



ARTICLE

TS-HEFS: Two-Stage Hybrid Ensemble Feature Selection for IP-Independent P2P Botnet Detection

Mohammad Syarifuz Zaim¹, Tohari Ahmad^{1,*} and Muhammad Aidiel Rachman Putra²

¹Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

²Department of Electrical and Information Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia

*Corresponding Author: Tohari Ahmad. Email: tohari@its.ac.id or tohari@if.its.ac.id

Received: 16 June 2026; Accepted: 09 September 2026; Published: 23 October 2026

ABSTRACT: Peer-to-peer (P2P) botnet detection remains challenging because its communication patterns closely resemble legitimate network traffic. In addition, communication identity-based features, such as SrcAddr and DstAddr, may increase the dependency of detection models on host identities rather than communication behavior. This study proposes a Two-Stage Hybrid Ensemble Feature Selection (TS-HEFS) method to identify representative non-IP features by integrating filtering and embedded feature selection through ranking aggregation and union feature selection. The method was evaluated on the CTU-13, NCC, NCC-2 sensor 1, NCC-2 sensor 3, and UNSW-NB15 (train-test set) datasets under five experimental scenarios. The best classification performance was obtained using the Random Forest classifier, achieving an F1-score of 99.48% on the NCC dataset and above 99% on both NCC-2 datasets after removing IP-based features. The union feature selection approach achieved higher classification performance than the intersection approach on the primary P2P botnet datasets, whereas the additional validation on UNSW-NB15 showed that the classification performance of the selected feature subset varied across datasets with different network traffic characteristics. These results indicate that communication behavior-based and data transfer-based features remain representative for P2P botnet detection after IP-based features are removed.

KEYWORDS: Botnet detection; cybersecurity; intelligent automation; network security; network infrastructure

1 Introduction

The development of computer network technology has increased the need for security systems capable of addressing various cyber threats, such as botnets [1,2]. In their operation, botnets utilize a collection of infected and controlled software to carry out malicious activities, ranging from Distributed Denial of Service (DDoS) attacks, malware distribution, and spam to data theft [3,4]. Along with the development of network infrastructure, botnets have also evolved into peer-to-peer (P2P) botnets, in which communication occurs in a distributed manner among nodes without relying on a single control center. This decentralized characteristic directly complicates the detection process, as P2P botnet communication patterns tend to resemble normal network traffic [5,6].

Machine learning-based methods have been widely used in botnet detection due to their ability to learn network traffic patterns and perform automatic classification. The use of classification algorithms such as Decision Tree, Random Forest, k -Nearest Neighbor, and Extreme Gradient Boosting has shown good performance in detecting botnets across various botnet datasets [6,7]. The application of ensemble learning has also been used to improve prediction stability through the combination of multiple algorithmic

models [8,9]. Nevertheless, model performance is influenced not only by the classification algorithm but also depends on the quality of the features selected from various network datasets [10,11].

Network traffic datasets generally contain a large number of features with diverse characteristics, but not all features contribute significantly to the classification process [12]. Irrelevant or redundant features not only increase computational burden but may also affect model stability and its ability to generalize to new data. Therefore, feature selection is one of the important stages in the botnet detection system process [10,13].

Various feature selection methods have been used to improve the quality of data representation in botnet detection systems. However, the feature selection results obtained often depend on the characteristics and evaluation mechanisms of each method, resulting in differences in the features considered important across methods [12,14]. Differences in network traffic characteristics can also affect the performance of feature selection methods and the features selected in each dataset [15]. This condition has encouraged the need for a feature selection method capable of utilizing the advantages of various methods while retaining information relevant to the classification process.

In addition to the selection method, the characteristics of the features used also affect the performance of botnet detection systems. Communication identity-based features, such as source IP address and destination IP address, are frequently used in the classification process because they can distinguish communication patterns among hosts [6]. Studies related to feature extraction and feature selection have also shown that the quality of feature representation affects the performance of classification models [11,12]. However, most previous studies have focused on improving detection performance through feature selection or feature transformation without evaluating the effect of using IP-based features on the classification process [6,16]. As a result, it remains unclear to what extent botnet activities can be represented using communication behavior-based features without relying on communication identity features. Therefore, a feature selection approach is needed that can retain relevant features to maintain classification performance even when communication identity information is not used.

Based on these limitations, this study proposes a Two-Stage Hybrid Ensemble Feature Selection (TS-HEFS) method that combines filtering and embedded feature selection methods through ranking aggregation and union feature selection. To evaluate the performance of the proposed method, this study used five testing scenarios, namely IP only, baseline, baseline without IP, TS-HEFS, and TS-HEFS without IP. These five scenarios were used to analyze the effect of feature selection and the use of IP-based features on P2P botnet classification performance.

The experiments were conducted using three P2P botnet datasets, namely CTU-13 scenario 12 dataset [17], NCC scenario 12 [18], NCC-2 sensor 1 and sensor 3 [19], together with an additional validation using the official training and testing sets of the UNSW-NB15 (UNSW) dataset [20]. Five classification algorithms, namely Decision Tree, Random Forest, Extra Trees, k -Nearest Neighbor, and Extreme Gradient Boosting, were used throughout the experiments. The P2P botnet datasets served as the primary evaluation datasets, whereas UNSW-NB15 was included only to provide additional validation under different network traffic characteristics. Model performance was evaluated using accuracy, precision, recall, and F1-score to analyze the impact of the feature selection method across different dataset characteristics. The main contributions of this study include:

- The development of a TS-HEFS method that combines filtering and embedded approaches through ranking aggregation and union feature selection.
- Evaluating the resilience of botnet classification models under IP-independent conditions through the removal of SrcAddr and DstAddr features.

The remainder of this paper is organized into several sections. [Section 2](#) discusses related studies on P2P botnet detection and feature selection. [Section 3](#) explains the method in this research. [Section 4](#) presents the experimental results and discussion. [Section 5](#) contains the conclusions and suggestions for future research.

2 Related Works

Research on P2P botnet detection has been widely conducted using various classification approaches, ensemble learning, and feature selection methods [5,16]. In addition to being used to improve classification performance, several studies have also developed hybrid feature selection and feature aggregation strategies to produce more stable feature representations [21–23]. This section discusses previous studies related to botnet detection, feature selection methods, feature aggregation strategies, and the use of communication identity-based features in the classification process.

Several studies have shown that the combination of classifiers and feature selection can improve botnet detection performance. Baruah et al. [6] used an ensemble classifier approach to detect P2P botnets by utilizing Chi-square-based feature selection, and the results showed that the combination of classification models could improve detection performance on P2P botnet datasets. Ramadhan et al. [24] proposed the SB-Net framework by integrating ensemble feature selection, ranking aggregation, and cascade learning to improve classification performance on a spam botnet dataset. These studies demonstrate that combining multiple classification or feature selection strategies can improve botnet detection performance.

In addition to classification models, feature quality is also an important factor in the botnet detection process. Commonly used feature selection methods include filtering, wrapper, and embedded feature selection. The filtering approach evaluates the statistical relationship between features and data labels, whereas embedded feature selection considers feature contributions during the model learning process [12,25]. Velasco-Mata et al. [26] showed that the use of feature selection can improve the stability of Decision Tree-based classification models in botnet detection.

The use of a single feature selection method remains a common approach in various studies, and each feature selection method has a different evaluation mechanism, resulting in different feature rankings [6,26]. This condition causes feature selection results to vary across datasets with different traffic characteristics. To address this limitation, several studies have begun combining multiple feature selection methods through hybrid and ensemble feature selection approaches [21,23]. Mahajan et al. [27] showed that the combination of several feature selection methods can produce a more stable feature subset compared to a single method. Budhraj et al. [23] proposed the Filter and Wrapper Stacking Ensemble (FWSE) approach as an effort to improve the quality of feature selection results through the integration of multiple methods.

In addition to combining multiple feature selection methods, several studies have also used feature aggregation strategies such as ranking aggregation, voting ensemble, and intersection feature selection. Hostiadi et al. [22] used the intersection method to retain features that appeared across all feature selection methods simultaneously. The intersection method can produce a more compact feature subset. However, it only retains features selected by all feature selection methods, so features that are relevant in certain methods may not be retained [21,22].

Several botnet detection studies have also used communication identity-based features such as source IP address, destination IP address, source port, and destination port as part of the classification process. These features are used because they can distinguish communication patterns among hosts and contribute substantially to model performance [5,6]. In addition, several studies have shown that communication identity-based features frequently appear as features with high importance in the feature selection process. Nevertheless, a high dependence on IP-based features may cause models to learn communication identities

rather than the behavioral characteristics of network traffic [26,28]. However, analyses of changes in model performance after the removal of IP-based features have rarely been conducted in previous studies.

A summary of previous studies on botnet detection and feature selection approaches is presented in Table 1. The comparison is based on feature selection methods, feature aggregation strategies, the use of IP-based features, datasets, and the limitations of each study. Based on previous studies, most research still uses a single feature selection method or a specific aggregation strategy in the feature selection process. In addition, most studies still retain communication identity-based features in the classification process without evaluating changes in model performance after these features are removed [21,22].

Table 1: Summary of previous studies on botnet detection and feature selection approaches.

Author	Feature Selection	Aggregation Strategy	IP Feature Usage	Dataset	Limitation
Velasco-Mata et al. [26]	Information gain and gini importance	No aggregation strategy	Without IP features	Botnet traffic dataset	Feature selection methods evaluated independently without feature aggregation
Das et al. [21]	Hybrid ensemble feature selection	Voting-based aggregation	Not evaluated	DDoS dataset	Limited to a single intrusion detection benchmark dataset
Hossain and Islam [25]	Hybrid feature selection	Not applied	Not evaluated	IoT botnet datasets	Limited to IoT botnet environments
Hostiadi et al. [22]	Hybrid feature selection	Intersection-based selection	Uses IP features	Botnet datasets	Does not evaluate the impact of IP feature removal
Ramadhan et al. [24]	Ensemble feature selection	Borda count	Uses IP features	Spam botnet datasets	Does not evaluate the impact of IP feature removal
Baruah et al. [6]	Chi-square feature selection	Not applied	Uses IP features	P2P botnet dataset	Limited to a single P2P botnet benchmark dataset
Proposed TS-HEFS	Two-stage hybrid ensemble feature selection	Ranking aggregation and union selection	With and without IP features	Botnet datasets	Evaluation focused on P2P botnet environments

Based on these conditions, several limitations remain in P2P botnet detection research. First, the use of a single feature selection method or a specific aggregation strategy may result in relevant features identified by other methods not being selected. Second, evaluations of model performance after the removal of IP features are still rarely conducted in previous studies. Therefore, a feature selection approach is needed that can utilize the advantages of various methods while also evaluating model resilience under IP-independent conditions.

3 Proposed Method

This study proposed a TS-HEFS method for IP-independent P2P botnet detection. The method integrated filtering and embedded feature selection through ranking aggregation and union feature selection

to produce a more stable feature representation across various scenarios. The research workflow is shown in Fig. 1.

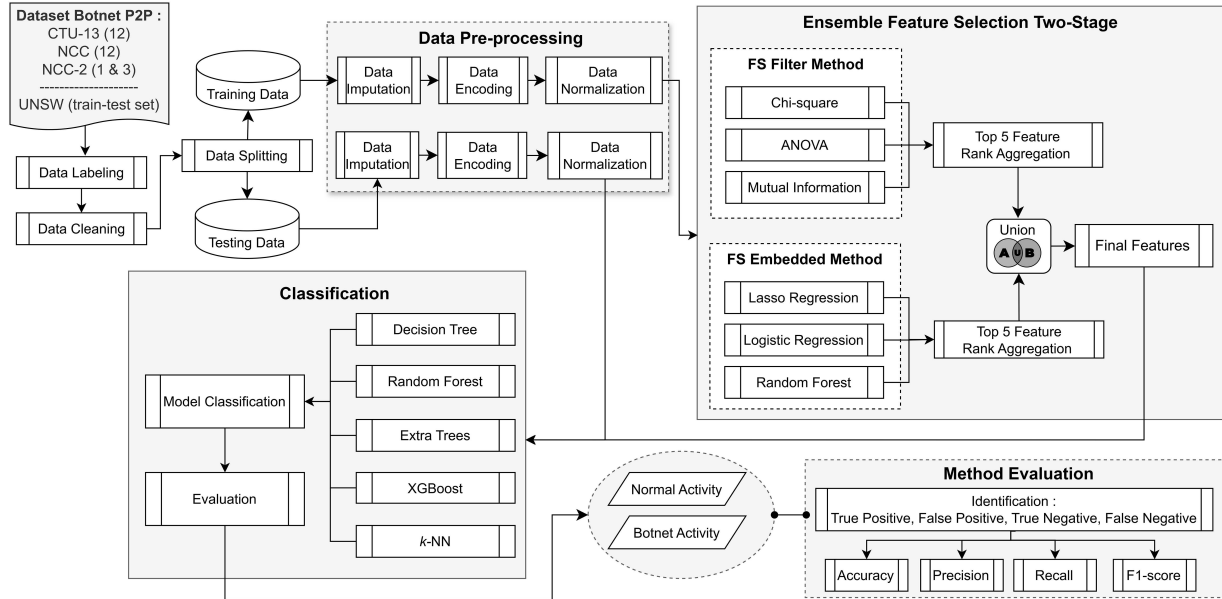


Figure 1: Proposed TS-HEFS method for IP-independent P2P botnet detection.

3.1 Data Preprocessing

The preprocessing stage was conducted to prepare the datasets before the feature selection and classification processes. The P2P botnet datasets used consisted of CTU-13 scenario 12 [17], NCC scenario 12 [18], and NCC-2 sensor 1 and sensor 3 [19], whereas the official training and testing sets of the UNSW-NB15 dataset were included as an additional validation dataset [20]. All datasets were converted into a binary classification scheme with label 0 representing normal traffic and label 1 representing attack traffic.

The preprocessing process included data cleaning, data splitting, missing value handling, encoding, and data normalization. During the data cleaning stage, duplicate records were removed to reduce redundancy. The StartTime, ActivityLabel, BotnetName, and SensorId features were removed because they only functioned as identity attributes or label indicators and did not represent network traffic characteristics [22,24]. In addition, the sTos and dTos features were also removed because they contained a high number of missing values [24]. For the UNSW-NB15 dataset, the id and attack_cat features were removed because they served as record identifiers and attack category labels, respectively, while binary classification was performed using the original class. In the P2P botnet datasets, the without IP scenario was created by removing the SrcAddr and DstAddr features before the data splitting process. This scenario was not applied to the UNSW-NB15 dataset because the official training and testing sets do not contain communication identity features. Consequently, the official training and testing sets were used only to provide additional validation of the proposed TS-HEFS approach under different network traffic characteristics.

After the data cleaning process, the P2P botnet datasets were divided into training and testing sets using stratified sampling with a ratio of 70:30. The split was performed before the subsequent preprocessing stages to prevent data leakage and to preserve the proportion of normal and botnet traffic in both subsets. In contrast, the official training and testing partitions provided with the UNSW-NB15 dataset were used directly without additional data splitting.

The missing value handling stage was conducted separately for categorical and numerical features. Categorical features were imputed using the mode value based on the data class, whereas numerical features were imputed using the median value based on the data class. This approach was used to preserve the data distribution while reducing the influence of outlier values in numerical features.

In the encoding stage, categorical features were converted into numerical form so that they could be processed by classification models. Frequency Encoding was applied to the Proto, Dir, and State features, which had high cardinality, whereas Label Encoding was used for the remaining categorical features. All numerical features were then normalized using Min-Max Scaling to a range of 0 to 1 before the feature selection process [12,24].

3.2 Two-Stage Hybrid Ensemble Feature Selection

This TS-HEFS method combined filtering and embedded feature selection approaches in a two-stage feature selection process [12,25]. The combination integrated statistical and model-based evaluations to produce a more comprehensive feature representation.

In the first stage, each feature selection method was executed independently to generate feature rankings according to its respective characteristics. The filtering methods used included Chi-square, ANOVA, and Mutual Information, which evaluated statistical relationships and information dependency between features and labels [6,12]. Meanwhile, the embedded methods used Lasso Regression, Logistic Regression, and Random Forest importance to evaluate feature contributions during the model training process [10,12]. Lasso Regression used $\alpha = 0.001$, Logistic Regression used $\text{max_iter} = 1000$, and Random Forest used 100 decision trees ($n_estimators = 100$). Mutual Information and all embedded methods used $\text{random_state} = 42$ to ensure reproducibility. The remaining hyperparameters followed the default settings of the corresponding library implementation. Subsequently, the ranking aggregation process was performed by selecting the top- k features from each group of feature selection methods [21,23]. In this study, the five best features from each feature selection method group were used to retain features with the highest contribution while reducing feature complexity.

In the second stage, the selected features from the filtering and embedded groups were combined using a union feature selection approach to retain features considered relevant by each feature selection method group [22,23]. The union feature selection process is shown in Eq. (1), where $F_{\text{filtering}}$ represents the set of features selected by the filtering methods and F_{embedded} represents the set of features selected by the embedded methods.

$$F_{\text{TS-HEFS}} = F_{\text{filtering}} \cup F_{\text{embedded}} \quad (1)$$

3.3 Classification and Evaluation

The classification process used five models: Decision Tree (DT), Random Forest (RF), Extra Trees (ET), k -Nearest Neighbor (k -NN), and Extreme Gradient Boosting (XGBoost). The use of multiple classification models was intended to analyze the stability of feature selection performance across different model characteristics.

Each model was selected based on its working mechanism. DT is a tree-structured classification model used to evaluate data separation capability using simple decision rules. Ensemble tree-based models such as RF and ET were used because they can handle non-linear relationships among features and reduce sensitivity to changes in feature representation. Meanwhile, XGBoost was used as a boosting-based approach, whereas k -NN was used as a distance-based method to evaluate the influence of feature representation on the classification process [24–26]. In this study, k -NN used a value of $k = 5$ to determine the number of nearest

neighbors during the classification process. RF and ET used 100 decision trees ($n_estimators = 100$), while DT, RF, ET, and XGBoost used $random_state = 42$ to ensure reproducibility. The remaining hyperparameters followed the default settings of the corresponding library implementation.

Model evaluation was conducted using accuracy, precision, recall, and F1-score (F1), which were calculated from the numbers of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). F1-score was used as the primary metric because it can evaluate the balance between precision and recall in imbalanced data distributions [2]. In addition to evaluation using testing data, 10-fold cross-validation was applied to the best-performing classification model to assess the consistency of classification performance across different data partitioning scenarios.

4 Results and Discussion

This section presents the experimental results and discussion of the proposed TS-HEFS method. The primary evaluation was conducted using three P2P botnet datasets, namely CTU-13, NCC, and NCC-2, whereas the UNSW-NB15 dataset was included as an additional validation dataset. The evaluation analyzes the effect of the proposed feature selection method on classification performance under different network traffic characteristics. The discussion focuses on the relationship among dataset characteristics, feature selection results, and classification model performance in each testing scenario.

4.1 Dataset Description

This study primarily used three groups of P2P botnet datasets, namely CTU-13 scenario 12 [17], NCC scenario 12 [18], NCC-2 sensor 1 and sensor 3 [19]. The datasets were selected because they have different network traffic characteristics and botnet communication patterns, allowing the evaluation of method performance under various network conditions. CTU-13 is a benchmark dataset that includes normal traffic, background traffic, and botnet activities in real network environments, with botnet communication patterns that tend to be non-periodic and therefore more difficult to distinguish from normal traffic [17]. The NCC dataset was developed to represent group activity-based botnet activities with more consistent communication patterns among nodes, resulting in a more structured pattern than CTU-13 [18]. NCC-2 represents multi-botnet activities with simultaneous attacks on multiple network sensors, containing a larger volume of data and greater variation in communication patterns than CTU-13 and NCC [19].

In addition to the primary P2P botnet datasets, the official training and testing sets of the UNSW-NB15 dataset [20] were included as an additional validation dataset. Unlike the primary datasets, UNSW-NB15 is a benchmark intrusion detection dataset containing normal traffic and multiple attack categories rather than focusing specifically on P2P botnet activities. Therefore, it was used only to provide additional validation of the proposed feature selection method under different network traffic characteristics. The official training and testing sets contain 175,341 and 82,332 network flow records, respectively. The training set consists of 119,341 attack and 56,000 normal instances, while the testing set contains 45,332 attack and 37,000 normal instances, indicating an imbalanced class distribution in both partitions.

All datasets were converted into a binary classification scheme consisting of normal and attack classes. Based on the data distribution, all datasets exhibited class imbalance conditions, where the number of normal or attack instances varied depending on the dataset characteristics. In addition, the size and communication characteristics of each dataset differed, resulting in different evaluation characteristics for the feature selection and classification processes. The characteristics of the primary P2P botnet datasets used in this study are presented in Table 2.

Table 2: Characteristics of P2P botnet datasets.

Dataset	Sub-Dataset	Duration	Bot Name	Type of Attack	Total Data	Botnet Data	Normal Data
CTU-13	12	1.21 h	NSIS.ay	P2P	325,471	2168	323,303
NCC	12	8 h	NSIS.ay	P2P	274,186	9000	265,186
NCC-2	1	8 h	Rbot, Neris, Sogo, NSIS.ay, Virut	IRC, PS, DDoS, US, SPAM, CF, HTTP, P2P	4,895,158	146,000	4,749,158
NCC-2	3	8 h	Rbot, Neris, Murlo, NSIS.ay, Virut	IRC, PS, DDoS, US, SPAM, CF, HTTP, P2P	3,885,792	294,000	3,591,792

4.2 Feature Selection Analysis

Table 3 summarizes the feature subsets obtained under each testing scenario for the primary P2P botnet datasets. In the IP only scenario, the classification process used only the two features, SrcAddr and DstAddr, to evaluate the contribution of communication identity-based features to classification performance. Meanwhile, the baseline scenario used all 11 features obtained after preprocessing without applying feature selection, whereas the baseline without IP scenario used 9 features after removing SrcAddr and DstAddr.

Table 3: Final selected features under different feature selection scenarios.

Dataset	IP Only	Baseline	Baseline without IP	TS-HEFS	TS-HEFS without IP
CTU-13	SrcAddr, DstAddr	Dur, Proto, SrcAddr, Sport, Dir, DstAddr, Dport, State, TotPkts, TotBytes, SrcBytes	Dur, Proto, Sport, Dir, Dport, State, TotPkts, TotBytes, SrcBytes	DstAddr, State, Sport, Dir, SrcAddr, Dport	Sport, State, Proto, Dir, Dur, Dport
NCC	SrcAddr, DstAddr	Dur, Proto, SrcAddr, Sport, Dir, DstAddr, Dport, State, TotPkts, TotBytes, SrcBytes	Dur, Proto, Sport, Dir, Dport, State, TotPkts, TotBytes, SrcBytes	SrcAddr, Sport, DstAddr, Dur, Dport	Sport, Dport, Dur, State, Dir, Proto
NCC-2 (Sensor 1)	SrcAddr, DstAddr	Dur, Proto, SrcAddr, Sport, Dir, DstAddr, Dport, State, TotPkts, TotBytes, SrcBytes	Dur, Proto, Sport, Dir, Dport, State, TotPkts, TotBytes, SrcBytes	SrcAddr, DstAddr, Proto, Sport, State, Dir, Dport	Sport, TotBytes, Dport, Dir, Proto, SrcBytes, State
NCC-2 (Sensor 3)	SrcAddr, DstAddr	Dur, Proto, SrcAddr, Sport, Dir, DstAddr, Dport, State, TotPkts, TotBytes, SrcBytes	Dur, Proto, Sport, Dir, Dport, State, TotPkts, TotBytes, SrcBytes	Dir, Dport, DstAddr, Sport, SrcAddr, State, Proto	Sport, State, Dport, Proto, Dir, Dur, TotBytes
UNSW (train-test set)	N/A	All features (42)		ct_state_ttl, sttl, dmean, dttl, dload, rate, is_sm_ips_ports	

The TS-HEFS and TS-HEFS without IP scenarios used a combination of filtering and embedded feature selection through ranking aggregation and union feature selection. The feature selection results show that

the TS-HEFS model was able to reduce the number of features to 5–7 features for each dataset. In the TS-HEFS scenario, IP-based features appeared in all datasets, whereas in the TS-HEFS without IP scenario, the feature selection process was dominated by communication behavior-based features such as Proto, State, Dir, Dur, Sport, and Dport. These features represent network communication behavior characteristics, including the protocol type used, connection status, communication direction, connection duration, and port usage patterns. This condition indicates that P2P botnet activities can still be represented using communication behavior-based features without relying on communication identity features.

Differences in feature subsets across datasets also indicate that botnet communication characteristics are not represented by the same features in all datasets. In CTU-13, communication pattern and network connectivity features appeared more frequently in the selected feature subsets. Meanwhile, in NCC-2 sensor 1 and sensor 3, data transfer-based features such as TotBytes and SrcBytes began to appear. This condition indicates that botnet activities in NCC-2 are influenced not only by network communication patterns but also by data transfer volume characteristics during the communication process.

Fig. 2 visualizes the relationship between features and data labels using the Pearson Correlation Coefficient (PCC) [29]. The correlation analysis results show that the contribution of each feature to the label differs across datasets, indicating that botnet activities are represented by different characteristics in each network environment. In CTU-13, all features exhibit relatively low correlation values with the label, indicating that no single feature dominantly represents botnet activities. In contrast, in NCC, the SrcAddr feature shows the highest positive correlation value compared to the other features, indicating that communication identity information still contributes substantially to distinguishing botnet traffic from normal traffic. Meanwhile, in NCC-2 sensor 1 and sensor 3, stronger correlations are observed in communication behavior-based features such as Proto, Sport, Dir, State, and Dport. These features represent protocol usage patterns, communication direction, connection status, and port usage characteristics associated with botnet activities. Although the DstAddr feature also showed a relatively high correlation in NCC-2, the stronger correlations of communication behavior-based features suggest that botnet activities in these datasets were associated with both communication identity and communication behavior characteristics. Overall, the differences in correlation patterns indicate that the contribution of individual feature categories varied across the evaluated P2P botnet datasets.

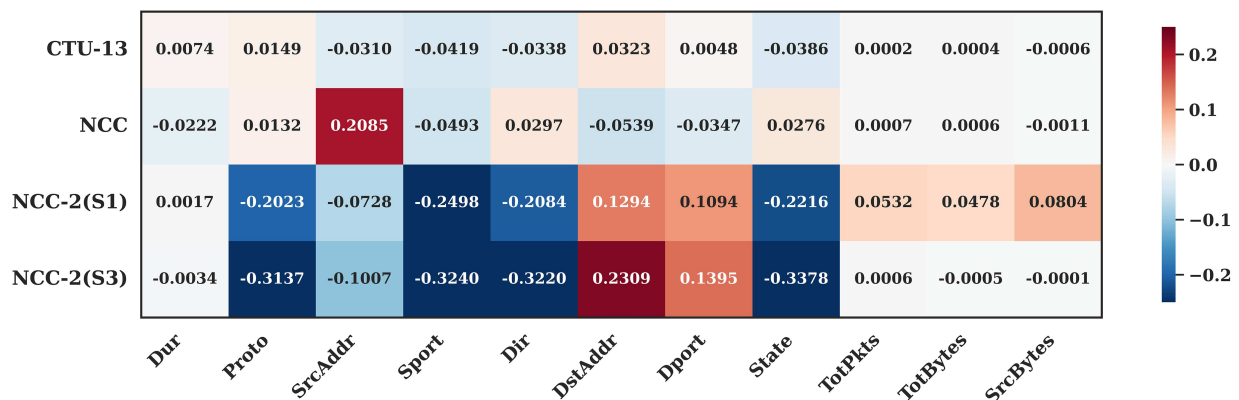


Figure 2: Feature correlation with label on P2P botnet datasets.

As an additional validation, the proposed TS-HEFS method was also evaluated using the official training and testing sets of the UNSW-NB15 dataset. Since the official dataset does not contain communication identity features, only the proposed TS-HEFS approach was applied. As shown in Table 3, the proposed

method reduced the original 42 predictive features to a compact feature subset. Unlike the primary P2P botnet datasets, the selected features were dominated by traffic statistics and connection state attributes rather than communication behavior-based features. This difference indicates that the selected feature subset varied according to the characteristics of the evaluated dataset rather than remaining the same across different network environments.

4.3 Classification Performance Evaluation

The classification results for each testing scenario are presented in Tables 4–6. The evaluation focused on performance changes across scenarios to analyze the stability of feature representations after feature selection and the removal of communication identity-based features were applied to different P2P botnet datasets and classification models.

Table 4 shows that the baseline scenario produced high classification performance on the NCC and NCC-2 datasets. Random Forest achieved an F1-score of 99.94% on NCC and 100% on NCC-2 sensor 1, while Decision Tree and Extra Trees also achieved F1-scores above 99% on most datasets. In contrast, model performance on the CTU-13 dataset showed greater variation across classification models. Extra Trees and XGBoost achieved F1-scores of 97.07% and 97.12%, respectively, whereas Random Forest, Decision Tree, and k -NN achieved F1-scores of only 78.43%, 78.26%, and 63.91%, respectively. These differences indicate that the contribution of the original feature set varied across datasets with different network traffic characteristics.

Table 4: Classification results for the baseline and baseline without IP scenarios on P2P botnet datasets.

Dataset	Model	Baseline				Baseline without IP			
		Acc. (%)	Pre. (%)	Rec. (%)	F1-score (%)	Acc. (%)	Pre. (%)	Rec. (%)	F1-score (%)
CTU-13	RF	99.64	65.04	98.77	78.43	99.89	91.18	92.31	91.74
	DT	99.63	64.88	98.62	78.26	99.90	92.76	92.62	92.69
	ET	99.96	97.22	96.92	97.07	99.84	86.21	91.38	88.72
	XGBoost	99.96	95.53	98.77	97.12	99.46	56.34	87.54	68.55
	k -NN	99.55	68.98	59.54	63.91	99.53	62.52	74.92	68.16
NCC	RF	99.99	99.89	100	99.94	99.98	99.74	99.74	99.74
	DT	99.97	99.30	99.89	99.59	99.93	98.28	99.70	98.99
	ET	99.98	99.78	99.67	99.72	99.94	98.75	99.48	99.11
	XGBoost	99.96	98.90	100	99.45	99.87	96.41	99.67	98.01
	k -NN	99.66	93.21	96.88	95.01	99.58	90.83	97.14	93.88
NCC-2 Sensor 1	RF	100	100	100	100	99.97	99.66	99.45	99.56
	DT	99.99	99.99	100	99.99	99.92	98.06	99.46	98.76
	ET	99.99	99.99	99.69	99.84	99.96	99.69	99.05	99.37
	XGBoost	99.99	99.96	99.99	99.98	99.84	97.66	96.87	97.26
	k -NN	99.85	97.59	97.38	97.49	99.57	94.94	90.42	92.63
NCC-2 Sensor 3	RF	99.99	99.99	99.98	99.99	99.91	99.50	99.35	99.42
	DT	99.99	99.98	99.99	99.99	99.78	97.87	99.29	98.57
	ET	99.98	99.98	99.78	99.88	99.90	99.54	99.12	99.33
	XGBoost	99.99	99.98	99.99	99.99	99.50	96.86	96.53	96.69
	k -NN	99.55	96.72	97.31	97.02	99.19	93.53	95.95	94.72

Table 5: Classification results for the IP only and TS-HEFS scenarios on P2P botnet datasets.

Dataset	Model	IP Only				TS-HEFS			
		Acc. (%)	Pre. (%)	Rec. (%)	F1-score (%)	Acc. (%)	Pre. (%)	Rec. (%)	F1-score (%)
CTU-13	RF	99.99	99.54	100	99.77	99.95	94.01	99.08	96.48
	DT	99.99	99.54	100	99.77	99.62	64.03	98.31	77.55
	ET	99.93	99.09	100	99.54	99.97	97.10	98.00	97.55
	XGBoost	99.94	96.98	94.00	95.46	99.96	95.39	98.77	97.05
	k-NN	99.54	91.98	33.54	49.15	99.67	77.72	71.38	74.42
NCC	RF	99.93	97.96	100	98.97	99.99	99.70	100	99.85
	DT	99.99	99.81	100	99.91	99.97	99.30	99.89	99.59
	ET	99.99	99.81	100	99.91	99.97	99.48	99.70	99.59
	XGBoost	99.96	98.93	100	99.46	99.96	98.86	100	99.43
	k-NN	99.82	98.51	95.99	97.24	99.65	93.35	96.44	94.87
NCC-2 Sensor 1	RF	99.99	99.99	100	99.99	99.99	99.99	100	99.99
	DT	99.99	99.99	100	99.99	99.93	99.97	97.76	98.85
	ET	99.98	99.98	99.28	99.63	99.99	99.99	99.57	99.78
	XGBoost	99.78	97.04	95.56	96.29	99.99	99.95	99.99	99.97
	k-NN	99.94	99.51	98.73	99.12	99.86	97.62	97.53	97.57
NCC-2 Sensor 3	RF	99.99	99.99	100	99.99	99.99	99.99	100	99.99
	DT	99.99	99.99	100	99.99	99.99	99.99	99.99	99.99
	ET	99.91	99.97	98.80	99.38	99.97	99.97	99.69	99.83
	XGBoost	99.70	97.98	97.99	97.98	99.99	99.95	99.88	99.91
	k-NN	99.85	99.65	98.42	99.03	99.53	96.51	97.24	96.88

Table 6: Classification results for the TS-HEFS without IP scenario on P2P botnet datasets.

Dataset	Model	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)	TP	FP	FN	TN
CTU-13	RF	99.82	83.57	90.00	86.67	585	115	65	96,877
	DT	99.90	92.82	91.54	92.18	595	46	55	96,946
	ET	99.76	78.90	88.00	83.20	572	153	78	96,839
	XGBoost	99.57	67.81	66.77	67.29	434	206	216	96,786
	k-NN	99.53	62.53	75.23	68.30	489	293	161	96,699
NCC	RF	99.96	99.26	99.70	99.48	2685	20	8	79,535
	DT	99.95	99.04	99.74	99.39	2686	26	7	79,529
	ET	99.95	98.93	99.59	99.26	2682	29	11	79,526
	XGBoost	99.90	98.19	98.89	98.54	2663	49	30	79,506
	k-NN	99.56	90.26	97.07	93.54	2614	282	79	79,273
NCC-2 Sensor 1	RF	99.90	99.13	99.19	99.25	43,232	299	354	1,424,321
	DT	99.90	98.06	98.57	98.31	42,961	851	625	1,423,769
	ET	99.94	99.32	98.67	98.99	43,008	296	578	1,424,324
	XGBoost	99.71	96.40	93.65	95.01	40,820	1524	2766	1,423,096
	k-NN	99.54	93.97	90.24	92.07	39,334	2524	4252	1,422,096
NCC-2 Sensor 3	RF	99.86	99.01	99.06	99.04	87,033	866	822	1,076,540
	DT	99.74	97.43	99.11	98.26	87,075	2297	780	1,075,109
	ET	99.84	99.19	98.73	98.96	86,738	704	1117	1,076,702
	XGBoost	99.29	95.49	95.01	95.25	83,474	3944	4381	1,073,462
	k-NN	99.19	93.52	95.94	94.72	84,286	5837	3569	1,071,569
UNSW (train-test set)	RF	83.26	78.41	96.02	86.33	43,528	11,982	1804	25,018
	DT	83.29	78.74	95.42	86.28	43,256	11,678	2076	25,322
	ET	83.29	78.50	95.94	86.35	43,491	11,914	1841	25,086
	XGBoost	81.88	75.60	99.07	85.76	44,912	14,497	420	22,503
	k-NN	52.27	61.98	34.46	44.29	15,622	9584	29,710	27,416

After the communication identity features were removed in the baseline without IP scenario, model performance on the NCC and NCC-2 datasets showed only relatively small performance changes. The F1-score of Random Forest on NCC decreased from 99.94% to 99.74%, while on NCC-2 it decreased from 100% to 99.56% on sensor 1 and from 99.99% to 99.42% on sensor 3. These results indicate that communication behavior-based and data transfer-based features continued to provide useful information for distinguishing botnet traffic after the removal of SrcAddr and DstAddr. This observation is consistent with the feature selection results in Table 3 and the correlation analysis in Fig. 2, where communication behavior-based features such as Proto, State, Dir, and Dport were consistently retained and exhibited relatively stronger correlations, particularly in the NCC-2 datasets. In contrast, the CTU-13 dataset showed greater performance variation after the communication identity features were removed. Random Forest and Decision Tree improved from 78.43% to 91.74% and from 78.26% to 92.69%, respectively, whereas XGBoost and Extra Trees decreased from 97.12% to 68.55% and from 97.07% to 88.72%, respectively. These results indicate that the contribution of communication identity features in CTU-13 differed from that in the NCC and NCC-2 datasets, with performance varying across both datasets and classification models.

The comparison presented in Table 5 shows that models in the IP only scenario achieved high performance on almost all datasets. On the NCC dataset, Random Forest achieved an F1-score of 98.97%, while Decision Tree and Extra Trees achieved 99.91%. On NCC-2 sensor 1 and sensor 3, Random Forest achieved an F1-score of 99.99%, while several other models also produced values close to 100%. In addition, identical performance metrics were obtained for several classifiers in the IP only scenario, including Random Forest and Decision Tree on CTU-13 and both NCC-2 sensors, and Decision Tree and Extra Trees on NCC. This pattern was observed among the tree-based classifiers when SrcAddr and DstAddr were used as the only input features. The consistently high F1-scores obtained using only SrcAddr and DstAddr indicate that communication identity-based features contributed substantially to the classification process and suggests that classification models may rely on communication identity information when these features are available.

After applying the TS-HEFS method, classification performance remained high even though the number of features was substantially reduced. On the NCC dataset, Random Forest achieved an F1-score of 99.85%, while on NCC-2 sensor 1 and sensor 3 it achieved 99.99%. In addition, XGBoost showed an increase in performance on NCC-2 sensor 1 from 96.29% in the IP only scenario to 99.97% after TS-HEFS was applied. The retained classification performance suggests that the TS-HEFS method preserved representative information while reducing the number of selected features. This interpretation is consistent with the feature selection results in Table 3, where communication behavior-based features, including Proto, State, Dir, Dur, Sport, and Dport, were consistently retained across multiple datasets.

Table 6 shows that the TS-HEFS without IP scenario maintained high classification performance for ensemble tree-based classifiers across the primary P2P botnet datasets. On the NCC dataset, Random Forest, Decision Tree, and Extra Trees achieved F1-scores of 99.48%, 99.39%, and 99.26%, respectively. Similar results were observed on NCC-2 sensor 1 and sensor 3, where these models consistently achieved F1-scores above 98%. In contrast, XGBoost and k -NN generally produced lower F1-scores across most datasets. On the CTU-13 dataset, Random Forest and Decision Tree achieved F1-scores of 86.67% and 92.18%, respectively, whereas XGBoost and k -NN achieved 67.29% and 68.30%. This difference indicates that classification performance on CTU-13 was more sensitive to changes in feature representation than on the NCC and NCC-2 datasets. This interpretation is supported by the feature selection results in Table 3 and the correlation analysis in Fig. 2, where communication behavior-based features were consistently retained and exhibited relatively stronger correlations than the other features, particularly in the NCC-2 datasets.

As an additional validation, the proposed TS-HEFS method was evaluated using the official training and testing sets of the UNSW-NB15 dataset. The evaluation was intended to examine the applicability of

the proposed method under different network traffic characteristics that differ from those of the primary P2P botnet datasets. Random Forest, Decision Tree, and Extra Trees produced comparable classification performance, with F1-scores of 86.33%, 86.28%, and 86.35%, respectively, whereas XGBoost and k -NN achieved lower F1-scores of 85.76% and 44.29%, respectively. These results indicate that the proposed feature selection method remained applicable to the additional validation dataset, although the classification performance differed from that observed on the primary P2P botnet datasets.

Across the evaluated datasets, each classification model responded differently to changes in feature representation introduced by the feature selection process. Based on the experimental configuration adopted in this study, Random Forest, Decision Tree, and Extra Trees generally produced more consistent classification performance than XGBoost and k -NN. Overall, the observed performance suggests that classification results were influenced by both the selected feature subset and the learning characteristics of the classification models.

4.4 Ablation Study

The ablation study evaluated the contribution of each feature selection component under the TS-HEFS without IP scenario. The evaluation used the Random Forest classifier because it achieved the highest classification performance across the previous experiments. The comparison results are presented in [Table 7](#) using five evaluation scenarios. The baseline scenario used all features available after preprocessing without applying feature selection. The filtering RA scenario applied filtering feature selection followed by Ranking Aggregation (RA) to select the top five features, whereas the embedded RA scenario applied embedded feature selection followed by Ranking Aggregation to select the top five features. The intersection scenario combined the top-5 feature subsets obtained from filtering RA and embedded RA using intersection feature selection to evaluate the effect of feature aggregation on classification performance. Meanwhile, the union scenario combined the top-5 feature subsets from filtering RA and embedded RA using union feature selection. The comparison of these scenarios was used to analyze the contribution of each feature selection component and its effect on the selected feature subsets and classification performance.

[Table 7](#) shows that each TS-HEFS component contributed differently to the classification performance. Compared with the baseline scenario, the filtering RA and embedded RA scenarios reduced the number of selected features while maintaining classification performance close to the baseline across most datasets. Across the primary P2P botnet datasets, the embedded RA scenario generally achieved higher F1-scores than the filtering RA scenario, increasing from 84.90% to 87.92% on CTU-13, from 93.70% to 99.07% on NCC-2 sensor 1, and from 90.70% to 98.85% on NCC-2 sensor 3. On the NCC dataset, both methods produced comparable performance, with F1-scores of 99.44% and 99.48%, respectively. In contrast, the additional validation on the UNSW-NB15 dataset showed a slightly higher F1-score for the filtering RA scenario 86.25% than for the embedded RA scenario 84.14%. This difference indicates that the contribution of filtering and embedded feature selection differed according to the characteristics of the evaluated datasets.

The combination of filtering RA and embedded RA resulted in different final feature subsets and classification performance. The intersection scenario consistently generated the smallest feature subsets because it only retained features selected by both ranking aggregation processes. Although this approach further reduced the number of selected features, the F1-score generally decreased compared with the baseline scenario, particularly on CTU-13, NCC-2 sensor 1, NCC-2 sensor 3, and the additional validation dataset. In contrast, the union scenario retained a larger number of selected features and generally produced classification performance closer to the baseline across the evaluated datasets. For example, the F1-score on NCC decreased only from 99.74% in the baseline scenario to 99.48%, while NCC-2 sensor 1 decreased from 99.56% to 99.25%, NCC-2 sensor 3 from 99.42% to 99.04%, and UNSW-NB15 from 89.44% to 86.33%. Across

the primary P2P botnet datasets, communication behavior-based features such as Proto, State, Sport, Dport, Dir, and Dur were consistently retained. The retained feature subsets are consistent with the feature selection results in Table 3 and the feature correlation analysis in Fig. 2, where Proto, State, and Dport exhibited relatively stronger correlations, particularly in the NCC-2 datasets. Although the additional validation on the UNSW-NB15 dataset produced a different feature subset, the union scenario also achieved the highest F1-score among the evaluated feature selection configurations. Overall, combining filtering and embedded feature selection through the union approach maintained classification performance closer to the baseline while reducing the number of selected features across different dataset characteristics.

Table 7: Ablation study of TS-HEFS components without IP features.

Dataset	Scenario	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)	Feature
CTU-13	Baseline	99.89	91.19	92.31	91.74	All features without IP (9)
	Filtering RA	99.80	87.32	82.62	84.90	State, Sport, Dir, Proto, Dur
	Embedded RA	99.84	85.80	90.15	87.92	Sport, State, Proto, Dur, Dport
	Intersection	99.81	88.36	82.92	85.56	Sport, State, Proto, Dur
	Union	99.82	83.57	90.00	86.67	Sport, State, Proto, Dir, Dur, Dport
NCC	Baseline	99.98	99.74	99.74	99.74	All features without IP (9)
	Filtering RA	99.96	99.19	99.70	99.44	Sport, Dur, Dport, State, Dir
	Embedded RA	99.97	99.26	99.70	99.48	Sport, Dport, Dur, State, Proto
	Intersection	99.96	99.15	99.70	99.43	Sport, Dport, Dur, State
	Union	99.97	99.26	99.70	99.48	Sport, Dport, Dur, State, Dir, Proto
NCC-2 Sensor 1	Baseline	99.97	99.66	99.45	99.56	All features without IP (9)
	Filtering RA	99.63	94.95	92.49	93.70	State, Dir, Sport, Proto, Dport
	Embedded RA	99.94	99.31	98.83	99.07	Sport, Dport, State, SrcBytes, TotBytes
	Intersection	99.62	94.96	92.06	93.49	Sport, State, Dport
	Union	99.96	99.31	99.19	99.25	Sport, State, Dport, Dir, Proto, SrcBytes, TotBytes
NCC-2 Sensor 3	Baseline	99.91	99.50	99.35	99.42	All features without IP (9)
	Filtering RA	98.61	91.48	89.94	90.70	State, Proto, Dir, Sport, Dport
	Embedded RA	99.83	98.86	98.85	98.85	Sport, Dport, State, Dur, TotBytes
	Intersection	98.59	91.45	89.62	90.52	Sport, State, Dport
	Union	99.86	99.01	99.06	99.04	Sport, State, Dport, Proto, Dir, Dur, TotBytes
UNSW (train-test set)	Baseline	87.18	81.85	98.57	89.44	All features (42)
	Filtering RA	83.17	78.40	95.85	86.25	sttl, ct_state_ttl, dload, rate, dmean
	Embedded RA	79.27	72.68	99.90	84.14	ct_state_ttl, dttl, dmean, sttl, is_sm_ips_ports
	Intersection	79.19	72.66	99.73	84.07	ct_state_ttl, sttl, dmean
	Union	83.26	78.41	96.02	86.33	ct_state_ttl, sttl, dmean, dttl, dload, rate, is_sm_ips_ports

4.5 Statistical Analysis

Statistical analysis was conducted following the classification and ablation experiments to evaluate the consistency of classification performance and the difference between the intersection and union feature aggregation approaches in the proposed TS-HEFS method. A 10-fold cross-validation was performed using the Random Forest classifier, with both approaches evaluated using the same data partitions. The resulting fold-wise F1-scores were used to compare the two approaches on each corresponding fold. The Nadeau and Bengio corrected resampled *t*-test was applied to evaluate whether the mean difference between the two approaches differed significantly from zero. This test was selected because the training sets in cross-validation overlap across folds, which can lead to an underestimated variance when fold results are treated as independent observations. The correction adjusts the variance estimate to account for this overlap [30].

A 95% confidence interval (CI) was also calculated based on the corrected variance estimate to indicate the range of plausible values for the mean difference.

Tables 8 and 9 summarize the results of the 10-fold cross-validation and statistical analysis. Both feature aggregation approaches showed relatively consistent classification performance, with standard deviations ranging from 0.05% to 2.27% across the evaluated datasets. On the primary P2P botnet datasets, the union approach achieved mean F1-scores of 90.78% on CTU-13, 99.18% on NCC-2 sensor 1, and 98.97% on NCC-2 sensor 3, corresponding to mean F1-score gains of 4.23%, 5.43%, and 8.49%, respectively, compared with the intersection approach. For the NCC dataset, the mean difference was only -0.04% , indicating very similar average performance between the two approaches. The corrected resampled t -test indicated statistically significant differences for CTU-13 ($p = 0.0015$), NCC-2 sensor 1 ($p < 0.0001$), NCC-2 sensor 3 ($p < 0.0001$), and UNSW-NB15 ($p = 0.0167$). The corresponding 95% CIs for the mean differences did not include zero. In contrast, the NCC dataset showed no statistically significant difference ($p = 0.4135$), with a 95% CI of $[-0.15, 0.07]$. For UNSW-NB15, the intersection approach obtained a slightly higher mean F1-score than the union approach, with values of 94.93% and 94.78%, respectively, resulting in a mean difference of -0.15% .

Table 8: F1-score (%) obtained from 10-fold cross-validation using the random forest classifier.

Dataset	Method	1	2	3	4	5	6	7	8	9	10
CTU-13	Intersection	87.04	88.89	85.71	84.83	84.07	88.18	89.70	86.20	84.69	86.18
	Union	89.63	94.39	91.28	86.12	92.76	91.56	92.31	90.20	89.11	90.43
NCC	Intersection	99.37	99.76	99.52	99.29	99.76	99.36	99.60	99.60	99.52	99.37
	Union	99.29	99.52	99.52	99.29	99.68	99.44	99.60	99.68	99.36	99.37
NCC-2 Sensor 1	Intersection	93.83	94.02	93.59	93.64	93.72	93.48	93.83	93.65	94.09	93.69
	Union	99.25	99.15	99.17	99.18	99.13	99.13	99.17	99.21	99.28	99.17
NCC-2 Sensor 3	Intersection	90.66	90.45	90.56	90.50	90.58	90.47	90.62	90.36	90.58	90.03
	Union	98.98	98.96	99.06	98.94	98.95	98.90	98.99	99.00	99.08	98.88
UNSW (train-test set)	Intersection	95.02	94.80	94.96	94.81	94.82	95.02	94.88	95.04	94.99	94.95
	Union	94.71	94.68	94.85	94.77	94.88	94.90	94.72	94.86	94.80	94.67

Table 9: Statistical comparison between the intersection and union feature aggregation approaches.

Dataset	Intersection (Mean $\pm$ SD)	Union (Mean $\pm$ SD)	Mean Gain (%)	95% CI	t-Statistic	p-Value
CTU-13	86.55 $\pm$ 1.88	90.78 $\pm$ 2.27	4.23	[2.09, 6.37]	4.4763	0.0015
NCC	99.52 $\pm$ 0.17	99.47 $\pm$ 0.15	-0.04	$[-0.15, 0.07]$	-0.8573	0.4135
NCC-2 Sensor 1	93.75 $\pm$ 0.19	99.18 $\pm$ 0.05	5.43	[5.25, 5.61]	69.7190	<0.0001
NCC-2 Sensor 3	90.48 $\pm$ 0.18	98.97 $\pm$ 0.06	8.49	[8.33, 8.65]	119.2033	<0.0001
UNSW (train-test set)	94.93 $\pm$ 0.09	94.78 $\pm$ 0.09	-0.15	$[-0.26, -0.03]$	-2.929	0.0167

An additional 20-fold cross-validation was also performed using the Random Forest classifier as supplementary validation. As this analysis was intended only as supplementary validation, only the mean F1-score across the 20 folds is presented. The proposed TS-HEFS method achieved mean F1-scores of 91.39% on CTU-13, 99.50% on NCC, 99.22% on NCC-2 sensor 1, 99.00% on NCC-2 sensor 3, and 94.84% on the UNSW-NB15 dataset. These results followed the same performance trend observed in the 10-fold cross-validation,

indicating that increasing the number of cross-validation folds did not substantially change the overall classification performance across the evaluated datasets. Overall, the statistical analysis indicates that the union feature aggregation approach achieved higher classification performance on CTU-13 and both NCC-2 datasets, whereas comparable performance was observed on the NCC dataset. The additional validation using the UNSW-NB15 dataset showed a slightly higher mean F1-score for the intersection approach.

4.6 Computational Complexity Analysis

Computational complexity analysis was conducted to provide a theoretical evaluation of the proposed TS-HEFS method. Let n denote the number of training samples, m the number of input features, and T the number of trees in the Random Forest model. Since the numbers of filtering and embedded feature selection methods are fixed in the proposed method, they are treated as constant factors and therefore omitted from the asymptotic Big-O notation. The filtering feature selection stage consists of Chi-square, ANOVA, and Mutual Information, each of which independently evaluates all input features over the training samples. Chi-square and ANOVA result in a computational complexity of $O(nm)$, while Mutual Information, estimated using a k-nearest-neighbor approach, results in $O(nm \log n)$. Therefore, the filtering feature selection stage results in a computational complexity of $O(nm \log n)$. The embedded feature selection stage consists of Lasso Regression, Logistic Regression, and Random Forest. Although all three methods are employed, the computational complexity is dominated by Random Forest, which constructs T decision trees over the training samples, resulting in $O(Tnm \log n)$, while the embedded ranking aggregation stage requires $O(m \log m)$ to aggregate and sort the feature rankings. Subsequently, the union feature selection stage combines the top-ranked features obtained from the filtering and embedded stages through a set union operation with a computational complexity of $O(m)$. The resulting feature subset is then reordered using the combined ranking scores, requiring an additional computational complexity of $O(m \log m)$. Therefore, the overall computational complexity of the proposed TS-HEFS method can be expressed as $O(nm \log n) + O(m \log m) + O(Tnm \log n) + O(m \log m) + O(m) + O(m \log m)$, which can be simplified to $O(Tnm \log n)$.

The computational complexity analysis indicates that the embedded feature selection stage contributes substantially to the overall computational complexity of the proposed TS-HEFS method because the Random Forest component constructs multiple decision trees and repeatedly evaluates the input features during the training process. In contrast, the filtering feature selection stage involves feature scoring using three statistical methods, followed by rank aggregation and sorting, while the subsequent union feature selection stage involves set operations and feature reordering. These operations introduce additional computational costs but do not involve iterative tree construction. Overall, the computational complexity of TS-HEFS is primarily influenced by the Random Forest-based embedded feature selection stage, while the other stages contribute additional feature scoring, ranking, and feature processing operations.

4.7 Comparison with Previous Studies

A comparative analysis was conducted by comparing the proposed TS-HEFS method with previously published studies on the CTU-13, NCC, NCC-2, and UNSW-NB15 benchmark datasets. The comparison aims to evaluate the classification performance of the proposed TS-HEFS method relative to existing studies on the same benchmark datasets. The comparison results are presented in Table 10. On the CTU-13 dataset, the proposed TS-HEFS method achieved its highest F1-score of 92.18% using the Decision Tree classifier. This result was higher than the F1-score of 81.19% reported by Hostiadi et al. [22] using the same classifier but remained lower than the highest F1-score of 99.32% reported by Baruah et al. [6], which employed Chi-square as a single feature selection method and a majority voting ensemble classifier. These results indicate that the classification performance on CTU-13 varied across different feature selection and classification strategies.

Table 10: Comparison of classification performance with previous studies on the evaluated datasets.

Authors	Datasets	Classification Method	Acc. (%)	Pre. (%)	Rec. (%)	F1-Score (%)
Baruah et al. [6]	CTU-13	Linear Discriminant Analysis	99.32	5.43	1.58	2.20
		<i>k</i> -NN	99.89	90.64	93.59	92.08
		Naïve Bayes	96.73	3.60	98.93	6.52
		Decision Tree	99.98	98.43	98.75	98.59
		Support Vector Machine	99.61	81.30	54.78	65.31
		Random Forest	99.62	97.89	98.65	99.02
		Majority Voting Ensemble	99.99	99.81	99.12	99.32
Hostiadi et al. [22]	CTU-13	Decision Tree	95.91	74.99	88.50	81.19
		Support Vector Machine	97.95	62.30	68.00	65.03
		<i>k</i> -NN	83.21	78.30	79.03	78.67
		Naïve Bayes	86.79	62.58	84.59	71.94
		Logistic Regression	97.94	76.18	67.69	71.68
	NCC	Decision Tree	99.81	97.23	95.33	96.27
		Support Vector Machine	98.55	79.13	70.88	74.78
		<i>k</i> -NN	99.75	95.36	95.57	95.47
		Naïve Bayes	84.27	63.08	75.27	68.64
		Logistic Regression	97.50	74.37	61.37	67.25
	NCC-2	Decision Tree	91.95	74.50	92.89	82.68
		Support Vector Machine	92.28	64.50	92.89	76.13
		<i>k</i> -NN	91.15	72.84	63.59	67.90
		Naïve Bayes	84.72	62.29	74.03	67.66
		Logistic Regression	94.79	70.13	75.70	72.81
	UNSW-NB15	Decision Tree	98.60	91.85	96.33	93.63
		Support Vector Machine	98.56	90.45	99.26	94.31
		<i>k</i> -NN	98.61	92.04	96.34	93.72
		Naïve Bayes	74.72	31.60	95.35	44.87
		Logistic Regression	97.84	89.22	94.74	91.34
TS-HEFS	CTU-13	Random Forest	99.82	83.57	90.00	86.67
		Decision Tree	99.90	92.82	91.54	92.18
		Extra Trees	99.76	78.90	88.00	83.20
		XGBoost	99.57	67.81	66.77	67.29
		<i>k</i> -NN	99.53	62.53	75.23	68.30
	NCC	Random Forest	99.96	99.26	99.70	99.48
		Decision Tree	99.95	99.04	99.74	99.39
		Extra Trees	99.95	98.93	99.59	99.26
		XGBoost	99.90	98.19	98.89	98.54
		<i>k</i> -NN	99.56	90.26	97.07	93.54
	NCC-2 (Average)	Random Forest	99.88	99.07	99.13	99.15
		Decision Tree	99.81	97.74	98.84	98.29
		Extra Trees	99.89	99.26	98.70	98.98
		XGBoost	99.49	95.94	94.33	95.13
		<i>k</i> -NN	99.37	93.75	93.09	93.39
	UNSW-NB15	Random Forest	83.26	78.41	96.02	86.33
		Decision Tree	83.29	78.74	95.42	86.28
		Extra Trees	83.29	78.50	95.94	86.35
		XGBoost	81.88	75.60	99.07	85.76
		<i>k</i> -NN	52.27	61.98	34.46	44.29

On the NCC dataset, the proposed TS-HEFS method achieved its highest F1-score of 99.48% using the Random Forest classifier, compared with the highest F1-score of 96.27% reported by Hostiadi et al. [22]. A similar trend was observed on the NCC-2 dataset, where TS-HEFS achieved an F1-score of 99.15%, exceeding the highest F1-score of 82.68% reported in the previous study. Unlike Hostiadi et al. [22], which employed an intersection-based feature selection strategy, the proposed method constructs the final feature subset using a hybrid feature selection method with union feature aggregation. These results indicate that the proposed TS-HEFS method achieved higher classification performance on both NCC datasets.

Different results were observed on the UNSW-NB15 dataset. The highest F1-score achieved by TS-HEFS was 86.35% using the Extra Trees classifier, whereas Hostiadi et al. [22] reported a highest F1-score of 94.31% using the Support Vector Machine classifier. Although the proposed method produced lower classification performance on this dataset, the comparison indicates that the relative performance of the evaluated approaches differed according to the dataset characteristics and the experimental settings adopted in each study. Overall, the comparison with previous studies demonstrates that the proposed TS-HEFS method achieved competitive classification performance across the evaluated benchmark datasets. Higher classification performance than previous studies was observed on the NCC and NCC-2 datasets, whereas different performance trends were obtained on the CTU-13 and UNSW-NB15 datasets. These results indicate that the relative classification performance of the evaluated methods varied across the benchmark datasets, reflecting differences in both the feature selection strategies and the experimental settings adopted in each study.

4.8 Discussion

Across the primary P2P botnet datasets, the proposed TS-HEFS method showed F1-score decreases of 0.26%, 0.31%, and 0.38% on the NCC, NCC-2 sensor 1, and NCC-2 sensor 3 datasets, respectively, after IP-based features were removed. On CTU-13, the F1-score decreased by 5.07%. These results indicate that the contribution of IP-based features differed across the evaluated datasets. These findings indicate that communication behavior-based and data transfer-based features remained representative for characterizing P2P botnet traffic after IP-based features were removed.

The ablation and statistical analyses also support the proposed union feature aggregation strategy. Compared with the intersection approach, the union approach achieved higher mean F1-scores on the CTU-13, NCC-2 sensor 1, and NCC-2 sensor 3 datasets, with improvements of 4.23%, 5.43%, and 8.49%, respectively. In contrast, only a 0.04% difference was observed on the NCC dataset, which is consistent with the corrected resampled *t*-test indicating no statistically significant difference ($p = 0.4135$). These results indicate that the union feature aggregation strategy retained complementary features selected by the filtering and embedded feature selection stages that were not preserved by the intersection approach, resulting in higher classification performance.

A different performance trend was observed during the additional validation using the UNSW-NB15 dataset. Unlike the primary P2P botnet datasets, the intersection approach produced a slightly higher mean F1-score than the union approach during cross-validation (94.93% and 94.78%, respectively), and the highest F1-score obtained by the proposed TS-HEFS method was 86.35%. These results indicate that the feature subset identified from the evaluated P2P botnet datasets did not consistently achieve similar classification performance on the additional intrusion detection dataset. This result suggests that the classification performance of the selected feature subset may vary according to the characteristics of the evaluated dataset.

5 Conclusion

The experimental results showed that the proposed TS-HEFS method produced compact feature subsets, with the effect of removing IP-based features varying across the evaluated datasets. The F1-score decreased by 0.26%, 0.31%, and 0.38% on the NCC, NCC-2 sensor 1, and NCC-2 sensor 3 datasets, respectively, and by 5.07% on CTU-13. Communication behavior-based and data transfer-based features were consistently selected across the evaluated datasets, indicating that these features remained representative for characterizing P2P botnet traffic without relying on IP-based attributes.

The proposed union feature selection approach achieved higher classification performance than the intersection approach across the primary P2P botnet datasets. The statistical analysis identified statistically significant performance differences on the CTU-13 and NCC-2 datasets, whereas no statistically significant difference was observed on the NCC dataset. The comparison with previous studies also showed that TS-HEFS achieved competitive classification performance under the evaluated experimental setting. Additional validation using the UNSW-NB15 dataset showed that the classification performance of the selected feature subset varied across the evaluated datasets, highlighting that the proposed method did not consistently achieve similar classification performance on datasets with different network traffic characteristics.

Future research may investigate more adaptive feature aggregation strategies and a wider range of feature selection methods to improve the adaptability of the proposed method across different network traffic characteristics. Further evaluation on network intrusion datasets containing more diverse attack categories beyond P2P botnets, including real-world network traffic from operational environments, would provide broader validation of the proposed method.

Acknowledgement: The authors express their sincere gratitude to all members of the Cyber Security Research Group, Net-Centric Computing (NCC) Laboratory, Department of Informatics, ITS, for their continuous support and insightful discussions. The authors declare that AI is used to improve the writing format and grammar.

Funding Statement: This research was funded by the Academic Excellent Research Grant, Department of Informatics, Institut Teknologi Sepuluh Nopember (ITS), Batch 1, Year 2026, under Research Agreement No. 3038/PKS/ITS/2026.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Mohammad Syarifuz Zaim, Tohari Ahmad, Muhammad Aidiel Rachman Putra; methodology, Mohammad Syarifuz Zaim, Tohari Ahmad, Muhammad Aidiel Rachman Putra; formal analysis, Mohammad Syarifuz Zaim, Muhammad Aidiel Rachman Putra; investigation, Mohammad Syarifuz Zaim, Muhammad Aidiel Rachman Putra; writing—original draft preparation, Mohammad Syarifuz Zaim; writing—review and editing, Tohari Ahmad, Muhammad Aidiel Rachman Putra; visualization, Mohammad Syarifuz Zaim; supervision, Tohari Ahmad; project administration, Tohari Ahmad; funding acquisition, Tohari Ahmad. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets are publicly available: The CTU-13 (<https://doi.org/10.1016/j.cose.2014.05.011>), the NCC (<https://doi.org/10.1016/j.dib.2021.107334>), NCC-2 (<https://doi.org/10.1016/j.dib.2022.108628>) and the UNSW-NB15 (<https://doi.org/10.1109/MilCIS.2015.7348942>).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ibrahim WNH, Anuar S, Selamat A, Krejcar O, González Crespo R, Herrera-Viedma E, et al. Multilayer framework for botnet detection using machine learning algorithms. *IEEE Access*. 2021;9:48753–68. doi:10.1109/ACCESS.2021.3060778.

2. Sattari F, Farooqi AH, Qadir Z, Raza B, Nazari H, Almutiry M. A hybrid deep learning approach for bottleneck detection in IoT. *IEEE Access*. 2022;10:77039–53. doi:10.1109/ACCESS.2022.3188635.
3. Wu G, Wang X, Zhang J. PeerG: a P2P botnet detection method based on representation learning and graph contrastive learning. *Comput Secur*. 2024;140:103775. doi:10.1016/j.cose.2024.103775.
4. Kornyo O, Asante M, Opoku R, Owusu-Agyemang K, Partey BT, Baah EK, et al. Botnet attacks classification in AMI networks with recursive feature elimination (RFE) and machine learning algorithms. *Comput Secur*. 2023;135:103456. doi:10.1016/j.cose.2023.103456.
5. Fortune DC, Mathurin SS, Kalita S. HTTP-based peer-to-peer botnet detection using a machine learning bagging classifier. In: *Proceedings of the 2024 2nd International Conference on Disruptive Technologies (ICDT)*; 2024 Mar 15–16; Greater Noida, India. p. 353–9. doi:10.1109/icdt61202.2024.10489499.
6. Baruah S, Borah DJ, Deka V. Detection of peer-to-peer botnet using machine learning techniques and ensemble learning algorithm. *Int J Inf Secur Priv*. 2023;17(1):1–16. doi:10.4018/ijisp.319303.
7. Gong D, Liu Y. A machine learning approach for botnet detection using LightGBM. In: *Proceedings of the 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*; 2022 May 20–22; Changchun, China. p. 829–33. doi:10.1109/CVIDLICCEA56201.2022.9824033.
8. Sunday B, Abdullahi YM, Matemilola AS, Sahabi Y, Abdullahi M. Voting ensemble learning model with dynamic weighting for improved network intrusion detection. *Frankl Open*. 2025;12:100375. doi:10.1016/j.fraope.2025.100375.
9. Alasmary F, Alraddadi S, Al-Ahmadi S, Al-Muhtadi J. ShieldRNN: a distributed flow-based DDoS detection solution for IoT using sequence majority voting. *IEEE Access*. 2022;10:88263–75. doi:10.1109/ACCESS.2022.3200477.
10. Kalakoti R, Nömm S, Bahsi H. In-depth feature selection for the statistical machine learning-based botnet detection in IoT networks. *IEEE Access*. 2022;10:94518–35. doi:10.1109/ACCESS.2022.3204001.
11. Sarhan M, Layeghy S, Moustafa N, Gallagher M, Portmann M. Feature extraction for machine learning-based intrusion detection in IoT networks. *Digit Commun Netw*. 2024;10(1):205–16. doi:10.1016/j.dcan.2022.08.012.
12. Oyelakin AM, G JR. A survey of feature extraction and feature selection techniques used in machine learning-based botnet detection schemes. *VAWKUM Trans Comput Sci*. 2021;9(1):1–7. doi:10.21015/vtcs.v9i1.604.
13. Kavitha G, Elango NM. Genetic algorithm-conditional mutual information maximization based feature selection for bot attack classification in IoT devices. *J Mob Multimed*. 2021;18(1):119–34. doi:10.13052/jmm1550-4646.1816.
14. Manivannan D. Recent endeavors in machine learning-powered intrusion detection systems for the Internet of Things. *J Netw Comput Appl*. 2024;229:103925. doi:10.1016/j.jnca.2024.103925.
15. Alhanaya M, Ateyeh Al-Shqeerat KH. Performance analysis of intrusion detection system in the IoT environment using feature selection technique. *Intell Autom Soft Comput*. 2023;36(3):3709–24. doi:10.32604/iasc.2023.036856.
16. Baruah S, Deka V, Das D, Barman U, Saikia MJ. Enhanced peer-to-peer botnet detection using differential evolution for optimized feature selection. *Future Internet*. 2025;17(6):247. doi:10.3390/fi17060247.
17. García S, Grill M, Stiborek J, Zunino A. An empirical comparison of botnet detection methods. *Comput Secur*. 2014;45:100–23. doi:10.1016/j.cose.2014.05.011.
18. Hostiadi DP, Ahmad T. Dataset for Botnet group activity with adaptive generator. *Data Brief*. 2021;38:107334. doi:10.1016/j.dib.2021.107334.
19. Putra MAR, Hostiadi DP, Ahmad T. Botnet dataset with simultaneous attack activity. *Data Brief*. 2022;45:108628. doi:10.1016/j.dib.2022.108628.
20. Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: *Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS)*; 2015 Nov 10–12; Canberra, ACT, Australia. p. 1–6. doi:10.1109/MilCIS.2015.7348942.
21. Das S, Venugopal D, Shiva S, Sheldon FT. Empirical evaluation of the ensemble framework for feature selection in DDoS attack. In: *Proceedings of the 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*; 2020 Aug 1–3; New York, NY, USA. p. 56–61. doi:10.1109/csccloud-edgecom49738.2020.00019.

22. Hostiadi DP, Atmojo YP, Rueankhong T, Pradipta GA, Ahmad T, Putra MAR. A new approach of optimizing machine learning classification using intersection-weighting feature selection for botnet attack detection. *Int J Intell Eng Syst.* 2025;18(11):677–98. doi:10.22266/ijies2025.1231.42.
23. Budhraj S, Doborjeh M, Singh B, Tan S, Doborjeh Z, Lai E, et al. Filter and wrapper stacking ensemble (FWSE): a robust approach for reliable biomarker discovery in high-dimensional omics data. *Brief Bioinform.* 2023;24(6):bbad382. doi:10.1093/bib/bbad382.
24. Ramadhan IF, Putra MAR, De La Croix NJ, Ahmad T, Kannole VE, Mtonga K, et al. SB-net: a novel spam botnet detection scheme with two-stage cascade learner and ensemble feature selection. *IEEE Open J Commun Soc.* 2025;6:7237–55. doi:10.1109/ojcoms.2025.3603425.
25. Hossain MA, Islam MS. A novel feature selection-driven ensemble learning approach for accurate botnet attack detection. *Alex Eng J.* 2025;118:261–77. doi:10.1016/j.aej.2025.01.042.
26. Velasco-Mata J, González-Castro V, Fernández EF, Alegre E. Efficient detection of botnet traffic by features selection and decision trees. *IEEE Access.* 2021;9:120567–79. doi:10.1109/ACCESS.2021.3108222.
27. Mahajan A, Kaushik B, Rahmani MKI, Banga AS. A hybrid feature selection and ensemble stacked learning models on multi-variant CVD datasets for effective classification. *IEEE Access.* 2024;12:87023–38. doi:10.1109/ACCESS.2024.3412077.
28. Calvo A, Escuder S, Ortiz N, Escrig J, Compastie M. RBD24: a labelled dataset with risk activities using log application data. *Comput Secur.* 2025;150:104290. doi:10.1016/j.cose.2024.104290.
29. Gong H, Li Y, Zhang J, Zhang B, Wang X. A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information. *Eng Appl Artif Intell.* 2024;131:107865. doi:10.1016/j.engappai.2024.107865.
30. Nadeau C, Bengio Y. Inference for the generalization error. *Mach Learn.* 2003;52(3):239–81. doi:10.1023/A:1024068626366.