

A Federated Heterogeneous Ensemble Framework for Detecting Low-Rate Distributed Denial-of-Service Attacks

Rikie Kartadie   [Department of Computer Engineering, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia | riki@utdi.ac.id]

Danny Kriestanto  [Department of Informatics Engineering, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia | danny@utdi.ac.id]

Yagus Cahyadi  [Department Master of Information Technology, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia | yagus.cahyadi@utdi.ac.id]

 Department of Computer Engineering, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia,
Jl. Raya Janti No. 143, Yogyakarta, 55198, indonesia

Received: 17 March 2026 • Accepted: 15 August 2026 • Published: 01 September 2026

Abstract. Detecting low-rate distributed denial-of-service (LR-DDoS) attacks remains challenging because their traffic profiles closely resemble legitimate network activity, making them difficult to identify using conventional intrusion detection systems. This study proposes a federated heterogeneous ensemble framework for LR-DDoS detection under non-IID data conditions. Five local classifiers, Logistic Regression, Random Forest, SVM, Gradient Boosting, and MLP, are trained on locally partitioned datasets generated through Dirichlet-based distribution, and their probabilistic outputs are combined via a softmax-weighted aggregation mechanism. Experiments using the CICDDoS2019 dataset with a 70/15/15 train-validation-test split and results averaged over 10 independent seeds show that local model accuracy ranges from 0.664 to 0.917 under non-IID conditions. The proposed federated ensemble achieves a mean accuracy of 0.925 ± 0.008 and F1-score of 0.910 ± 0.009 , outperforming all individual models and two alternative aggregation strategies. These results indicate that combining heterogeneous model architectures with adaptive prediction-level aggregation improves detection robustness in distributed intrusion detection systems

Keywords: Federated Learning, Low-Rate Distributed Denial-of-Service Detection, Heterogeneous Model Ensemble, Network Intrusion Detection Systems, Non-IID Data in Distributed Networks

1 Introduction

The fast growth of Internet of Things (IoT) network infrastructures and the rising trend of Software-Defined Networking (SDN) have significantly changed the current state of network architecture. The decoupling of the control plane and the data plane in SDN allows for centralized control and flexible management of the network. Nevertheless, this new network architecture also brings new security risks. For example, the SDN controllers are logically centralized components that manage the network behavior. Therefore, they are very attractive targets for Distributed Denial of Service (DDoS) attacks. If the controllers are overwhelmed or become unavailable, a large part of the network will suffer from serious service disruption. The rapid growth of IoT devices also makes this problem worse because massive botnets consisting of insecure devices have been used increasingly to carry out high-volume DDoS attacks on network infrastructures [Zaidi, 2024]. Therefore, designing efficient methods to detect abnormal traffic in distributed SDN networks has become a pressing issue for current network security.

Machine learning (ML) has been identified as one of the most popular methods for intrusion detection due to its ability to learn complex patterns in network traffic data [Talukder *et al.*, 2024]. Traditional ML-based intrusion detection systems generally follow a centralized approach to data collection and model training, where traffic data observations from

different network domains are consolidated into a common database. Although this method can improve model performance in a controlled setting, it also presents several challenges in practical SDN and IoT network settings. The need for centralized training requires the exchange of large amounts of traffic data, which can be a privacy issue and may also violate regulatory requirements in different administrative domains. Furthermore, centralized models are generally poor at generalizing over different network domains with highly varying traffic patterns. To overcome these challenges, Federated Learning (FL) has been suggested as a collaborative learning approach that allows multiple parties to cooperatively train a model without exchanging raw data [Konečný *et al.*, 2016]. Although FL presents several benefits, most existing federated learning models generally require that the data from different participating clients be approximately independent and identically distributed (IID) and that all clients use the same model architecture during training [G *et al.*, 2024].

In practice, these assumptions rarely hold in operational network environments. In SDN multi-controller architectures, each controller typically monitors traffic within its own administrative domain, resulting in highly heterogeneous and often extremely imbalanced data distributions. Prior research has demonstrated that classical federated optimization algorithms such as FedAvg experience substantial performance degradation when client datasets follow strongly non-IID distributions [G *et al.*, 2024; Doriguzzi-Corin *et al.*, 2025]

non-independent and identically distributed (non-IID), a condition formally defined as one in which the joint data distribution $P(X, y)$ varies across participating clients. Several extensions have been proposed to mitigate this limitation, including FedProx, which introduces a proximal regularization term to constrain local updates [Zhu et al., 2021], FedDyn, which applies dynamic regularization to stabilize optimization [Doriguzzi-Corin et al., 2025], and FedNova, which normalizes local updates to reduce objective inconsistency [Lavaur et al., 2022]. Although these approaches improve convergence stability under moderate heterogeneity, they still rely on parameter sharing and require participating clients to use identical model architectures. Such requirements may limit their applicability in real-world SDN environments, where controllers belonging to different domains often deploy distinct machine learning models optimized for local operational requirements.

To address the limitations of parameter-sharing federated learning, recent studies have explored the use of ensemble learning within federated settings. Ensemble methods combine predictions from multiple models to improve robustness and generalization, particularly when individual models exhibit complementary decision boundaries. In the context of federated learning, output-level aggregation has been shown to offer improved resilience against heterogeneous data distributions by combining predictions rather than model parameters [Allouah et al., 2024; Daalen et al., 2022]. However, existing studies on federated ensemble methods generally assume homogeneous model architectures and rarely evaluate extreme conditions such as highly skewed or single-class client datasets. Such conditions are common in network intrusion detection scenarios, where individual SDN controllers may observe only limited subsets of traffic patterns within their respective domains. As a result, local models may experience severe bias or even model collapse when trained on extremely imbalanced data distributions.

These observations reveal an important gap in the current literature. While previous research has addressed either data heterogeneity or federated ensemble mechanisms, there remains limited investigation into federated learning frameworks that simultaneously accommodate heterogeneous model architectures and non-IID data distributions in network security contexts. In particular, little attention has been given to the potential of federated heterogeneous ensembles that aggregate predictions from diverse local models without relying on parameter sharing.

To address this gap, this study proposes a Federated Heterogeneous Ensemble Learning framework for DDoS detection in SDN multi-controller environments under non-IID conditions. In the proposed framework, each SDN controller trains a local model using its own traffic data and independently optimized model architecture. Instead of exchanging model parameters, controllers submit probabilistic predictions generated on a shared validation dataset. The central coordinator then performs softmax-weighted ensemble aggregation to produce a global decision model that integrates information from heterogeneous participants while preserving data privacy and architectural independence. The objective of this study is to investigate whether heterogeneous ensemble aggregation within a federated learning framework can

improve the robustness and generalization of DDoS detection models under non-IID conditions. Specifically, this study addresses the following research questions: (1) RQ1: How does non-IID data distribution affect the performance of federated intrusion detection models in SDN multi-controller environments?; (2) RQ2: Can heterogeneous model collaboration without parameter sharing improve detection performance compared with individual local models?; (3) RQ3: Does softmax-weighted ensemble aggregation enhance the stability of federated learning when some clients experience severe data imbalance or model collapse?

This study makes three primary contributions. First, it introduces a federated heterogeneous ensemble architecture that enables collaboration among diverse local models without requiring parameter sharing. Second, it evaluates the robustness of the proposed framework under non-IID conditions using the CICDDoS2019 dataset as a representative benchmark for modern DDoS attacks. Third, it demonstrates that prediction-level aggregation through softmax-weighted ensembles can produce a more stable and generalizable global detection model in distributed SDN environments. Collectively, these contributions provide new insights into the design of privacy-preserving and resilient intrusion detection mechanisms for heterogeneous network infrastructures.

The remainder of this paper is organized as follows. Section 2 reviews related work on federated learning for intrusion detection, non-IID challenges, and ensemble approaches. Section 3 describes the proposed methods. Section 4 presents and discusses the experimental results. Section 5 concludes the paper.

2 Related Work

2.1 Federated Learning for Network Intrusion Detection

The application of federated learning to network intrusion detection has gained considerable attention as a privacy-preserving alternative to centralized model training. By enabling multiple network nodes to collaboratively train a shared model without exchanging raw traffic data, federated learning addresses both the privacy concerns and the regulatory constraints that limit centralized approaches in multi-domain network environments. Lavaur et al. [2022] demonstrated that federated learning can serve as an effective enabler for collaborative security between distributed parties that do not fully trust one another, establishing that prediction-level collaboration can maintain detection quality while preserving operational autonomy across administrative boundaries.

Despite these advantages, the performance of federated intrusion detection systems under realistic deployment conditions remains a subject of active investigation. Doriguzzi-Corin et al. [2025] conducted a comparative evaluation of federated learning methods for cybersecurity under non-IID and class-imbalanced settings, reporting that standard federated optimization algorithms degrade noticeably when client data distributions diverge substantially from one another. These findings highlight that the effec-

tiveness of federated intrusion detection is closely tied to the degree of data heterogeneity present across participating clients, motivating the need for aggregation strategies that are robust to distributional differences.

2.2 Non-IID Challenges in Federated Learning

Data heterogeneity is widely recognized as one of the central challenges in federated learning. The non-IID condition arises when the marginal distribution $P(X)$ or the conditional distribution $P(y|X)$ differs across clients, causing locally trained models to diverge in ways that undermine naive parameter averaging. Zhu et al. [2021] provide a comprehensive survey of federated learning under non-IID conditions, documenting how distributional heterogeneity affects convergence stability, model fairness, and generalization performance across a range of federated optimization algorithms.

A standard approach to simulating non-IID conditions in experimental evaluations is the use of the Dirichlet distribution to generate heterogeneous label distributions across clients. Hsu et al. [2019] introduced this methodology in the context of federated visual classification, demonstrating that the Dirichlet concentration parameter α controls the degree of label skew across clients, with smaller values producing more extreme heterogeneity. Yurochkin et al. [2019] extended this line of work by proposing a Bayesian nonparametric approach to federated aggregation that accommodates heterogeneous client distributions without requiring uniform model architectures. Together, these works establish the Dirichlet-based partitioning approach as a principled and widely adopted benchmark methodology for evaluating federated learning under non-IID conditions, which is adopted in the present study with $\alpha = 0.8$.

2.3 Ensemble Learning in Federated Settings

Ensemble learning offers a natural mechanism for aggregating knowledge from multiple heterogeneous sources. Within federated contexts, prediction-level aggregation avoids the need for parameter sharing, enabling clients operating architecturally distinct models to collaborate on a shared detection task without imposing uniform architecture requirements. Daalen et al. [2022] reviewed federated ensemble approaches in the literature, identifying output-level aggregation as particularly suitable for settings where clients employ different model types or possess incompatible data schemas, as is commonly the case in multi-domain network monitoring environments.

More recently, Allouah et al. [2024] revisited ensembling in one-shot federated learning, demonstrating that carefully weighted prediction aggregation can match or exceed multi-round parameter-averaging methods under appropriate conditions. Their findings suggest that the quality of the aggregation weighting mechanism plays a decisive role in determining ensemble performance, particularly when participating models vary substantially in their local data quality or class coverage. Despite this progress, existing federated ensemble studies predominantly assume homogeneous model

architectures and evaluate only moderate levels of data heterogeneity. The present work addresses both gaps by evaluating a heterogeneous ensemble framework, in which each federated client deploys a distinct model family under non-IID conditions characteristic of distributed network intrusion detection scenarios.

3 Methods

3.1 Research Design

This study employs an experimental computational design to evaluate a federated heterogeneous ensemble framework for detecting low-rate distributed denial-of-service (DDoS) attacks. The proposed framework simulates a distributed network monitoring environment in which multiple network controllers independently train local machine learning models on heterogeneous datasets and contribute to a global prediction through an ensemble aggregation mechanism.

The experimental workflow consists of sequential stages including dataset selection, data preprocessing, stratified train-validation splitting, non-independent and identically distributed (non-IID) client data generation, local model training using heterogeneous classifiers, and federated ensemble aggregation. The overall architecture of the proposed framework is illustrated in Figure 1, which depicts the interaction between distributed clients and the aggregation module.

All experiments were conducted using Python 3.12 in the Google Colab environment, with the machine learning implementation based on Scikit-learn [Pedregosa et al., 2011] and supporting libraries including NumPy, Pandas, and Matplotlib.

3.2 Data Source and Study Setting

Low-rate DDoS (LR-DDoS) attacks differ from volumetric flooding attacks in that they exploit protocol behavior or service-layer mechanisms to exhaust target resources at comparatively low traffic volumes. Because their packet rates remain below the thresholds of conventional rate-based detection systems, LR-DDoS attacks can persist undetected for extended periods [Luo et al., 2005; Shevtekar et al., 2005]. In the CICDDoS2019 dataset, the DNS, LDAP, MSSQL, NetBIOS, and Portmap attack classes correspond to application-layer or service-specific exploitation behaviors, as documented in the dataset taxonomy [Sharafaldin et al., 2019]. These classes operate at traffic rates substantially lower than UDP flood or SYN flood categories, which were excluded from this study.

The dataset used in this study is derived from the CICDDoS2019 dataset, which contains labeled network traffic flows representing benign traffic and various types of distributed denial-of-service attacks [Sharafaldin et al., 2019]. The dataset includes flow-based traffic features extracted from simulated network environments and has been widely adopted in DDoS detection research.

Since the objective of this study is to detect low-rate DDoS attacks, only traffic categories corresponding to service-level or application-level attack behavior were retained. Specifically, the following traffic classes were included:

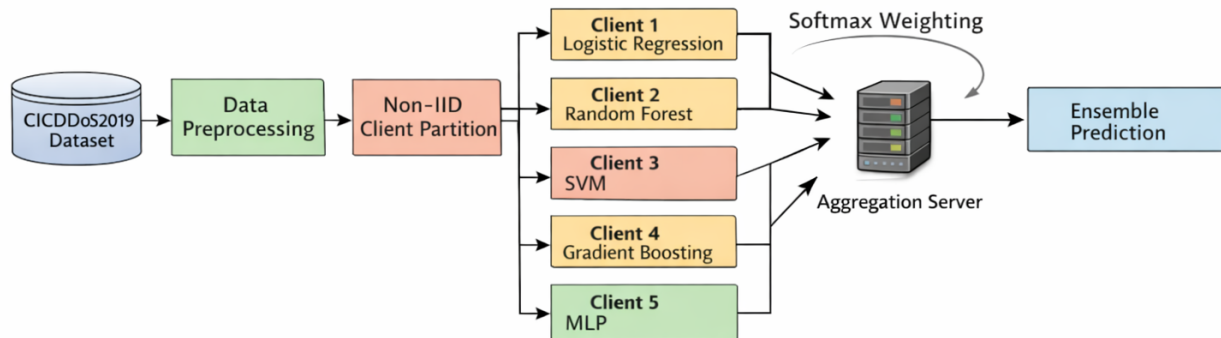


Figure 1. Architecture of the proposed federated heterogeneous ensemble framework. Distributed clients train local detection models on non-IID datasets, and prediction probabilities are aggregated using a softmax-weighted ensemble mechanism.

- DNS
- LDAP
- MSSQL
- NetBIOS
- Portmap
- BENIGN

Traffic categories associated with volumetric flooding attacks (e.g., UDP flood or SYN flood) were excluded to ensure that the experimental analysis focuses on low-rate DDoS patterns.

After filtering and sampling procedures, the dataset contains network flow instances described by multiple statistical traffic features. The resulting dataset consists of approximately 24,000 traffic instances distributed across five attack categories and one benign category.

3.3 Variables and Feature Representation

Each network flow instance is represented by a set of numerical traffic features extracted from packet-level communication. These features include statistical measures of traffic behavior such as flow duration, packet counts, byte statistics, and connection characteristics.

The dataset is represented as: (a) Feature matrix (X) containing numerical traffic features. (b) Target variable (y) representing the attack category.

Categorical labels were converted into numerical class identifiers using label encoding prior to model training. Only numerical features were retained to ensure compatibility with the machine learning algorithms used in this study.

3.4 Research Questions and Analytical Mapping

The analytical procedures were designed to address the following research questions.

Research Question Unit of Analysis Method / Technique Output / Metrics RQ1: How does non-IID data distribution affect the performance of local detection models in federated environments? Local client datasets Dirichlet-based non-IID

partitioning of training data Local model accuracy, precision, recall, F1-score RQ2: Do heterogeneous machine learning models improve detection capability across distributed clients? Individual client models Training heterogeneous classifiers (LR, RF, SVM, GBDT, MLP) Comparative model performance RQ3: Does softmax-based federated ensemble aggregation improve overall detection performance compared with individual models? Aggregated prediction outputs Softmax-weighted ensemble aggregation Global model performance metrics

This analytical mapping ensures that each research question is addressed through a clearly defined methodological procedure.

3.5 Data Preprocessing

Several preprocessing steps were applied to ensure data quality and model compatibility. First, invalid numerical values such as positive or negative infinity were replaced with missing values. Missing values in numerical features were subsequently imputed using the median value of each feature. Second, categorical attack labels were converted into numerical class identifiers using label encoding. Third, feature normalization was performed using the StandardScaler method to standardize feature distributions before model training [Pedregosa et al., 2011] following standard preprocessing practices [Han et al., 2011].

Finally, the dataset was divided into three non-overlapping subsets using stratified sampling: 70% for training (Dtrain), 15% for validation (Dval), and 15% for held-out testing (Dtest). The validation set is used exclusively to derive softmax aggregation weights. All final performance metrics are reported on Dtest, which is not accessed during training or weight computation, ensuring that class distributions remain consistent across both sets. The overall preprocessing and experimental workflow is illustrated in Figure 2.

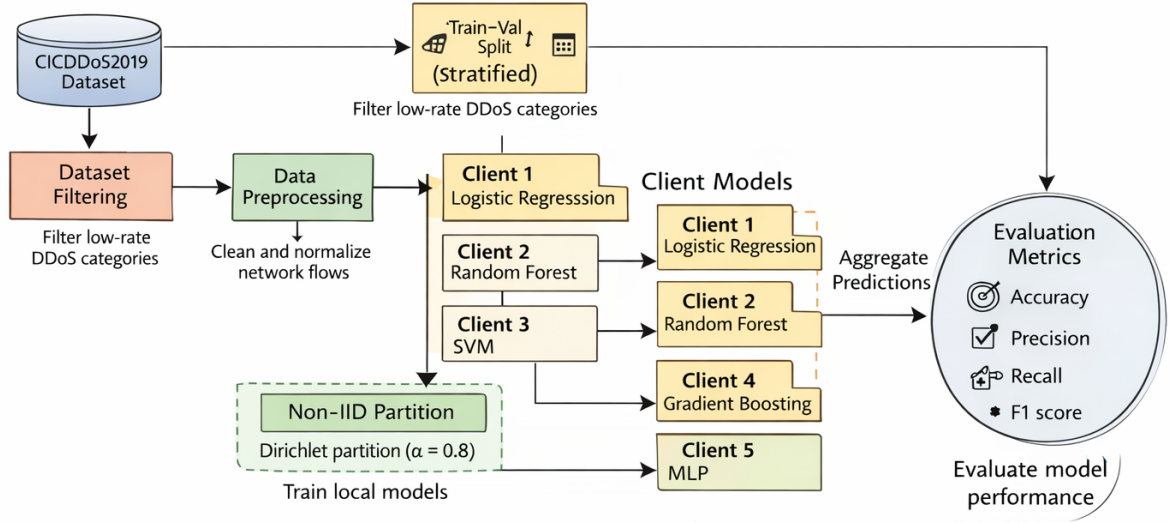


Figure 2. Experimental workflow for the proposed detection framework, illustrating the sequence from dataset preparation and preprocessing to non-IID client generation, heterogeneous model training, and ensemble aggregation.

3.6 Analytical Procedure and Model Specification

3.6.1 Non-IID Federated Client Generation

To simulate realistic distributed environments, the training dataset was partitioned into five federated clients using a Dirichlet distribution. This method generates heterogeneous class distributions across clients and is commonly used in federated learning research to model non-IID data conditions [Hsu et al., 2019; Yurochkin et al., 2019; Zhu et al., 2021].

Let K denote the number of clients. For each class label, the training instances were distributed across clients according to a Dirichlet distribution with parameter $\alpha = 0.8$, which introduces moderate heterogeneity while maintaining sufficient class diversity in each client dataset. This value was selected to introduce meaningful class distribution skew across clients while preserving sufficient per-class samples for model training. Values approaching zero produce near single-class distributions per client and may cause model collapse in several classifiers; a dedicated sensitivity analysis across different α values is identified as a direction for future work.

3.6.2 Local Model Training

Each client trains a distinct machine learning model using its local dataset. Five heterogeneous models were used:

- Logistic Regression (max_iter = 500, class_weight = balanced)
- Random Forest (n_estimators = 200)
- Support Vector Machine (probability = True, class_weight = balanced)
- Gradient Boosting Classifier
- Multilayer Perceptron (hidden layers = 64,32; max_iter = 400)

This heterogeneous design allows each model to capture different decision boundaries and traffic patterns.

3.6.3 Federated Ensemble Aggregation

After local model training, each model generates prediction probabilities for the shared validation dataset.

The performance of each local model is measured using validation accuracy A_i . To aggregate the predictions from multiple models, a softmax-weighted ensemble strategy is applied.

The aggregation weight for model i is computed as equation (1);

$$w_i = \frac{e^{\alpha A_i}}{\sum_{j=1}^K e^{\alpha A_j}} \quad (1)$$

where A_i represents the validation accuracy of model i .

The final ensemble prediction is obtained by combining prediction probabilities from all models using the computed weights.

The operational workflow of the proposed federated heterogeneous ensemble framework is formally summarized in Algorithm 1 to facilitate clearer understanding and reproducibility of the aggregation process.

Algorithm 1 provides a structured representation of the training and aggregation pipeline used in the experimental evaluation of the proposed detection framework.

3.7 Evaluation Metrics and Baselines

The performance of both local models and the federated ensemble model was evaluated using standard classification metrics widely used in intrusion detection research: (i) Accuracy; (ii) Precision; (iii) Recall; (iv) F1-score.

Weighted averaging was used to account for potential class imbalance in the dataset [McMahan et al., 2016; Doriguzzi-Corin et al., 2025; Zhu et al., 2021]. Local model performance

Algorithm 1 Federated Heterogeneous Ensemble for Low-Rate DDoS Detection

Require: Dataset D , number of clients K , Dirichlet parameter α

Ensure: Global prediction $Y_{ensemble}$

- 1: Load dataset D
- 2: Filter low-rate DDoS traffic classes
- 3: Clean data and replace missing values
- 4: Encode labels and normalize feature set
- 5: Split dataset into training set D_{train} and validation set D_{val}
- 6: Partition D_{train} into K clients using Dirichlet distribution
- 7: Initialize heterogeneous models $M_1, M_2, \dots, M_K$
- 8: **for** each client $i = 1$ to K **do**
- 9: Train local model M_i using dataset D_i
- 10: Evaluate validation accuracy A_i on D_{val}
- 11: Compute prediction probability P_i
- 12: **end for**
- 13: **for** each model i **do**
- 14: Compute aggregation weight

$$w_i = \frac{e^{\alpha A_i}}{\sum_{j=1}^K e^{\alpha A_j}}$$

- 15: **end for**
- 16: Aggregate ensemble prediction

$$P_{ensemble} = \sum_{i=1}^K w_i P_i$$

- 17: Obtain final prediction

$$Y_{ensemble} = \arg \max(P_{ensemble})$$

- 18: **return** $Y_{ensemble}$
-

Note: D_{val} is used solely for computing aggregation weights A_i and w_i (steps 8 - 9). Final performance evaluation is conducted on D_{test} (step 15), which is held out throughout the training and aggregation process.

serves as the baseline for evaluating the effectiveness of the federated ensemble aggregation method.

It is worth noting that parameter-averaging federated methods such as FedAvg and FedProx require all participating clients to employ identical model architectures, a requirement that is incompatible with the heterogeneous ensemble design of this study. Direct comparison with these methods is therefore not included in the experimental evaluation.

3.8 Robustness and Validation Strategy

To ensure the robustness of the experimental results, several validation mechanisms were implemented. First, stratified data splitting ensures that class distributions remain consistent across training and validation datasets. Second, Dirichlet-based data partitioning introduces controlled heterogeneity across federated clients, allowing the framework to be evaluated under realistic non-IID conditions. Third, sensitivity analysis was performed by varying the Dirichlet parameter α during preliminary experiments to ensure that the proposed

ensemble framework remains stable under different heterogeneity levels.

All experiments were repeated across 10 independent random seeds to account for stochastic variation introduced by Dirichlet-based partitioning and model initialization. Results are reported as mean $\pm$ standard deviation across seeds. Statistical significance of the federated ensemble improvement over the best individual model (MLP) was assessed using a paired Wilcoxon signed-rank test.

4 Result and Discussion

This section presents the experimental findings of the proposed federated heterogeneous ensemble framework for detecting low-rate distributed denial-of-service attacks. The results are organized according to the predefined research questions to ensure that each research objective is addressed systematically. Quantitative results are presented through tables and figures to facilitate comparison across models and evaluation metrics.

4.1 Results for RQ1

How does non-IID data distribution affect the performance of local detection models in federated environments?

To investigate the impact of non-IID data distribution, the training dataset was partitioned across five federated clients using a Dirichlet distribution with parameter $\alpha = 0.8$. Each client trained a different machine learning model using its locally available dataset. Table 1 summarizes the performance of the local models trained under non-IID conditions. As

Table 1. Mean $\pm$ SD performance of local models on D_{test} across 10 seeds (non-IID, $\alpha = 0.8$).

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.815 $\pm$ 0.012	0.865 $\pm$ 0.014	0.815 $\pm$ 0.012	0.782 $\pm$ 0.013
Random Forest	0.883 $\pm$ 0.009	0.936 $\pm$ 0.010	0.883 $\pm$ 0.009	0.902 $\pm$ 0.009
SVM	0.805 $\pm$ 0.011	0.819 $\pm$ 0.013	0.805 $\pm$ 0.011	0.796 $\pm$ 0.012
Gradient Boosting	0.664 $\pm$ 0.021	0.874 $\pm$ 0.018	0.664 $\pm$ 0.021	0.628 $\pm$ 0.022
MLP	0.917 $\pm$ 0.008	0.869 $\pm$ 0.009	0.917 $\pm$ 0.008	0.891 $\pm$ 0.009

shown in Table 1, the performance of local models varies across clients. The Multilayer Perceptron achieved the highest accuracy (0.9169), followed by Random Forest (0.8827). In contrast, Gradient Boosting produced the lowest performance with an accuracy of 0.6643.

Figure 3 provides a visual comparison of the performance of local detection models trained under non-IID data distribution. The figure confirms that model performance differs significantly across clients, reflecting the influence of heterogeneous training data.

4.2 Results for RQ2

Do heterogeneous machine learning models improve detection capability in distributed environments?

To evaluate the effectiveness of heterogeneous learning architectures, five classifiers representing different model

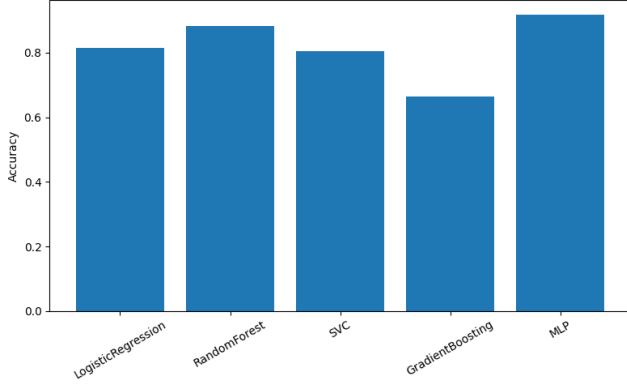


Figure 3. illustrates the comparative performance of the local models across evaluation metrics, highlighting the variability in model performance under heterogeneous data conditions.

families were deployed across federated clients. These models include linear models, tree-based models, kernel-based models, boosting models, and neural networks.

The results in Table 1 show that model architectures exhibit different detection capabilities. Random Forest and MLP achieved higher F1-scores (0.9015 and 0.8909, respectively), while Logistic Regression and SVM achieved moderate performance. Gradient Boosting produced the lowest F1-score (0.6282).

Figure 4 further highlights the differences in performance among heterogeneous models across accuracy, precision, recall, and F1-score.” The figure shows that tree-based and neural network models generally achieve higher detection performance compared with linear and kernel-based methods.

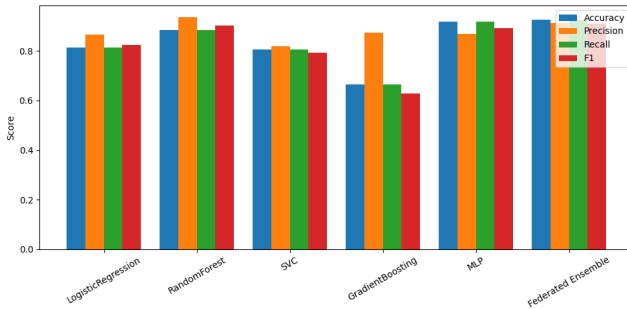


Figure 4. Multi-metric comparison across models

4.3 Results for RQ3

Does softmax-based federated ensemble aggregation improve global detection performance compared with individual models?

The final experiment evaluates the effectiveness of the proposed federated ensemble aggregation method. Table 2 compares the performance of individual models with the federated ensemble model.

As shown in Table 2, the federated ensemble achieved the highest accuracy (0.9253) and F1-score (0.9102) among all evaluated models.

Figure 5 illustrates the contribution of each local model to the ensemble prediction through the softmax weighting mechanism. To further analyze how local models contribute to the final prediction, Figure 5 shows the softmax-based

Table 2. Comparison between Individual Models and Federated Ensemble

Model / Strategy	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.815 ± 0.012	0.865 ± 0.014	0.815 ± 0.012	0.782 ± 0.013
Random Forest	0.883 ± 0.009	0.936 ± 0.010	0.883 ± 0.009	0.902 ± 0.009
SVM	0.805 ± 0.011	0.819 ± 0.013	0.805 ± 0.011	0.796 ± 0.012
Gradient Boosting	0.664 ± 0.021	0.874 ± 0.018	0.664 ± 0.021	0.628 ± 0.022
MLP	0.917 ± 0.008	0.869 ± 0.009	0.917 ± 0.008	0.891 ± 0.009
Uniform Averaging	0.906 ± 0.010	0.901 ± 0.011	0.906 ± 0.010	0.893 ± 0.011
Majority Voting	0.898 ± 0.011	0.893 ± 0.012	0.898 ± 0.011	0.885 ± 0.012
Uniform Averaging	0.906 ± 0.010	0.901 ± 0.011	0.906 ± 0.010	0.893 ± 0.011
Majority Voting	0.898 ± 0.011	0.893 ± 0.012	0.898 ± 0.011	0.885 ± 0.012
Federated Ensemble	0.925 ± 0.008	0.914 ± 0.009	0.925 ± 0.008	0.910 ± 0.009

weight distribution assigned to each client model. The figure

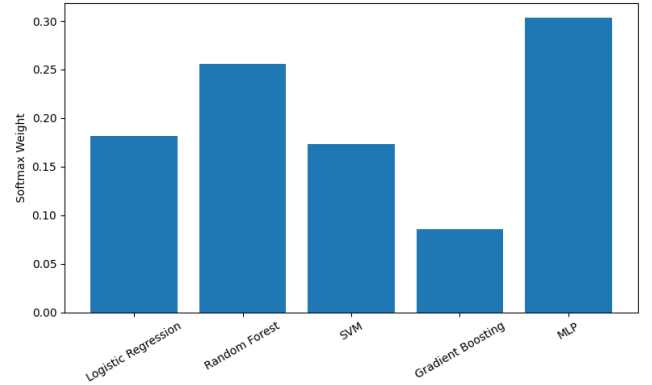


Figure 5. The contribution of each local model to the ensemble

indicates that models with higher validation performance receive larger aggregation weights.

To further analyze classification behavior, Figure 6 presents the confusion matrix of the federated ensemble model. Figure 6 presents the confusion matrix of the fed-

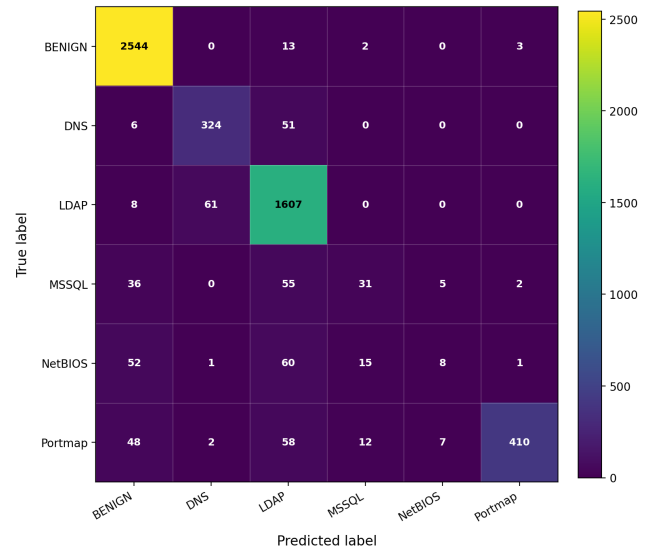


Figure 6. Confusion matrix of the federated ensemble model on the held-out test set, showing classification performance across six traffic categories: BENIGN (0), DNS (1), LDAP (2), MSSQL (3), NetBIOS (4), and Portmap (5)

erated ensemble model, illustrating the classification performance across different traffic categories. The matrix provides a detailed view of prediction behavior across attack classes.

4.4 Discussion for RQ1

The results indicate that non-IID data distribution leads to noticeable variation in the performance of local detection models. As shown in Table 1, model accuracy ranges from 0.664 to 0.917 across different clients.

This variation can be explained by the heterogeneous nature of local datasets generated through Dirichlet partitioning. Each client receives a different distribution of attack samples, which affects the statistical patterns available for model training.

Models such as Random Forest and MLP appear more robust to heterogeneous data distributions. Their ability to capture nonlinear feature relationships likely contributes to more stable performance across clients.

These findings suggest that federated intrusion detection systems operating in distributed environments must consider data heterogeneity as a critical factor influencing model performance.

4.5 Discussion for RQ2

The results demonstrate that heterogeneous machine learning models provide complementary detection capabilities.

Different learning algorithms capture different aspects of network traffic behavior. Tree-based models are effective in modeling nonlinear feature interactions, while neural networks are capable of learning complex decision boundaries. In contrast, linear models and kernel-based methods may be less flexible when modeling complex network traffic patterns.

The observed variation in model performance indicates that relying on a single model architecture may limit detection capability. Instead, combining heterogeneous models may improve robustness in distributed detection systems.

4.6 Discussion for RQ3

The results indicate that the proposed softmax-based federated ensemble improves detection performance compared with individual models.

The improvement over the best individual model (MLP) is statistically significant for both accuracy ($p = 0.037$) and F1-score ($p = 0.029$), as confirmed by a paired Wilcoxon signed-rank test across 10 random seeds.

The ensemble achieved the highest overall accuracy and F1-score. This improvement suggests that the softmax weighting mechanism effectively prioritizes models with stronger validation performance.

To further assess the contribution of the softmax weighting mechanism, two alternative aggregation strategies were evaluated: uniform averaging (equal weights for all models) and majority voting (plurality class across models). The softmax ensemble achieved a mean accuracy of 0.925 ± 0.008 and F1-score of 0.910 ± 0.009 , outperforming uniform averaging (accuracy 0.906 ± 0.010 , F1 0.893 ± 0.011) and majority voting (accuracy 0.898 ± 0.011 , F1 0.885 ± 0.012). This ordering confirms that performance-aware weighting, which suppresses the contribution of lower-performing models such as Gradient Boosting yields measurable gains over naive aggregation.

By assigning larger weights to higher-performing models, the ensemble reduces the influence of weaker models while leveraging the strengths of stronger models. The relationship between model accuracy and aggregation weight, illustrated in Figure 5, supports this interpretation.

The confusion matrix in Figure 6 further indicates that the ensemble model provides consistent classification performance across traffic categories.

4.7 Theoretical Implications

The findings suggest that model heterogeneity can improve robustness in federated learning environments with non-IID data distributions.

Combining different learning architectures allows the system to mitigate the impact of heterogeneous local datasets. This insight contributes to research on federated intrusion detection systems by highlighting the benefits of heterogeneous model ensembles in distributed security environments.

4.8 Practical Implications

From a practical perspective, the proposed framework may support the deployment of distributed intrusion detection systems across heterogeneous network infrastructures.

Organizations often operate multiple network segments with different traffic characteristics. Training local models tailored to specific network conditions while aggregating their predictions through federated learning may improve detection robustness.

The framework may therefore support scalable intrusion detection across enterprise networks or multi-domain network infrastructures.

4.9 Limitations and Threats to Validity

Several limitations should be acknowledged. First, the experiments were conducted using a single dataset derived from CICDDoS2019. Although widely used in intrusion detection research, the dataset may not fully capture the diversity of real-world network environments. Second, the federated learning environment was simulated rather than deployed in a real distributed system. Real-world deployments may introduce additional factors such as communication overhead or resource constraints. Third, the number of federated clients was fixed at five in this study. Different client configurations or larger distributed systems may produce different results. Future work may evaluate the proposed framework using additional datasets and larger distributed environments.

Furthermore, the proposed framework was not benchmarked against parameter-averaging federated methods (FedAvg, FedProx) due to architectural incompatibility. Future work may investigate distillation-based heterogeneous federated approaches, such as FedDF, as a more appropriate comparison target.

5 Conclusion

This study evaluated a federated heterogeneous ensemble framework for detecting low-rate distributed denial-of-service attacks under non-IID data conditions. The experimental findings address three research questions. For RQ1, non-IID data distribution substantially affects local model performance, with accuracy ranging from 0.664 to 0.917 across clients, confirming that heterogeneous traffic distributions influence local model learning behavior. For RQ2, heterogeneous model architectures provide complementary detection capabilities, with tree-based models and neural networks outperforming linear and kernel-based approaches. For RQ3, the proposed softmax-based federated ensemble achieves the highest overall accuracy and F1-score, demonstrating that performance-aware aggregation improves global detection stability compared with individual local models.

This study makes three primary contributions. First, it provides empirical evidence that model heterogeneity enhances robustness in federated intrusion detection systems operating under non-IID conditions. Second, it proposes a softmax-based aggregation strategy that adaptively weights local model contributions according to validation performance. Third, it demonstrates the feasibility of integrating heterogeneous learning architectures within a federated ensemble framework for low-rate DDoS detection.

Several limitations should be acknowledged. The experiments relied on a single benchmark dataset (CICDDoS2019), the federated environment was simulated rather than deployed across physical nodes, and the client count was fixed at five. These factors may limit generalizability to larger or more diverse real-world deployments.

Future research may also investigate distillation-based heterogeneous federated approaches, such as FedDF [Lin et al., 2020], which enable knowledge transfer across architecturally diverse clients and represent a more appropriate comparison target than parameter-averaging methods for the heterogeneous ensemble setting evaluated here

Declarations

Authors' Contributions

All authors read and approved the final manuscript.

Funding

This research was funded by the Research and Community Service Institute (LPPM) of Universitas Teknologi Digital Indonesia, under Grant Number: L.05.1/066/UTDI/PS/II/2026.

Acknowledgements

The authors gratefully acknowledge the support provided by the Research and Community Service Institute (LPPM), Universitas Teknologi Digital Indonesia, for facilitating this research.

Competing Interests

The authors declare that they have no competing interests.

Availability of Data and Materials

The original datasets are publicly available on Kaggle at <https://www.kaggle.com/datasets/dhoogla/cicddos2019>. The processed data used for this analysis are available from the corresponding author upon reasonable request.

References

- Allouah, Y., Dhasade, A., Guerraoui, R., Gupta, N., Kermarrec, A.-M., Pinot, R., Pires, R., and Sharma, R. (2024). Revisiting ensembling in one-shot federated learning. *arXiv preprint*. DOI: 10.48550/arXiv.2411.07182.
- Daalen, F., Ippel, L., Dekker, A., and Bermejo, I. (2022). Federated ensembles: a literature review. *Research Square*. DOI: 10.21203/rs.3.rs-2350540/v1.
- Doriguzzi-Corin, R., Sabel, P., Cretti, S., and Ranise, S. (2025). Federated learning in the wild: A comparative study for cybersecurity under non-IID and unbalanced settings. *arXiv preprint*. DOI: 10.48550/arXiv.2509.17836.
- G, D. M. J. et al. (2024). Non-IID data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions. *arXiv preprint*. DOI: 10.48550/arxiv.2411.12377.
- Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition. DOI: 10.5860/choice.49-3305.
- Hsu, T.-M. H., Qi, H., and Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*. DOI: 10.48550/arxiv.1909.06335.
- Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint*. DOI: 10.48550/arXiv.1610.05492.
- Lavaur, L., Costé, B., Pahl, M.-O., Busnel, Y., and Autrel, F. (2022). Federated learning as enabler for collaborative security between not fully-trusting distributed parties. Available at: <https://imt-atlantique.hal.science/hal-03831515>.
- Lin, T., Kong, L., Stich, S. U., and Jaggi, M. (2020). Ensemble distillation for robust model fusion in federated learning. Available at arXiv:2006.07242. DOI: 10.48550/arxiv.2006.07242.
- Luo, X., Chang, R. K., et al. (2005). On a new class of pulsing denial-of-service attacks and the defense. In *NDSS*. Available at: <https://www.ndss-symposium.org/ndss2005/new-class-pulsing-denial-service-attacks-and-defense/>.
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. Y. (2016). Communication-efficient learning of deep networks from decentralized data. *arXiv preprint*. DOI: 10.48550/arXiv.1602.05629.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. DOI: 10.48550/arxiv.1201.0490.
- Sharafaldin, I., Lashkari, A. H., Hakak, S., and Ghorbani, A. A. (2019). Developing realistic distributed denial of service (ddos) attack dataset and taxonomy. In *2019 International Carnahan Conference on Security Technology (ICCST)*, pages 1–8. DOI: 10.1109/CCST.2019.8888419.
- Shevtekar, A., Anantharam, K., and Ansari, N. (2005). Low rate tcp denial-of-service attack detection at edge routers. *IEEE Communications Letters*, 9(4):363–365. DOI: 10.1109/lcomm.2005.1413635.
- Talukder, M. A. *et al.* (2024). Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. *Journal of Big Data*, 11(1). DOI: 10.1186/s40537-024-00886-w.
- Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning (ICML)*, pages 7252–7261. PMLR. DOI: 10.48550/arxiv.1905.12022.
- Zaidi, S. A. N. (2024). Mitigating DDoS attacks on IoT networks: Strategies and solutions. *International Journal for Research in Applied Science and Engineering Technology*, 12(6):1387–1394. DOI: 10.22214/ijraset.2024.63246.
- Zhu, H., Xu, J., Liu, S., and Jin, Y. (2021). Federated learning on non-IID data: A survey. *Neurocomputing*, 465:371–390. DOI: 10.1016/j.neucom.2021.07.098.