Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30154

Research on DDoS Attack Detection Method Fusing Dynamic Feature Selection and Heterogeneous Model Ensembling

Haotong Zhang[1], Yufei Ye[2*†], Yaoyang Zhang[3†]

[1]*College of Information Engineering, Zhejiang University of Technology, Hangzhou, China*
[2]*College of Mechanical Engineering, Zhejiang University, Hangzhou, China*
[3]*Dulwich International High School Suzhou, Suzhou, China*
**Corresponding Author. Email: 3230102886@zju.edu.cn*
[†]*These authors contributed equally to this work and should be considered as co-second authors.*

Abstract. As distributed denial-of-service (DDoS) attacks and their derivatives grow increasingly complex and diverse, traditional approaches relying on single global feature selection and single-model classifiers often struggle to balance accuracy in identifying hybrid attacks with training and inference efficiency. This paper proposes a unified framework of "Dynamic Feature Selection and Multi-Model Ensemble," systematically validated on the CICDDoS2019 benchmark dataset. Specifically, we first use XGBoost as the baseline, compute global feature importance via SHAP, and construct a Top-N dynamic feature subset based on this to quantitatively analyse the trade-off between accuracy and training time. Second, to optimize accuracy, we design differentiated feature engineering and data processing workflows for different models: applying standardization and PCA dimensionality reduction for tree models (XGBoost, Random Forest), introducing port range binning, protocol indicators, and TCP flag-derived features for neural networks (MLP), combined with low-cardinality one-hot encoding and high-cardinality frequency encoding; finally, fusing the probability outputs of three models via soft voting. Experiments demonstrate that retaining only the top-20 most important features achieves a system accuracy of 69.10%, comparable to or slightly exceeding the full-feature baseline while reducing training time to 6.9 seconds—approximately a 47% acceleration over the baseline. In terms of accuracy, the three-model ensemble further improves upon the baseline XGBoost by about 1 percentage point. SHAP-based interpretability analysis indicates that approximately 20 features represent the "inflection point" between performance and efficiency. Top-ranked features are primarily correlated with source/destination ports, packet counts in forward/reverse flows, bytes, and frequency—consistent with the traffic mechanisms of typical amplified reflection-based DDoS attacks. This paper contributes a reusable dual-objective evaluation framework for speed and accuracy, proposes differentiated feature engineering strategies for heterogeneous models, and significantly reduces training overhead while steadily improving detection accuracy—all while maintaining interpretability.

Keywords: DDoS Detection, Feature Selection, SHAP, Ensemble Learning, Feature Engineering, CICDDoS2019

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ30154

1. Introduction

Adversarial activities in cyberspace are evolving from single, static attack modes to hybrid, phased, and confrontational scenarios. Existing response methods for large-scale Distributed Denial of Service (DDoS) attacks rely heavily on fixed feature sets and single classifiers, making it difficult to adapt to significant differences in feature distribution among different attack groups [1]. This results in insufficient recall and increased false positive rates in hybrid attack scenarios. Meanwhile, due to training costs and engineering complexity, many methods overlook ensemble learning or differentiated feature engineering, thereby limiting upper-bound performance and robustness. This paper aims to address the core challenge: how to significantly reduce model training and inference time while maintaining or even improving detection accuracy, ensuring practical feasibility and cost-effective deployment. To solve this problem, we propose combining "dynamic feature selection" with "multi-model differentiated engineering + ensemble": while preserving interpretability, SHAP scores are used to identify a set of most contributory features to reduce dimensionality and computational overhead. Simultaneously, customized feature engineering pipelines are constructed for tree-based models and neural networks respectively, and their respective advantages are integrated through soft voting. Our key findings indicate that retaining only approximately 20 key features on the CICDDoS2019 dataset maintains or slightly improves accuracy compared to the full-feature baseline, while reducing training time by nearly half. Probabilistic ensemble of the three models further enhances accuracy. This demonstrates that in typical high-dimensional traffic feature spaces, reasonable feature dimensionality reduction and differentiated engineering can effectively advance the goal of unifying speed and accuracy.

2. Related work

The CICDDoS2019 dataset, constructed by the Canadian Institute for Cybersecurity, follows a rigorous data collection and annotation process. Combining rich feature dimensions with broad attack coverage, it has established a recognized standard in academia for evaluating DDoS detection algorithms [2]. In recent years, interpretable machine learning methods have been widely applied in security scenarios for feature selection and evidence auditing. Among them, SHAP provides global and local interpretations for tree models and certain neural networks through unified game-theoretic decomposition. This makes it suitable for measuring feature contributions in Intrusion Detection System (IDS) scenarios, thereby guiding the selection of feature subsets. Additionally, ensemble learning enhances generalization ability and stability by aggregating multiple weakly correlated classifiers [3,4]. Comprehensive studies have shown that fusion methods generally outperform independent models and exhibit particular robustness in cases of class imbalance or data distribution drift [5]. Different from previous studies, this paper combines SHAP-based dynamic feature selection with heterogeneous engineering for different models and probabilistic soft voting on top of a mature benchmark, forming a unified and operable paradigm for balancing speed and accuracy.

3. Dataset and preprocessing

This paper uses benign data and multi-class DDoS/DrDoS attack samples contained in the CICDDoS2019 dataset, including but not limited to DNS, LDAP, MSSQL, NTP, NetBIOS, SNMP, UDP, Portmap, Syn, TFTP, UDPLag, and WebDDoS. To mitigate class imbalance and control training scale, we sample a maximum of 13,000 records per class: classes exceeding the upper limit adopt a subsampling strategy with a fixed random seed, while classes below the upper limit retain all

DOI: 10.54254/2755-2721/2026.TJ30154

records. During data cleaning, all parsable fields are converted to numerical types as much as possible. Infinite values are mapped to missing values and filled with zeros, and outliers are suppressed through threshold clipping to reduce abnormal noise. Constant columns are removed before training to ensure each feature carries meaningful information across samples. In the data exploration phase, various visualization methods—including scatter plots, box plots, time series plots, correlation heatmaps, violin plots, and port distribution pie charts—are used to comprehensively evaluate traffic duration, forward/reverse packet and byte counts, temporal evolution, and port distribution. This confirms the overall distribution and potential abnormal patterns, laying the foundation for subsequent feature engineering. For the multi-class classification task, we construct a mapping dictionary from filenames to numerical labels according to the above strategy, and merge samples of all classes into a unified clean_data.csv file for subsequent experiments.

4. Methodology

This paper's methodology comprises two main approaches: Experiment 1: SHAP-driven dynamic feature selection focuses on accelerating model training. Experiment 2: Multi-model differentiated feature engineering combined with ensemble learning emphasizes improving model training accuracy. Both approaches can be evaluated independently or integrated for final system construction.

4.1. Experiment 1

In dynamic feature selection, we use XGBoost training with all features as the baseline, recording baseline accuracy and training time. Subsequently, we compute SHAP values for the training set using TreeExplainer. To accommodate the different data structures SHAP may return for multi-class classification, we perform compatibility processing. We then use the average absolute SHAP value across sample dimensions as the global feature importance metric, obtaining feature rankings in descending order. Based on this ranking, we constructed Top-N subsets with $N \in \{5, 10, 20, 30, 40, 50\}$. We retrained XGBoost on each subset, systematically comparing training time and test set accuracy to identify the minimal sufficient feature set. We plotted curves showing accuracy and time as a function of N. The flowchart for Experiment 1 is shown in Figure 1 below.

4.2. Experiment 2

In multi-model differentiation engineering and ensemble training, we perform customized processing to address structural differences between tree models and neural networks. For tree models (XGBoost, Random Forest), we removed strongly discrete or non-numeric attributes (e.g., timestamps, raw protocol enumerations, raw flag count categorical columns), retained continuous numerical variables, standardized the retained variables, and then reduced dimensions to 30 principal components using PCA. Although tree models theoretically do not require standardization, PCA effectively reduces redundancy and improves training efficiency in high-dimensional, strongly correlated count and rate feature spaces, while having a limited impact on accuracy in empirical tests. XGBoost employs mlogloss as the multi-class evaluation metric, disables legacy label encoding, and sets a fixed random seed. Random Forest controls variance and training time by limiting tree depth, minimum node split samples, and feature sampling ratios. For neural networks (MLP), we introduce structured feature engineering oriented toward network traffic. Specifically: -

DOI: 10.54254/2755-2721/2026.TJ30154

Source and destination ports are bucketed by system ports, registered ports, and dynamic ports to derive port segment indicator variables. - Derived features such as total_flags, flag_diversity, syn_ack_ratio, abnormal_termination, bidirectional_psh, and bidirectional_urg are constructed from TCP flags and bidirectional flag combinations to capture connection establishment and abnormal termination behaviors. Constructing indicator variables is_tcp, is_udp, and is_icmp for protocol types. Subsequently, low-cardinality discrete variables undergo one-hot encoding, high-cardinality categorical variables use frequency encoding to suppress dimensionality explosion, and remaining numerical variables are standardized. This yields an MLP input matrix with consistent scaling and high information density. To balance fitting capability and stability, we employed a three-hidden-layer ReLU network (128-64-32) with an adaptive learning rate and early stopping strategy, incorporating an internal validation set to prevent overfitting. After training each of the three models separately, we perform ensemble prediction using probabilistic soft voting. This involves taking the arithmetic mean of the respective category posterior probabilities and selecting the category with the highest probability as the final prediction. This fusion approach enhances robustness when category boundaries are ambiguous or sample heterogeneity is pronounced. It is concluded that the boundary conditions at the ends of the specimens have a great effect on the static mechanical properties of the composite slabs presented here, and the composite slab with restraint at four corners has the largest bearing capacity. The flowchart for Experiment 2 is shown in Figure 2 below.

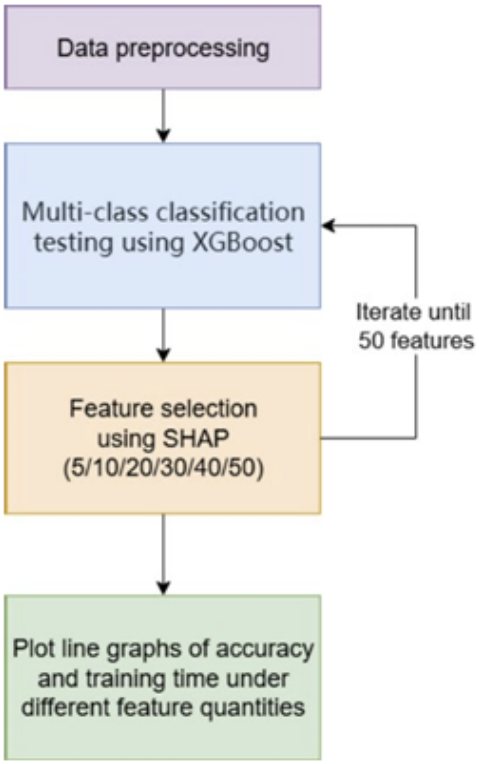


Figure 1. Flowchart for experiment 1

DOI: 10.54254/2755-2721/2026.TJ30154

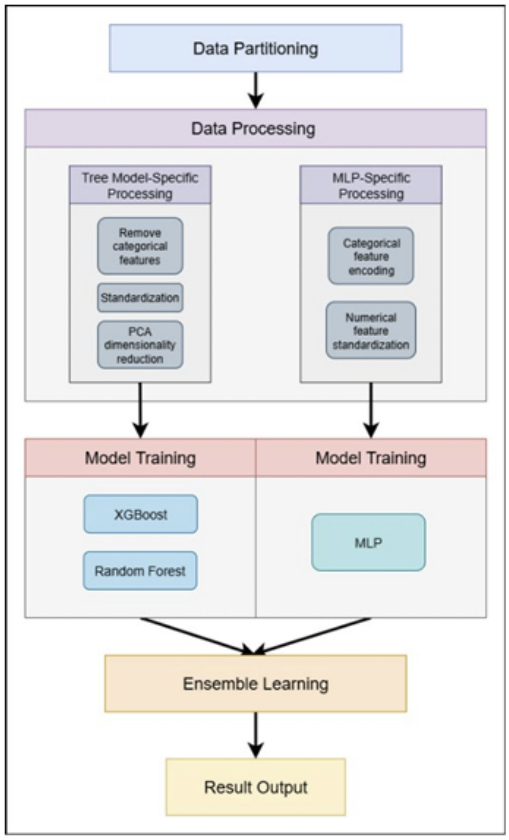


Figure 2. Flowchart for experiment 2

5. Experimental setup

We used the merged and cleaned clean_data.csv as the data source, employing a fixed random seed to split the dataset into 80% training and 20% testing. For metrics, we report the overall accuracy and export a complete classification report, including precision, recall, and F1 scores for each category to facilitate subsequent analysis. For dynamic feature selection experiments, we additionally output graphs showing accuracy versus feature count and training time versus feature count. Based on SHAP global importance, we generate a full feature ranking and a Top-20 bar chart. All experiments use the same random seed to minimize variance, and model evaluation results are exported as CSV files for subsequent plotting and comparison.

6. Experiment results

6.1. Experiment 1 (acceleration experiment)

XGBoost achieved baseline accuracy and training time using all features. Subsequently, we retrained the model on the Top-N subsets ranked by SHAP. Results showed that using only the top 20 most important features achieved a test accuracy of 69.10%, which was comparable to or slightly better than the full-feature baseline, while reducing training time to just 6.9 seconds—approximately 47% faster than the baseline. When N was below 20 (e.g., Top-10), significant loss of key discriminative information led to a noticeable drop in accuracy. When N exceeded 20 (e.g., Top-30 or Top-50), accuracy improvements were negligible or even slightly declined, while training time continued to increase. This indicates that approximately 20 features form an efficiency-effectiveness inflection

DOI: 10.54254/2755-2721/2026.TJ30154

point, maximally reducing training overhead while preserving primary discriminative signals. We simultaneously provide two curves illustrating accuracy and training time versus N, along with visualizations of SHAP global importance (full and Top-20), intuitively presenting this trend. The Time-Num graph is shown in Figure 3 below, and the Accuracy-Num graph is shown in Figure 4 below.

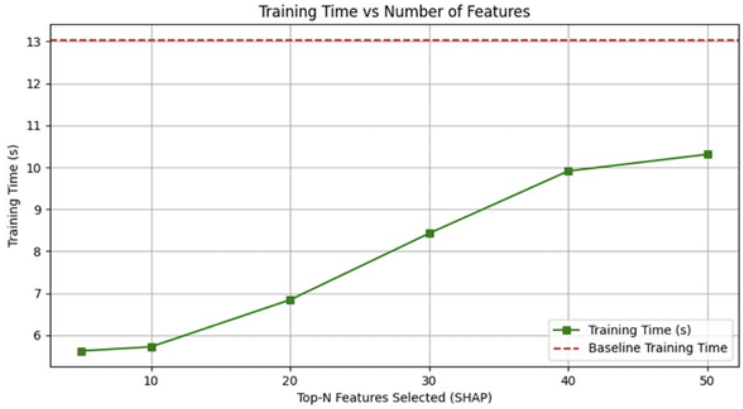


Figure 3. Time-Num graph

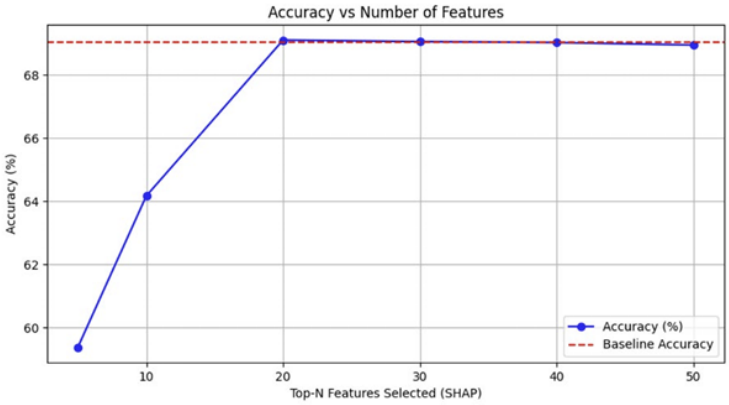


Figure 4. Accuracy-Num graph

6.2. Experiment 2 (quality enhancement)

These three models were trained and evaluated under their respective distinct feature engineering workflows. XGBoost and Random Forest learned from inputs standardized and compressed to 30 principal components via Principal Component Analysis (PCA), both demonstrating excellent overall stability and training efficiency. XGBoost achieved an overall accuracy of 66.31%, while Random Forest reached 65.32%. The Multilayer Perceptron (MLP) was trained using structured features including port range binning, protocol metrics, TCP flag derivation, and bidirectional statistics. However, its current accuracy, at 63.96%, is still slightly lower than that of the tree-based models, indicating room for further hyperparameter adjustment. Through a soft voting ensemble that averages the probabilistic outputs of the three models with equal weights, the overall accuracy improved by approximately 1 percentage point compared to the XGBoost baseline, reaching 67.28%. This result is consistent with conclusions from existing research that ensemble learning outperforms single models in both accuracy and stability. It also indicates that engineering methods

DOI: 10.54254/2755-2721/2026.TJ30154

such as dimensionality reduction and feature selection can effectively mitigate the additional training overhead associated with ensemble methods.

7. Explainability analysis

We ranked features based on SHAP global feature importance. The SHAP global feature importance (all features) ranking bar chart is shown in Figure 5 below. As seen, features beyond the top 20 exhibit SHAP values approaching zero. Therefore, we again selected the top 20 features for analysis. The SHAP global feature importance (top-20) ranking bar chart is shown in Figure 6 below.

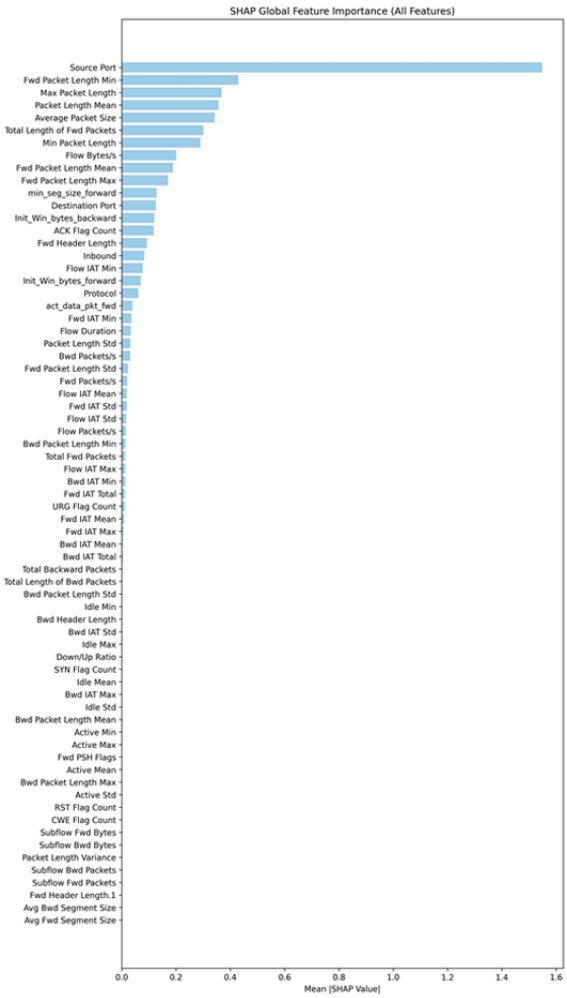


Figure 5. SHAP global feature importance (all features) ranking bar chart

DOI: 10.54254/2755-2721/2026.TJ30154

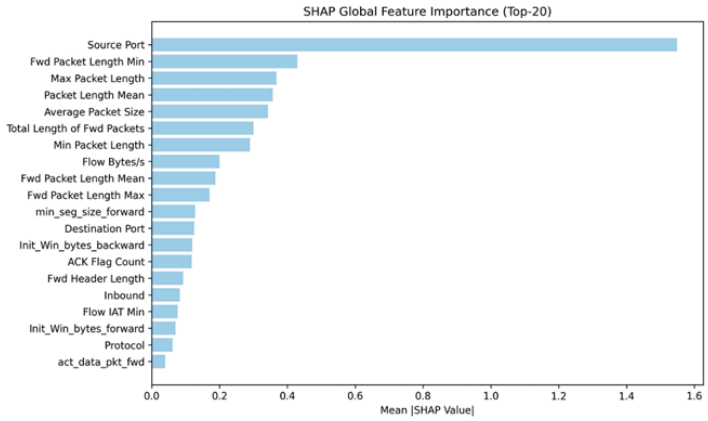


Figure 6. SHAP global feature importance (top-20) ranking bar chart

Top-ranked features are port statistics, packet counts, bytes, and frequency for all traffic. The "source port" feature has the highest importance. Amplified reflection-based DDoS attacks use small queries to create large server responses. A characteristic pattern emerges: minimal forward traffic and high-volume reverse traffic. These attacks aim at specific service ports. Port features and traffic statistics are critical evidence for detection. TCP behaviour features like syn_ack_ratio capture abnormal connection patterns. They help distinguish anomalous traffic from normal traffic. Feature importance decreases sharply after the top 20. More features add little accuracy and may introduce noise. Using only the top 10 features removes key signals. This results in a large drop in accuracy. For the CICDDoS2019 task, about 20 features form the minimal sufficient set. This set holds the main decision signals. It is the inflection point between performance and efficiency.

8. Conclusions

This paper proposes and validates a unified framework for network attack detection, combining dynamic feature selection, multi-model differentiation engineering, and soft voting integration. On the CICDDoS2019 dataset, feature pruning based on SHAP enables the use of only the top-20 features to maintain or even slightly improve accuracy while reducing training time by approximately 47%. Building upon this foundation, probabilistic ensemble integration of XGBoost, Random Forest, and MLP further improved accuracy by approximately 1 percentage point compared to the baseline. The methodological core lies in leveraging interpretable feature importance to guide feature subset construction, coupled with customized feature engineering and preprocessing tailored to each model's structural characteristics. This approach achieves overall gains in both speed and accuracy without compromising interpretability. Future work will proceed along two main lines: On one hand, further expanding the dataset scale and refining the Top-N value grid (e.g., 15, 25, 35), while reporting inference latency and throughput to evaluate end-to-end performance; Second, we will conduct more systematic hyperparameter search and regularization design for MLP, explore stronger fusion methods (such as Stacking and probabilistic calibration based on meta-learners), and investigate adaptive feature subset selection by attack family and dynamic switching mechanisms for online scenarios to continuously enhance the system's accuracy, robustness, and deployment efficiency.

DOI: 10.54254/2755-2721/2026.TJ30154

Acknowledgments

Haotong Zhang is the sole first author with the largest contribution to this work, Yufei Ye is the second and corresponding author, and Yaoyang Zhang is a co-second author.

References

[1] Rashid, S. M. K.-U.-R. Raazi, and S. Ali, "Detection and mitigation of DDoS attacks on IoT networks using machine learning, " in 2024 26th International Multi-Topic Conference (INMIC), Karachi, Pakistan, 2024, pp. 1-6, doi: 10.1109/INMIC64792.2024.11004398.

[2] K. Saluja, S. Bagchi, V. Solanki, M. N. A. Khan, E. Dhamija, and S. K. Debnath, "Exploring robust DDoS detection: A machine learning analysis with the CICDDoS2019 dataset, " in 2024 IEEE 5th India Council International Subsections Conference (INDISCON), Chandigarh, India, 2024, pp. 1–6, doi: 10.1109/INDISCON62179.2024.10744272.

[3] G. Kunapuli, Ensemble Methods for Machine Learning. Shelter Island, NY: Manning, 2023.

[4] A_Student_2020, "Ensemble Learning Methods: A Practical Guide, " CSDN Blog, May 2024. [Online]. Available: https: //blog.csdn.net/a_student_2020/article/details/139784117.

[5] S. Abimannan, E.-S. M. El-Alfy, Y.-S. Chang, S. Hussain, S. Shukla, and D. Satheesh, "Ensemble multifeatured deep learning models and applications: A survey, " IEEE Access, vol. 11, pp. 107194–107217, 2023, doi: 10.1109/ACCESS.2023.3320042.