 [6], [7]. These datasets provide realistic network traffic scenarios that enable comprehensive evaluation of intrusion detection models.

Among the available machine learning algorithms, Random Forest has become one of the most widely adopted methods because of its robustness, resistance to overfitting, and effectiveness in handling high-dimensional datasets [7]. As an ensemble learning algorithm, Random Forest combines multiple decision trees to improve classification accuracy and generalization performance. Previous studies have shown that Random Forest-based intrusion detection models can achieve competitive performance while maintaining computational efficiency, making them suitable for practical cybersecurity applications [5]-[8].

Despite their promising detection capabilities, many machine learning-based intrusion detection models are often regarded as black-box systems because their internal decision-making processes are difficult to interpret [9]. Although these models can achieve high predictive accuracy, cybersecurity analysts frequently face challenges in understanding why a particular network traffic instance is classified as malicious or benign. The lack of interpretability can reduce user trust, hinder model validation, and limit the deployment of machine learning systems in security-critical environments where transparency and accountability are essential [10], [11].

To address this challenge, Explainable Artificial Intelligence (XAI) has emerged as an important research field that aims to improve the transparency and interpretability of machine learning models [11]. XAI techniques provide insights into how model predictions are generated and enable users to understand the contribution of individual features to classification outcomes. Among various explainability approaches, SHapley Additive exPlanations (SHAP) has become one of the most widely adopted methods due to its strong theoretical foundation based on cooperative game theory and its capability to provide both global and local explanations of model behavior [12]. By assigning contribution scores to individual features, SHAP allows researchers and practitioners to identify the factors that most significantly influence prediction outcomes.

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

Table 1. Comparison of Previous Studies and Research Gap

Study	Dataset	Explainability Method	Main Contribution	Limitation
Patil et al.[13]	CICIDS2017	LIME	Explainable intrusion detection using machine learning and ensemble learning	Focused on model explainability without discussing operational cybersecurity implications
Capuano et al.[14]	Multiple Cybersecurity Domains	XAI Survey	Comprehensive survey of XAI applications in cybersecurity	Did not provide an implementation-oriented IDS framework
Rjoub et al.[15]	Multiple Cybersecurity Domains	XAI Survey	Comprehensive review of XAI techniques for cybersecurity	Limited discussion of practical deployment in intrusion detection operations
Neupane et al.[16]	IDS Studies	X-IDS Survey	Analysis of explainable intrusion detection systems and stakeholder requirements	Focused on challenges and opportunities rather than practical implementation
Proposed Study	CICIDS2017 DDoS	SHAP	Explainable DDoS detection using Random Forest and SHAP with operational interpretation	Provides practical insights for SOC analysts, network administrators, and incident response teams

As shown in Table 1, previous studies have successfully demonstrated the effectiveness of SHAP and explainable artificial intelligence techniques for intrusion detection. However, most studies primarily focused on model performance evaluation and feature importance ranking. Limited attention has been given to translating explainability results into actionable insights for cybersecurity practitioners. Therefore, this study extends existing research by discussing how SHAP-based interpretations can support Security Operations Center (SOC) analysts, network administrators, and incident response teams in understanding attack behavior and improving operational decision-making processes [14]-[16].

Recent studies have demonstrated the effectiveness of integrating SHAP into intrusion detection systems to improve model transparency while maintaining competitive classification performance [17], [13], [18]. SHAP-based explanations help security analysts understand attack characteristics, validate model predictions, and support more informed cybersecurity decision-making processes. Furthermore, explainable machine learning approaches are increasingly recognized as a critical requirement for trustworthy artificial intelligence systems in cybersecurity applications [11].

However, although numerous studies have focused on improving intrusion detection accuracy, research addressing both predictive performance and model interpretability remains relatively limited [13], [18]. Therefore, there is a need to develop intrusion detection models that not only achieve effective attack detection but also provide meaningful explanations for prediction outcomes. Based on this motivation, this study

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

proposes an explainable machine learning framework for network intrusion detection using the Random Forest algorithm and SHAP-based feature interpretation. The CICIDS2017 dataset is utilized to evaluate model performance and identify the contribution of network traffic features to intrusion classification. The results are expected to improve transparency, trustworthiness, and practical applicability of machine learning-based intrusion detection systems in cybersecurity environments.

2. METHOD

This study proposes an explainable machine learning framework for network intrusion detection using the Random Forest algorithm and SHAP-based feature interpretation. The proposed framework consists of five main stages: dataset collection, data preprocessing, model development, performance evaluation, and explainability analysis. The overall research workflow is presented in Figure 1.

Figure 1. Research Framework

The CICIDS2017 dataset was utilized in this study to evaluate the proposed intrusion detection model. The dataset was developed by the Canadian Institute for Cybersecurity (CIC) and contains both normal and malicious network traffic records. Several attack categories are included in the dataset, such as Distributed Denial of Service (DDoS), PortScan, Botnet, Brute Force, and Web Attacks [6]. The dataset contains network flow features extracted using CICFlowMeter, including flow duration, packet length statistics, packet rate, and traffic volume characteristics. These features provide valuable information for distinguishing malicious traffic from normal network activities.

Data preprocessing was conducted to improve data quality and prepare the dataset for machine learning analysis. The preprocessing procedure consisted of the following stages:

1. Data Cleaning Missing values, duplicate records, and invalid observations were identified and removed to ensure data consistency.
2. Label Encoding Class labels were converted into numerical representations suitable for machine learning algorithms.

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

3. Dataset Splitting The processed dataset was divided into training and testing subsets using an 80:20 ratio.

Random Forest was employed as the primary classification algorithm due to its robustness, high predictive performance, and resistance to overfitting [7]. Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their outputs through majority voting.

$$\hat{y} = mode(h_1(x), h_2(x), \ldots, h_n(x)) \quad (1)$$

The Random Forest model was trained using the preprocessed CICIDS2017 dataset and implemented using the Scikit-Learn library in Python. The trained model was subsequently evaluated using the testing dataset to assess its effectiveness in distinguishing malicious network traffic from normal activities.

The performance of the proposed intrusion detection model was evaluated using four widely adopted classification metrics, namely Accuracy, Precision, Recall, and F1-Score. These metrics were selected because they provide a comprehensive assessment of the model's ability to correctly classify network traffic instances [5].

Accuracy measures the proportion of correctly classified instances among all observations and is calculated using Equation (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Precision evaluates the proportion of correctly predicted attack instances among all instances predicted as attacks and is defined in Equation (3).

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall measures the ability of the model to correctly identify actual attack instances and is calculated using Equation (4).

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1-Score represents the harmonic mean of Precision and Recall, providing a balanced measure of classification performance, as shown in Equation (5).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

where TP (True Positive) denotes the number of attack instances correctly classified as attacks, TN (True Negative) represents the number of normal instances correctly classified as normal traffic, FP (False Positive) indicates normal traffic incorrectly classified as attacks, and FN (False Negative) refers to attack instances incorrectly classified as normal traffic.

To improve model interpretability, SHAP (SHapley Additive exPlanations) was applied to explain the predictions generated by the Random Forest classifier [12]. SHAP computes feature contribution scores based on Shapley values derived from cooperative game theory. The SHAP value for each feature indicates how much the feature contributes to

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

increasing or decreasing the prediction outcome. Two levels of interpretation were generated:

1. Global Interpretation, Global SHAP analysis was used to identify the most influential features affecting the overall behavior of the intrusion detection model.
2. Local Interpretation, Local SHAP analysis was used to explain individual prediction outcomes by showing how specific network traffic features contributed to the classification decision.

The SHAP analysis provides transparency into the model decision-making process and enables cybersecurity analysts to better understand the factors influencing attack detection outcomes [12]-[13].

3. RESULT AND DISCUSSIONS

The CICIDS2017 dataset was utilized to evaluate the proposed explainable intrusion detection framework. The selected subset, Friday-WorkingHours-Afternoon-DDos, contains both benign and DDoS network traffic instances. After preprocessing, the dataset consisted of 225,745 network flow records, including 128,027 DDoS samples and 97,718 benign samples.

Table 2. Distribution of Network Traffic Classes

Class	Number of Samples
DDoS	128,027
BENIGN	97,718
Total	225,745

The dataset distribution indicates that both attack and normal traffic were sufficiently represented, enabling the machine learning model to learn discriminative patterns for intrusion detection.

The Random Forest classifier was trained and evaluated using the preprocessed CICIDS2017 dataset. The performance was assessed using Accuracy, Precision, Recall, and F1-Score. The obtained results are presented in Table 3.

Table 3. Classification Performance of the Random Forest Model

Metric	Value (%)
Accuracy	99.9889
Precision	99.9922
Recall	99.9883
F1-Score	99.9902

The exceptionally high classification performance obtained in this study can be attributed to several factors. First, the CICIDS2017 DDoS dataset contains distinctive traffic characteristics that clearly differentiate attack traffic from benign network flows. Features such as packet length statistics, flow duration, and packet transmission behavior exhibit substantial differences between malicious and normal traffic.

Second, the Random Forest algorithm effectively captures complex nonlinear relationships among network traffic features through its ensemble learning mechanism. By combining multiple decision trees, Random Forest reduces overfitting while maintaining strong generalization capability.

Third, the selected dataset focuses specifically on DDoS attack scenarios, which generate highly recognizable traffic patterns compared with more diverse multi-class

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

intrusion detection problems. Consequently, the classifier can more effectively learn discriminative patterns associated with malicious activities, resulting in extremely high accuracy, precision, recall, and F1-score values.

The experimental results demonstrate that the Random Forest classifier achieved excellent performance in distinguishing malicious network traffic from normal activities. The model attained an accuracy of 99.9889%, indicating that nearly all network traffic instances were correctly classified. Furthermore, the precision score of 99.9922% suggests that the number of false positive classifications was extremely low. The recall value of 99.9883% indicates that the model successfully identified almost all attack instances present in the testing dataset. In addition, the F1-Score of 99.9902% reflects a strong balance between precision and recall, demonstrating the reliability of the proposed intrusion detection model.

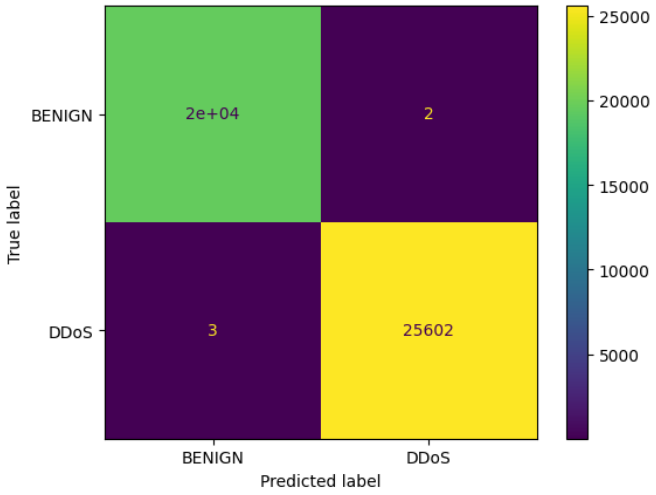

Figure 2. Confusion Matrix of Random Forest Classification

The confusion matrix analysis further confirms the effectiveness of the model. Only two benign traffic instances were incorrectly classified as attacks, while three DDoS instances were misclassified as benign traffic. These findings indicate that the Random Forest classifier provides highly reliable intrusion detection performance for DDoS attack identification.

Feature importance analysis was conducted to identify the most influential network traffic characteristics contributing to the classification process. The top ten features ranked by importance are presented in Table 4.

Table 4. Top Features Contributing to Intrusion Detection

Rank	Feature	Importance
1	Avg Fwd Segment Size	0.0774
2	Init_Win_bytes_forward	0.0756
3	Fwd Packet Length Max	0.0736
4	Fwd Packet Length Mean	0.0730
5	Subflow Fwd Packets	0.0569
6	Subflow Fwd Bytes	0.0563
7	Total Length of Fwd Packets	0.0496
8	Bwd Packet Length Min	0.0475
9	Destination Port	0.0453
10	act_data_pkt_fwd	0.0394

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

The results indicate that packet size characteristics, forward packet statistics, and flow-level traffic measurements significantly influence intrusion detection decisions. These features effectively capture behavioral differences between normal and DDoS traffic patterns, enabling the Random Forest model to achieve high classification accuracy.

To improve the interpretability of the proposed intrusion detection model, SHAP (SHapley Additive exPlanations) was employed to analyze the contribution of individual features to classification outcomes. SHAP provides feature-level explanations by assigning contribution scores that quantify how each feature influences the prediction generated by the Random Forest classifier.

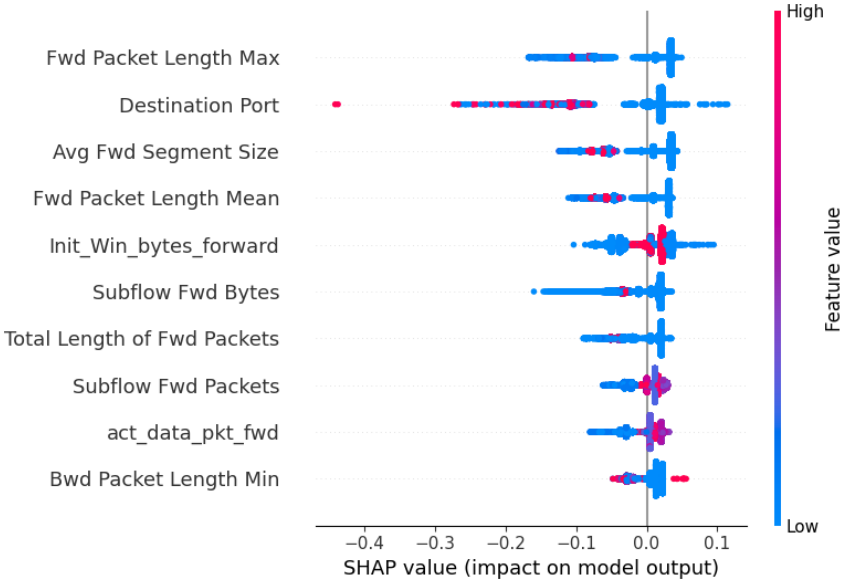

Figure 3. SHAP Summary Plot of Feature Contributions

Table 5. Top Features Based on SHAP Importance

Rank	Feature	SHAP Importance
1	Fwd Packet Length Max	0.046821
2	Destination Port	0.040170
3	Avg Fwd Segment Size	0.039894
4	Fwd Packet Length Mean	0.035248
5	Init_Win_bytes_forward	0.029374
6	Subflow Fwd Bytes	0.025861
7	Total Length of Fwd Packets	0.025323
8	Subflow Fwd Packets	0.017534
9	act_data_pkt_fwd	0.016560
10	Bwd Packet Length Min	0.016413

The experimental results demonstrate that the proposed Random Forest-based intrusion detection model achieved outstanding classification performance on the CICIDS2017 DDoS dataset. The model obtained an accuracy of 99.9889%, precision of 99.9922%, recall of 99.9883%, and F1-score of 99.9902%. These results indicate that the model was capable of accurately distinguishing malicious traffic from normal network activities while maintaining a very low misclassification rate. The confusion matrix analysis further confirmed the effectiveness of the proposed approach. Only a small number of network

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

traffic instances were incorrectly classified, indicating that the Random Forest algorithm successfully learned discriminative traffic patterns associated with DDoS attacks. The ensemble learning mechanism and random feature selection strategy contributed to the model's strong generalization capability and robustness. Furthermore, SHAP analysis provided valuable insights into the internal decision-making process of the classifier. The results revealed that packet-related features such as *Fwd Packet Length Max*, *Destination Port*, and *Avg Fwd Segment Size* played dominant roles in attack classification. These findings are consistent with the characteristics of DDoS attacks, which typically generate abnormal packet transmission behavior and target specific network services. The integration of SHAP into the intrusion detection framework significantly improved model interpretability. By identifying the contribution of individual features, SHAP enables cybersecurity analysts to understand why a particular network traffic instance is classified as malicious. Consequently, the proposed explainable machine learning framework offers not only high predictive performance but also enhanced transparency and trustworthiness for real-world cybersecurity applications.

Practical Implications of SHAP for Cybersecurity Operations

Beyond improving model interpretability, SHAP provides practical benefits for cybersecurity professionals. For Security Operations Center (SOC) analysts, SHAP explanations help clarify why a particular network flow is classified as malicious. Instead of relying solely on prediction labels, analysts can identify which network traffic attributes contribute most significantly to the detection decision, thereby improving trust in automated intrusion detection systems.

For network administrators, the identified influential features can support proactive security monitoring. For example, the importance of Destination Port indicates that specific network services may be more frequently targeted during DDoS attacks. Monitoring these services can help administrators strengthen defensive configurations and reduce potential attack surfaces.

For incident response teams, SHAP explanations provide valuable contextual information during forensic investigations. Features such as Fwd Packet Length Max, Avg Fwd Segment Size, and Subflow Fwd Bytes can be used to understand attack behavior and prioritize suspicious traffic patterns. As a result, the explainable intrusion detection framework can accelerate incident analysis and improve operational decision-making.

Therefore, the integration of SHAP into the intrusion detection process not only enhances model transparency but also provides actionable insights for real-world cybersecurity operations.

The SHAP analysis revealed that packet-related characteristics were the most influential factors affecting intrusion detection decisions. The feature Fwd Packet Length Max achieved the highest importance score, indicating that the maximum size of forward packets significantly contributed to distinguishing DDoS traffic from benign network flows. Similarly, Destination Port demonstrated substantial influence, suggesting that attack traffic tends to target specific service ports more frequently than normal traffic.

In addition, features related to packet size distribution, including Avg Fwd Segment Size and Fwd Packet Length Mean, were identified as critical indicators for attack classification. These findings suggest that DDoS traffic exhibits distinct packet transmission patterns that can be effectively captured by the Random Forest model. Furthermore, flow-based characteristics such as Subflow Fwd Bytes and Subflow Fwd Packets also contributed significantly to classification performance, reflecting differences in traffic volume between normal and malicious network activities.

The SHAP-based interpretation enhances the transparency of the intrusion detection system by providing meaningful explanations for model predictions. Unlike conventional

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

black-box machine learning approaches, the proposed framework allows security analysts to identify which network traffic attributes contribute most significantly to attack detection decisions. Therefore, SHAP not only improves model interpretability but also increases trustworthiness and practical applicability in cybersecurity environments.

Local SHAP Explanation for a DDoS Sample

To complement the global SHAP analysis, a local explanation was conducted to examine how individual network traffic features contributed to a specific DDoS prediction. Local SHAP explanations provide instance-level interpretability by quantifying the contribution of each feature to the final classification outcome.

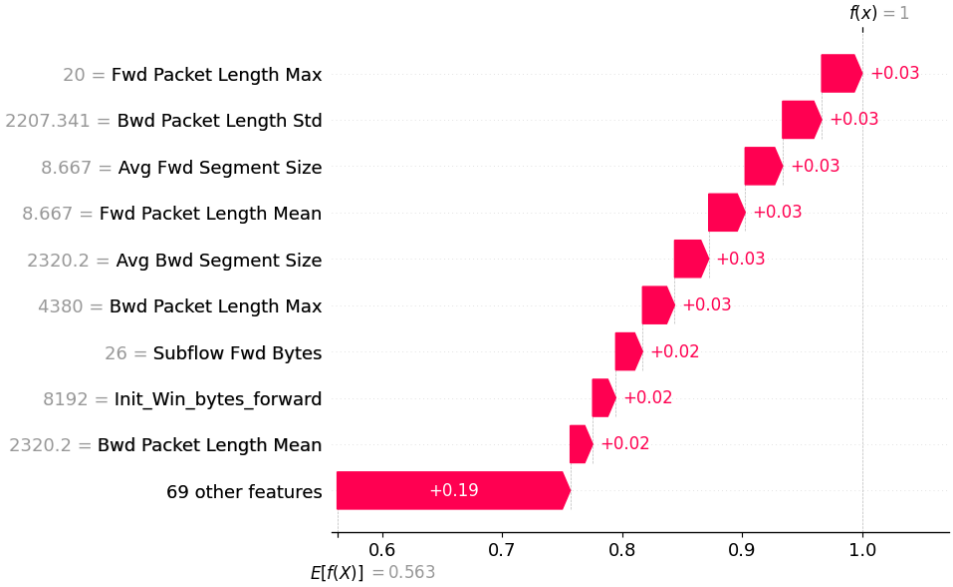

Figure 4. Local SHAP Explanation for a DDoS Sample Using SHAP Waterfall Plot

The local SHAP explanation presented in Figure 4 illustrates how individual network traffic features contributed to the classification of a specific DDoS sample. The waterfall plot shows that several packet-related characteristics positively influenced the prediction outcome, including Fwd Packet Length Max, Bwd Packet Length Std, Avg Fwd Segment Size, and Fwd Packet Length Mean. Each of these features increased the prediction score toward the attack class by approximately 0.02–0.03 SHAP units.

The baseline prediction value of the model was 0.563. After considering the contribution of all relevant features, the prediction score increased to 1.0, indicating a very high confidence level that the analyzed network flow represented a DDoS attack. The results demonstrate that packet size statistics and traffic volume characteristics played dominant roles in the attack classification process.

This local explanation complements the global SHAP analysis by providing instance-level interpretability. While the global analysis identifies generally important features across the entire dataset, the local explanation reveals how specific feature values influence an individual prediction. Such information can assist cybersecurity analysts in validating intrusion detection alerts and understanding the reasoning behind automated attack detection decisions.

The findings of this study are consistent with previous research demonstrating the effectiveness of Random Forest and explainable artificial intelligence techniques for intrusion detection. Compared with Le et al [17], who employed ensemble tree models

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

with SHAP explanations, the proposed framework achieved comparable interpretability while maintaining excellent detection performance. Furthermore, unlike previous studies that primarily focused on feature importance ranking, this study extends the discussion by highlighting the practical implications of SHAP explanations for SOC analysts, network administrators, and incident response teams.

The integration of SHAP into the intrusion detection process enables cybersecurity professionals to understand not only which features are important but also how these features contribute to attack detection decisions. Consequently, the proposed approach provides both predictive performance and operational transparency, which are critical requirements for trustworthy cybersecurity systems.

4. CONCLUSION

This study proposed an explainable machine learning framework for network intrusion detection using the Random Forest algorithm and SHAP-based feature interpretation. The CICIDS2017 Friday-WorkingHours-Afternoon-DDos dataset was utilized to evaluate the effectiveness of the proposed approach in detecting malicious network traffic.

Experimental results demonstrated that the Random Forest classifier achieved excellent performance, obtaining an accuracy of 99.9889%, precision of 99.9922%, recall of 99.9883%, and F1-score of 99.9902%. These results indicate that the proposed model was highly effective in distinguishing DDoS attacks from benign network traffic while maintaining a very low misclassification rate.

Furthermore, SHAP analysis enhanced the interpretability of the intrusion detection model by identifying the contribution of individual network traffic features to classification decisions. The results revealed that Fwd Packet Length Max, Destination Port, Avg Fwd Segment Size, and Fwd Packet Length Mean were among the most influential features affecting intrusion detection outcomes. The SHAP-based explanations provided valuable insights into the internal decision-making process of the Random Forest classifier and improved the transparency of attack detection results.

Overall, the integration of Random Forest and SHAP offers a practical and trustworthy solution for network intrusion detection by combining high predictive performance with model interpretability. Future research may explore the application of advanced ensemble learning and deep learning techniques combined with explainable artificial intelligence methods to improve intrusion detection performance across more diverse cyberattack scenarios.

5. REFERENCES

[1] P. W. Singer, A. Friedman, P. W. Singer, and A. Friedman, *Cybersecurity and Cyberwar: What Everyone Needs to Know*®. in What Everyone Needs To Know®. Oxford, New York: Oxford University Press, 2014.
[2] "ENISA Threat Landscape 2023 | ENISA." Accessed: Jun. 03, 2026. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023
[3] "Cost of a data breach 2025 | IBM." Accessed: Jun. 03, 2026. [Online]. Available: https://www.ibm.com/reports/data-breach
[4] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *2010 IEEE Symposium on Security and Privacy*, Oakland, CA, USA: IEEE, 2010, pp. 305–316. doi: 10.1109/SP.2010.25.
[5] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Commun. Surv. Tutor.*, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

Dery Yuswanto Jaya: *Coresponding Author

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, July 2026
DOI : **https://doi.org/10.58602/chain.v4i3.283**
Page 206-217

[6] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization:," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018, pp. 108–116. doi: 10.5220/0006639801080116.
[7] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[8] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*, Fourth edition. Cambridge, MA, United States: Morgan Kaufmann Publishers, an imprint of Elsevier, 2023. doi: 10.1016/C2013-0-18660-6.
[9] Z. C. Lipton, "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018, doi: 10.1145/3236386.3241340.
[10] D. Gunning and D. W. Aha, "DARPA's Explainable Artificial Intelligence Program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019, doi: 10.1609/aimag.v40i2.2850.
[11] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
[12] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017, *arXiv*. doi: 10.48550/ARXIV.1705.07874.
[13] S. Patil *et al.*, "Explainable Artificial Intelligence for Intrusion Detection System," *Electronics*, vol. 11, no. 19, p. 3079, Sep. 2022, doi: 10.3390/electronics11193079.
[14] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," *IEEE Access*, vol. 10, pp. 93575–93600, 2022, doi: 10.1109/ACCESS.2022.3204171.
[15] G. Rjoub *et al.*, "A Survey on Explainable Artificial Intelligence for Cybersecurity," *IEEE Trans. Netw. Serv. Manag.*, vol. 20, no. 4, pp. 5115–5140, Dec. 2023, doi: 10.1109/TNSM.2023.3282740.
[16] S. Neupane *et al.*, "Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities," 2022, *arXiv*. doi: 10.48550/ARXIV.2207.06236.
[17] T.-T.-H. Le, H. Kim, H. Kang, and H. Kim, "Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method," *Sensors*, vol. 22, no. 3, p. 1154, Feb. 2022, doi: 10.3390/s22031154.
[18] I. C. Obagbuwa, M. N. Ngafeeson, O. F. Obagbuwa, and A. Tsetse, "Machine Learning and Explainable Artificial Intelligence for Network Intrusion Detection:," *Int. J. Inf. Secur. Priv.*, vol. 20, no. 1, pp. 1–22, Feb. 2026, doi: 10.4018/IJISP.402900.

Dery Yuswanto Jaya: *Coresponding Author