Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Feature-Centric Framework for Network Traffic Management and Optimization

Aymen Hasan Alawadi[1]

[1]Department of Computer Science, Faculty of Education University of Kufa 54001 Najaf Governorate, Iraq
Corresponding author E-mail: aymen@uokufa.edu.iq

Abstract. *While complex network development and dynamic traffic are continuing to evolve, there has been great optimization and management of network traffic. This paper develops a feature-based approach based on intelligent selection and reduction to improve traffic flow management, optimization of resource utilization, and enhancement of network security. Methods of statistical correlation and clustering are employed to retain only the relevant features, filtering out irrelevant ones to optimize operational efficiency and the decision-making process. Advanced correlation algorithms are proposed in a cross-selection set, such as Spearman and Kendall, besides dendrogram-based clustering analysis. This makes the solution robust for designing and monitoring modern network environments. It can be seen from the experimental results that the proposed framework can remove redundancy and enhance the feature interpretability, scalability, and adaptability to domain-specific network requirements. Furthermore, the framework proved its efficiency in capturing nonlinear feature relationships, making it robust for real-world network applications.*

Keywords: *Network Feature Selection, Spearman correlation, Kendall correlation, Dendrogram clustering*

https://doi.org/10.46649/fjiece.v4.1.25a.26.3.2025

1. INTRODUCTION

Modern network architectures are characterized by increasing complexity and dynamic traffic patterns. Such a challenge presents significant difficulties in effective network management and optimization. The enormous volume of data passing via these networks requires innovative ways to assure network performance, security, and optimal resource allocation. Traditional methods in network management methods, such as port-based classification [1], static traffic analysis [2], and deep packet inspection (DPI) [1] often struggle to fulfill the nature of network traffic due to their dependency on static rules, inability to handle encrypted traffic, and challenges with the high dimensionality of network traffic. Hence, developing more sophisticated techniques that can adapt to these dynamic environments is essential. This paper introduces a feature-centric framework developed to address these challenges by leveraging advanced feature correlation algorithms, and behavioral clustering. The core idea of the proposed framework lies in its emphasis on finding the most relevant network features that can be used for several network design aspects such as understanding network feature relationships, network performance monitoring, prioritizing security efforts, network design and optimization, and future-proofing the network.
This framework is crucial since networks generate massive data, much of which is redundant or irrelevant for certain tasks such as traffic management, resource allocation optimization, or network monitoring.

Hence, this research is motivated by the need for a systematic approach that can effectively identify network traffic features in a structured manner.

The proposed framework builds upon a comprehensive dataset called CIC-IDS2017 [3]. Such a dataset is designed to validate the proposed solutions related to intrusion detection systems (IDSs) and intrusion prevention systems (IPSs). The data collection is 51.1 Giga Bytes in size and contains both benign and real-world network attacks. It provides network traffic analysis data (i.e., HTTP, HTTPS, FTP, SSH, and email protocols) based on the timestamp, source, and destination IP addresses, source and destination ports, protocols, and attacks, all of which are stored in separate CSV (comma-separated values) files.

This research proposes a feature-centric framework to improve network traffic management and handling employing state-of-the-art correlation analysis and clustering techniques. The proposed framework not only is lightweight in computational overhead but also improves the quality of anomaly detection and resource allocation. The key contributions of the paper are as follows:

1. Applying advanced cross-selection correlation analysis methods (Kendall and Spearman) to eliminate redundant features and improve interpretability.
2. Incorporating dendrogram-clustering to visualize the obtained traffic patterns and validate the correlated features.
3. Providing a comprehensive methodology that data streams and new network technologies such as understanding network behaviors, optimizing intrusion detection, monitoring network performance, and future-proofing network infrastructure.

The rest of the paper is organized as follows: First, a literature review of existing works is discussed in section 2, then the methodology of the framework is illustrated in section 3. In section 4, the experimental results and analysis are explained. Finally, section 5 concludes the proposed framework.

2. LITERATURE REVIEW

Traffic classification has widely adopted in research and attacted much attention lately. This section reviews the existing literature on feature selection, network traffic analysis, and anomaly detection, demonstrating how the proposed feature-centric method can improve network performance. In terms of network traffic analysis, several studies emphasized the critical role of feature selection in the dimensionality reduction of network traffic features to enhance the efficiency of machine learning models [4]. Here, the selection process aims to identify the most relevant features that contribute most to accurate classification and anomaly detection, reducing the computation overhead of collecting and computing several features in the classification model [4]. H. Nguyen, et al, in [5] have employed filter-based methods such as correlation-based feature selection (CFS) besides linear programming and linear dependence to find and reduce redundant features by assessing their correlation with the target label and with each other for the KDD CUP'99 IDS dataset. I. Sharafaldin et al. in [3] proposed a method to reduce the redundant features of the UNSW-NB15 dataset using the Pearson correlation coefficient and feature clustering analysis. Wrapper-based methods such as tabu search have used by A. Nazir and R. A. Khan in [4]. The tabu search was employed with a Random Forest to identify the optimal set of features from the UNSW-NB15 dataset, but this method is more computationally expensive. M. A. Bouke et al., introduced BukaGini in [6], which is a Gini index feature selection method to eliminate redundant features enhanced by Random Forest as an ensemble approach. The performance of BukaGini may be affected by the chosen ensemble learning methods, especially in the case when dealing with high-dimensional datasets. Several ensemble learning methods have tested on the CIC-IDS2017 dataset by Y. Zhang et al., in [7]. After evaluating the feature importance utilizing various base approaches, then the XGBoost and ET (Extra Trees) algorithms were used to evaluate the performance of the chosen feature selection methods. However, such methods require careful parameter tuning and are sensitive to overfitting, and computational complexity. E. Jaw et al., introduced a hybrid feature selection method in [8] based on genetic search, rule-based engine, and CfsSubsetEval to

select the most effective features in the CIC-IDS2017 dataset. However, CfsSubsetEval may not perform well in a dataset with high dimensionality, noisy, or irrelevant features. The Pearson correlation coefficient method is employed by P. Chen et al, in [9] to efficiently process large amounts of data and reduce the dimensionality of numerous features. The method included three steps: search, evaluation, and classification, and it includes techniques such as step forward selection, and step backward selection. Spearman and Pearson correlation methods have utilized for security analysis of Internet of Things (IoT) systems by revealing underlying network features. Li et al, [10] employed Spearman and Pearson correlation coefficients to analyze the source and destination ports features and how they relate to distributed denial of service (DDoS) attacks. The results showed that the port features are not associated enough with the attack but rather with protocol types such as TCP and UDP. Similarly, Pires and Mascarenhas introduced in [11] a practical approach for cyber threat detection, focusing on exploratory analysis for feature reduction. Their results showed that the characteristics of a network such as attack type, source port, destination port, and duration of the attack are significant for system vulnerabilities detection. However, the linear relationships assumption between the features can be limited for complicated, evolving, and sophisticated attacks.

Although several studies have taken feature selection into consideration in network traffic, However, they often lack a comprehensive framework that integrates general network functions such as traffic flow control, resource utilization optimization, and network security improvement. This work addresses these gaps by proposing a comprehensive feature extraction and selection framework based on Spearman and Kendall correlation, in addition to dendrogram-based clustering analysis. By capturing the nonlinear relationships between the features, the proposed framework enhances the features' interpretability and adaptability for various network scenarios. This makes the proposed framework a “feature-centric” solution for understanding network behaviors, optimizing intrusion detection, monitoring network performance, and future-proofing network infrastructure.

3. METHODOLOGY

Generally, the network traffic dataset contains several unrelated and redundant features. In this study, the CIC-IDS2017 dataset is used. Such a dataset includes ٨٤ features besides the labeled target of the attack. Furthermore, the dataset incorporates realistic background network traffic generated using the B-Profile system, to simulate the normal behavior of 25 users based on protocols such as HTTP, HTTPS, FTP, SSH, and email [3]. The data was generated and collected for five days, and the traffic features were saved in several CSV files based on the capture day.

Figure 1 illustrates the proposed centric feature framework flowchart. The framework follows a structured process, beginning with dataset preprocessing, followed by Kendall and Spearman correlation methods to estimate the features’ pair relationships. Dynamic thresholds for both correlation methods are applied to determine the feature significance. The thresholds check that the selected features' correlation significance meets the minimum set criteria, if not, the threshold is then adjusted accordingly. When the threshold is satisfied, strong correlations are identified. Then, cross-selection further refines the feature sets. Finally, dendrogram-based clustering is performed to visualize the feature grouping, ensuring that the correlation features are optimized for further analysis.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

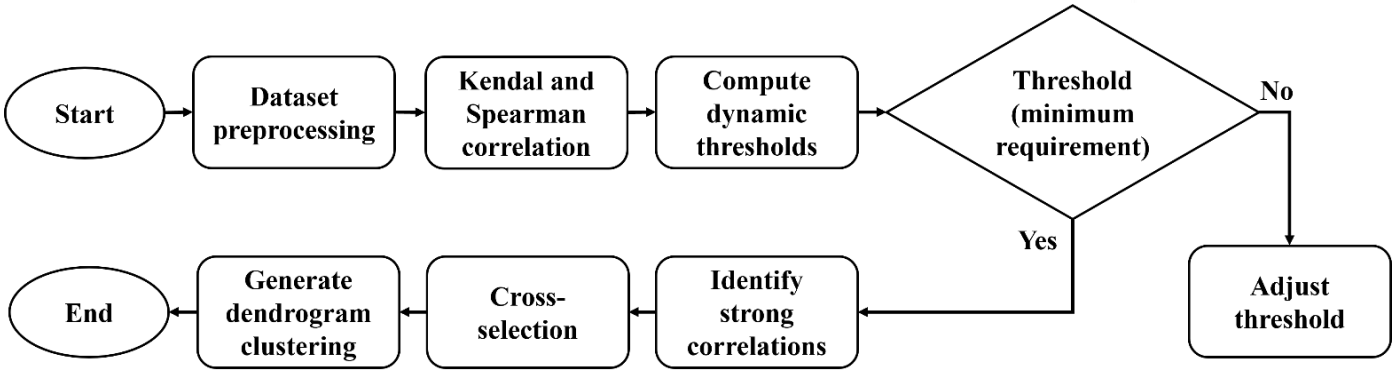

Fig. 1. The centric feature framework flowchart.

The main purpose of the proposed method is to analyze the IDS2017 dataset and identify any correlated features related to traffic flow control, resource utilization optimization, and network security improvement. To analyze the entire dataset, every CSV file of the traffic day has combined into one CSV file. The resulting file has 3,119,345 entities (rows). The input of the proposed model is the complete dataset, and the output will be the highly correlated feature sets to provide valuable insights, which can significantly aid network analysts in understanding network behavior and optimizing network security and performance. Therefore, the dataset features have analyzed as presented in Table 1. Note that the sub-features such as (Max, Min, and Std (Standard Deviation) related to the feature (Fwd Packet Length) have removed, instead, only mean values have been kept. The reasons for keeping only the mean values are as follows. Firstly, to prevent computational overhead, as it is faster to compute and takes less memory. Secondly, min values are near zero, particularly for measurements such as packet size or inter-arrival times [10], and thus are less helpful in differentiating between traffic types. Third, std is highly correlated with mean and thus is redundant for high-level network traffic analysis.

Table 1. CIC-IDS2017 dataset features and their meaning.

Feature	Meaning
Flow Duration	The total time that a network flow lasts.
Total Fwd Packets	The total number of packets sent in the forward direction (source to destination).
Total Backward Packets	The total number of packets sent in the backward direction (destination to source).
Flow Bytes/s	The throughput of the flow is measured in bytes per second.
Flow Packets/s	The rate at which packets are transmitted during the flow.
Total Length of Fwd Packets	The cumulative size (in bytes) of all forward packets in the flow.
Total Length of Bwd Packets	The cumulative size (in bytes) of all backward packets in the flow.
Fwd Packet Length Mean	The average length of forwarding packets.
Bwd Packet Length Mean	The average length of backward packets.
Flow IAT Mean	The average inter-arrival time between consecutive packets.
Flow IAT Std	The standard deviation of inter-arrival times indicates variability in packet arrival timings.
Fwd Header Length	The average length of the header in forwarding packets.
Bwd Header Length	The average length of the header in backward packets.
SYN Flag Count	The number of SYN flags observed indicates the frequency of connection initiation attempts.
PSH Flag Count	The number of PSH flags can indicate application-layer push activity and potential attacks.
ACK Flag Count	The number of ACK flags, representing the acknowledgment of packet receipt.
FIN Flag Count	The number of FIN flags, used to signal the end of a TCP connection.
Down/Up Ratio	The ratio of downward to upward traffic can reveal imbalances in data flow.
Avg Fwd Segment Size	The average size of data segments (amount of data carried in a single packet) is sent in the forward direction.
Avg Bwd Segment Size	The average size of data segments sent in the backward direction.
Subflow Fwd Packets	The number of packets in forward subflows provides a finer granularity of traffic volume.
Subflow Bwd Packets	The number of packets in backward subflows.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Init_Win_bytes_forward	The initial TCP window size in bytes for forwarding traffic is related to transmission control.
Init_Win_bytes_backward	The initial TCP window size in bytes for backward traffic.
act_data_pkt_fwd	The number of actual data packets sent in the forward direction.
Active Mean	The average period during which a flow is active (transmitting data).
Active Std	The variability (standard deviation) of active periods within flows.
Idle Mean	The average period during which a flow is idle (not transmitting data).
Idle Std	The variability (standard deviation) of idle periods within flows.
Label	The categorical label indicates the type of traffic (e.g., benign, various attacks).

3.1 Preprocessing Steps

The proposed method includes various preprocessing steps to prepare the dataset for correlation and clustering:

1. **Type conversion:** All the object types of the columns, except the label, are converted into numeric type. Such conversion enables subsequent correlation calculations are performed on numeric data.
2. **Label handling:** The label column is cleaned by dropping rows with missing labels and replacing any literal "nan" strings with actual NaN values. Then, the label values are encoded using LabelEncoder to convert the attack types into numerical values.
3. **Missing and infinite values:** To avoid infinite values resulting from division operations (e.g., total_bytes / duration) in case of duration = 0, such values are replaced with NaN. Furthermore, rows (or columns) with excessive missing data are dropped. Then, these NaN will be dropped. Note that the NaN value is recognized as a missing value in Python data science libraries such as Pandas, NumPy, and most machine learning frameworks.
4. **Resulting dataset:** The cleaned and pre-processed dataset is then filtered to keep only the specific set of features ready to be analyzed.

3.2 Correlation Analysis

In this paper, two statistical methods, Spearman, and Kendall correlation, have employed to evaluate the relationships among dataset features. These methods eliminate redundant features by correlating the feature pairs and reducing them to only one representative feature. Furthermore, such analysis provides deeper insights into the dataset, by exploring the behavior patterns, data trends, possible relationships, and anomalies [11]. Hence, researchers, data analysts, and network administrators can better understand network traffic and focus on certain features of their applications.

- **Spearman correlation:**
  It measures the relationships between every dataset pair based on ranked data. Unlike other correlation methods such as Pearson correlation, the Spearman method is less sensitive to data noise and outliers, and it is used to assess monotonic relationships (linear and nonlinear), making it popular in psychology, social science, and economics [11]. However, to apply the Spearman correlation to the CIC-IDS2017 dataset, the following steps will be applied:

1. Rank the numerical values of each feature.
2. Calculate the rank differences for each feature pair.
3. Calculate the squared rank differences and then sum them up.
4. Apply the Spearman equation and obtain the correlation values.
5. Visualize the correlation matrix that maps the feature relationships.
6. Define the highly correlated features and remove redundant ones.

Spearman correlation is computed using a coefficient (ρ) that measures the strength of a monotonic relationship between two certain features. The coefficient is calculated using Equation (1) [11]:

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

$$p = 1 - \frac{6\sum di^2}{n(n^2-1)} \qquad (1)$$

Where d_i represents the rank difference between the corresponding feature values (x_i and y_i), and n is the number of observed features. Note that the correlation coefficient scale is measured within the range [−1,1], where -1 represents the weakest correlation and 1 is the strongest one. However, applying the Spearman equation to the CIC-IDS2017 dataset will reveal both the linear and nonlinear relationships between the specified features. This type of analysis helps to guarantee that one feature represents two or more features so that the total number of features is eliminated.

- **Kendall's correlation**:

  Kendall correlation is employed to ensure a more reliable feature selection procedure by ensuring robust dependency analysis while mitigating sensitivity to outliers [12], especially in a dynamic and noisy network environment. Furthermore, Kendall's correlation provides a better probability-based measure of correlation by comparing concordant and discordant pairs [13]. Generally, Kendall's correlation is calculated as tau (τ) which measures the correlation between two dependencies based on the number of concordant and discordant pairs, as presented in Equation 2.

$$\tau = \frac{(C-D)}{\frac{1}{2}n(n-1)} \qquad (2)$$

  Where *C* indicates the number of concordant groups (i.e., the pairs with the same order after ranking), *D* is the number of discordant groups (i.e., the pairs that go opposite in the order), and *n* is the number of observed entities. However, the Kendal correlation applied to the dataset with the following steps:

1. Define the possible feature groups.
2. Calculate concordant and discordant pairs.
3. Apply Kendall's equation to calculate the correlation values (range [−1, 1]).
4. Build the correlation matrix and analyze relationships.
5. Identify the highly correlated features and remove the redundant ones.

- **Behavioral clustering**:

  A dendrogram-based clustering method is a type of hierarchical clustering mainly used to explore features' relationships and reduce the number of redundant ones. The root node of the hierarchy denotes the entire dataset, whereas the leaf node reflects a single feature [14]. This step helps identify potential anomalies in the dataset. Dendrogram-based clustering provides a hierarchical visualization of the relationships, allowing an effective categorization of the network behaviors that support real-time analysis. The advantage of using clustering is using the simple distanced measure, such as the Euclidean distance, between the correlated entities of the CIC-IDS2017 dataset. As clusters merge, the inter-cluster distance is monotone, ensuring that the height of each merger in the dendrogram represents increasing dissimilarity. This analysis provides a deep insight into how the features are closely linked to each other so that one can represent the other effectively. The results of the clustering produce a group of clusters from the linked features, ready to be analyzed.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed feature-centric framework thorough validation through extensive experiments on the CIC-IDS2017 dataset, which assessed the proposed feature correlation, clustering accuracy, and overall system performance. Firstly, the observations of using Spearman and Kendall correlation-based selection will be presented in the following:

4.1 Feature Selection Effectiveness

Figures 2 and 3 show the heatmap for the conducted Spearman correlation, ranging from -0.75 for the weakest correlation up to 1 for the strongest correlation. The correlation between each feature represented by colored squares. The color intensity of each square represents the correlation coefficient between tow features (deep blue for strong negative correlation [0.75], white for no correlation [0.00], and deep red for strong positive corralito [1.00] for Sperman and Kendall). The figures reveal the relationships of the features and how they are related to each other, especially within traffic flow metrics. For instance, there is a very strong positive correlation between 'Total Fwd Packets' and 'Subflow Fwd Packets' in both figures. After applying correlation-based reduction, the improved collection preserved only the most important features while removing redundant ones, resulting in a more compact and meaningful feature set.

To best understand the relationship of the features and indicate the correlation threshold values, both Spearman and Kendall correlations are statistically summarized in Table 2. Note that the variation in correlation numbers between 808 and 804 is due to the different approaches used and the robustness of the Kendall and Spearman correlation estimation. The analysis also reveals that the Spearman method has a wide coefficient range, so the results are sensitive to monotonic but nonlinear relationships compared with the Kendall method.

Depending on static thresholds (0.3 and 0.7 as suggested by [10]) might not help identify significant positive and negative correlations. Hence, dynamic thresholds will be estimated based on the 90th percentile, ensuring only the strongest correlated features are considered. Such a method is adaptable and more robust against varying data scales and distributions, and it reduces the impact of extreme values by focusing on the top 10% of correlations. Furthermore, fixed thresholds (e.g., 0.7) may not suit all datasets. For a given correlation method (Kendall or Spearman), the computation will be according to Equation 3:

$$Dynamic\_threshold = P90(C) \quad (3)$$

Where $P90(C)$ = 90th percentile of the correlation values (C), and C is the array of all computed correlation coefficients between feature pairs.

To make sure that the chosen threshold is too small, which could lead to false positives, the process of the threshold calculations included a minimum threshold (*min_threshold* =0.3). This value acting as a floor value to check the process with the *max(dynamic_threshold, min_threshold)* function ensures that the effective threshold is never below 0.3. Additionally, *min_threshold* serves as the minimum boundary to identify the strong negative correlations. In this paper, the negative correlations are used to understand the inverse relationship between the correlated features.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

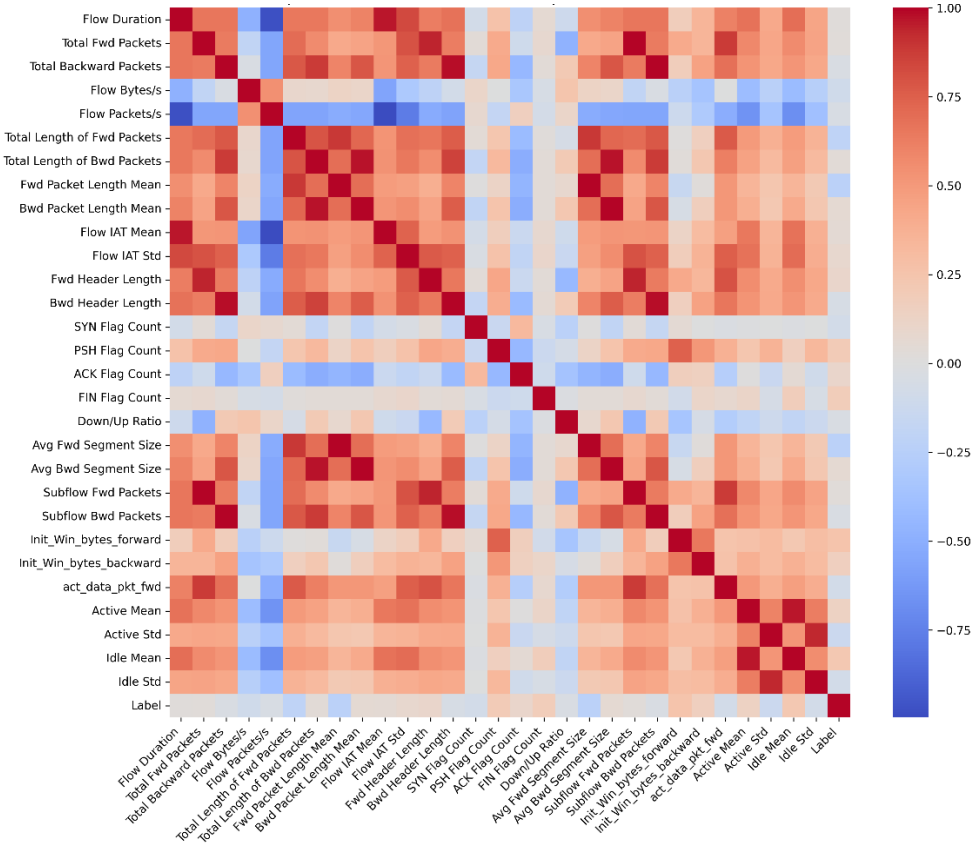

Fig. 2. Spearman correlation heatmap for the selected features of the CIC-IDS2017 dataset.

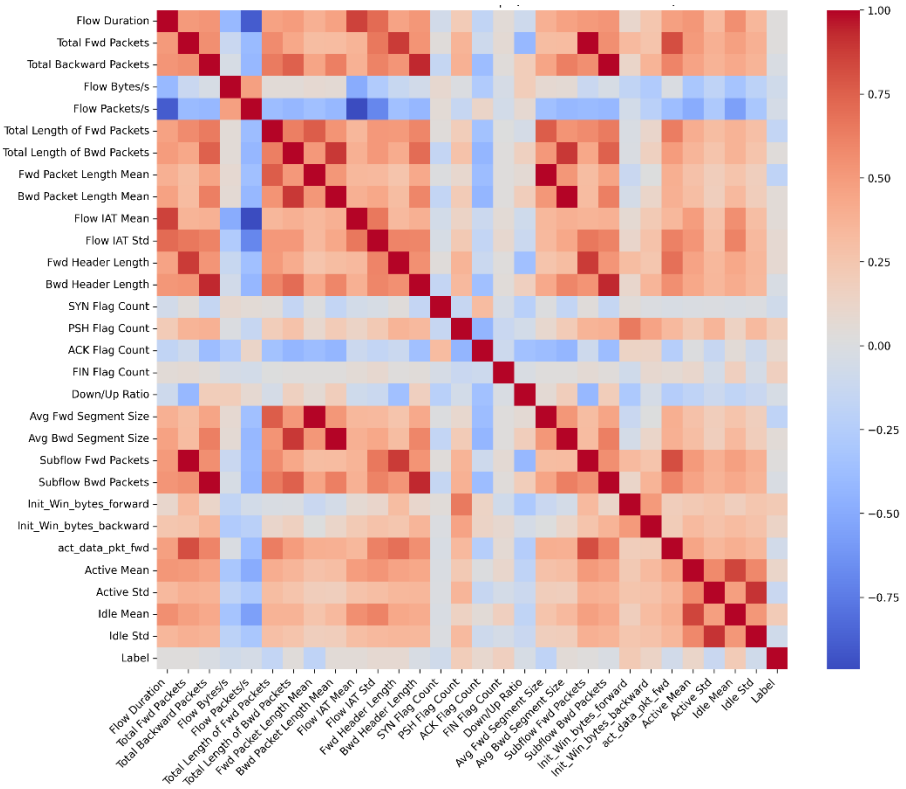

Fig. 3. Kendall correlation heatmap for the selected features of the CIC-IDS2017 dataset.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Table 2. Spearman and Kendall correlations analysis.

Statistic	Kendall Correlation	Spearman Correlation
Count	808	804
Mean	0.2058	0.2445
Standard Deviation	0.3548	0.4145
Minimum	-0.9584	-0.9951
25th Percentile	-0.0718	-0.0782
Median (50%)	0.2734	0.3543
75th Percentile	0.4557	0.5449
Maximum	1.0000	0.9954

The calculation shows that the significant positive values for the Kendall threshold equals 0.6733 and the Spearman threshold equals 0.7819. Similarly, the Kendall strong negative correlations are (<-0.6733), and (<-0.7819) in the case of Spearman correlation. Tables 3 and 4 show the correlation analysis conducted using the Spearman and Kendall methods. The results reveal multiple instances of strong positive and negative correlations for the selected features of the dataset. To improve the reliability of feature selection, a cross-validation step has applied to the Spearman and Kendall correlations, keeping only consistently significant pairs from both approaches. This method reduces potential biases inherent in each correlation and enhances the robustness of the selected features, as demonstrated in Table 5. These correlations provide deep insights into feature redundancy, which is important for feature selection and dimensionality reduction. In the following, the strong positive and negative correlations will be discussed:

4.1.1. Cross-Selected Correlated Feature Pairs

The correlated features were structured in pairs that appeared in both Spearman and Kendall correlations with correlation values larger than the obtained Spearman (>0.7819) and Kendall (>0.6733) thresholds. Each pair will be mentioned with the related correlation coefficients (Spearman (spr_cof) and Kendall (ken_cof)).

1. **Total Fwd Packets and Subflow Fwd Packets (spr_cof and ken_cof= 1)**, and **Total Backward Packets and Subflow Bwd Packets (spr_cof and ken_cof= 1)**: These pairs show that the subflow Fwd Packets can directly represent the total packet counts, therefore keeping both features may be unnecessary. From these pairs, the retained feature would be the Subflow Fwd packets, since it can be generalized to different traffic types better than the forward and backward packets features**.** For instance, the subunits of the flow become more manageable, making it adaptive to different flow durations and behaviors.
2. **Fwd Packet Length Mean and Avg Fwd Segment Size (spr_cof and ken_cof= 1)**, and **Bwd Packet Length Mean and Avg Bwd Segment Size (spr_cof and ken_cof= 1)**: These pairs indicate that the segment size features (backward or forward), and packet length features are equal in both forward and backward directions. Hence, the retained features from this group are the Fwd Packet Length Mean. Such a feature produces meaningful information about different network flows whether TCP or UPD.
3. **Active Std and Idle Std (spr_cof=0.9954 and ken_cof= 0.9546), Active Mean and Idle Mean (spr_cof=0.9859 and ken_cof=0.9013)**: These pairs show that the active and idle states are tightly coupled, which indicates the network activities are periodically being active and idle. The retained features are the active mean and active Std. These features are important to present the active time of the network services rather than the idle status.
4. **Flow Duration and Flow IAT Mean (spr_cof=0.9579 and ken_cof=0.8442)**: This pair indicates that the flow duration feature is strongly related to the inter-arrival time and flow duration.

Therefore, one of them is sufficient to represent the time-based characteristics of network flows. But Flow IAT Mean will be kept. Such a feature provides significant information about the average time gap between consecutive packets, which will help to analyze the traffic behavior.

5. **Total Length of Bwd Packets and Bwd Packet Length Mean (spr_cof=0.9374 and ken_cof=0.8396)**, and **Total Length of Bwd Packets and Avg Bwd Segment Size (spr_cof=0.9374 and ken_cof=0.8396)**: This group of pairs indicates high redundancy in capturing the backward packet statistics (segment size, packet length). Again, this group confirms that the Bwd Packet Length Mean is important, hence, it will be retained.
6. **Total Fwd Packets and Fwd Header Length (spr_cof=0.9341 and ken_cof=0.8726)**, and **Fwd Header Length and Subflow Fwd Packets (spr_cof=0.9341 and ken_cof=0.8726)**, and **Fwd Header Length and act_data_pkt_fwd (spr_cof=0.8258 and ken_cof=0.7357):** These pairs indicate that when the number of forwarded packets increases, the total number of forwarded header length also increases. Similarly, subflow forward packets and forward header length present a strong dependency on the forwarded actual data packets. Therefore, the redundant features could be safely removed without affecting the data meaning or model prediction. The kept features that come from this group is the act_data_pkt_fwd feature since it represents the actual sent data packets excluding the control packets. Subflow Fwd Packets has already selected in group 1 and Fwd Header Length is redundant and does not provide extra predictive information.
7. **Total Fwd Packets and act_data_pkt_fwd (spr_cof=0.9119 and ken_cof=0.8704)**, and **Subflow Fwd Packets and act_data_pkt_fwd (spr_cof=0.9119 and ken_cof=0.8704):** These pairs suggest that keeping all three features could be unnecessary, hence Total Fwd Packets feature might be the most informative feature to retain.
8. **Total Backward Packets and Total Length of Bwd Packets (spr_cof=0.8621 and ken_cof=0.7478)**, and **Total Length of Bwd Packets and Subflow Bwd Packets (spr_cof=0.8621 and ken_cof=0.7478)**, and **Total Length of Bwd Packets and Bwd Header Length (spr_cof=0.8227 and ken_cof=0.6794):** This group of correlation indicates that the total length of backward packets is directly connected to the total number of backward packets, subflow backward packets, and the backward header length. Hence, the total Backward Packets feature can be used to represent them efficiently**.**
9. **Total Backward Packets and Total Length of Fwd Packets (spr_cof=0.8621 and ken_cof=0.6800)**, and **Total Length of Fwd Packets and Subflow Bwd Packets (spr_cof=0.8012 and ken_cof=0.6800):** In this group, the forward and backward packet lengths are closely related. This could mean several flows operate symmetrically, with equal forward and backward traffic patterns. Such redundancy suggests that using either the forward or backward packet length (but not both) could be sufficient. This behavior is expected in normal communications, such as web browsing or real-time applications, where responses are proportionate to the data sent. However, in DDoS attacks, this behavior could be disrupted when the attacker floods the destination with a massive number of forward packets but receives fewer responses. Hence, the Total Length of Fwd Packets feature will be kept, since it provides meaningful information for traffic analysis and attack detection.

3.1.2. Spearman Strong Negative Correlations

1. **Flow Packets/s and Flow IAT Mean (spr_cof= -0.9951 and ken_cof= -0.9584)**: This relationship suggests that when the inter-arrival time increases, the transmission rate decreases, which is considered normal network traffic behavior.

2. **Flow Duration and Flow Packets/s (spr_cof= -0.9749 and ken_cof= -0.8854)**: This correlation shows that longer flow durations mostly have lower packet transmission rates. This is highly expected in burst traffic attacks.

From the above pairs, Flow Packets/s could be retained, since it captures the rate of packet transmission.

Table 3: Strong positive correlations (>0.7819) and Strong negative correlations (<-0.7819) in the Spearman method

No.	Feature Pair	Correlation Value
1.	Total Fwd Packets and Subflow Fwd Packets	1
2.	Total Backward Packets and Subflow Bwd Packets	1
3.	Fwd Packet Length Mean and Avg Fwd Segment Size	1
4.	Bwd Packet Length Mean and Avg Bwd Segment Size	1
5.	Active Std and Idle Std	0.9954
6.	Active Mean and Idle Mean	0.9859
7.	Total Backward Packets and Bwd Header Length	0.9627
8.	Bwd Header Length and Subflow Bwd Packets	0.9627
9.	Flow Duration and Flow IAT Mean	0.9579
10.	Total Length of Bwd Packets and Bwd Packet Length Mean	0.9374
11.	Total Length of Bwd Packets and Avg Bwd Segment Size	0.9374
12.	Total Fwd Packets and Fwd Header Length	0.9341
13.	Fwd Header Length and Subflow Fwd Packets	0.9341
14.	Total Fwd Packets and act_data_pkt_fwd	0.9119
15.	Subflow Fwd Packets and act_data_pkt_fwd	0.9119
16.	Total Backward Packets and Total Length of Bwd Packets	0.8621
17.	Total Length of Bwd Packets and Subflow Bwd Packets	0.8621
18.	Fwd Header Length and act_data_pkt_fwd	0.8258
19.	Total Length of Bwd Packets and Bwd Header Length	0.8227
20.	Total Length of Fwd Packets and Total Length of Bwd Packets	0.8163
21.	Total Length of Fwd Packets and act_data_pkt_fwd	0.8022
22.	Total Backward Packets and Total Length of Fwd Packets	0.8012
23.	Total Length of Fwd Packets and Subflow Bwd Packets	0.8012
24.	Total Fwd Packets and Flow IAT Std	0.7998
25.	Flow IAT Std and Subflow Fwd Packets	0.7998
26.	Flow Packets/s and Flow IAT Mean	-0.9951
27.	Flow Duration and Flow Packets/s	-0.9749

Table 4: Strong positive correlations (>0.6733) and Strong negative correlations (<-0.6733) in the Kendall method

No.	Feature Pair	Correlation Value
1.	Total Fwd Packets and Subflow Fwd Packets	1
2.	Fwd Packet Length Mean and Avg Fwd Segment Size	1
3.	Total Backward Packets and Subflow Bwd Packets	1
4.	Bwd Packet Length Mean and Avg Bwd Segment Size	1
5.	Active Std and Idle Std	0.9546
6.	Total Backward Packets and Bwd Header Length	0.9129
7.	Bwd Header Length and Subflow Bwd Packets	0.9129
8.	Active Mean and Idle Mean	0.9013
9.	Total Fwd Packets and Fwd Header Length	0.8726
10.	Fwd Header Length and Subflow Fwd Packets	0.8726
11.	Total Fwd Packets and act_data_pkt_fwd	0.8704
12.	Subflow Fwd Packets and act_data_pkt_fwd	0.8704

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

13.	Flow Duration and Flow IAT Mean	0.8442
14.	Total Length of Bwd Packets and Bwd Packet Length Mean	0.8396
15.	Total Length of Bwd Packets and Avg Bwd Segment Size	0.8396
16.	Total Length of Bwd Packets and Subflow Bwd Packets	0.7478
17.	Total Backward Packets and Total Length of Bwd Packets	0.7478
18.	Fwd Header Length and act_data_pkt_fwd	0.7357
19.	Active Mean and Active Std:	0.7140
20.	Active Mean and Idle Std	0.6930
21.	Total Backward Packets and Total Length of Fwd Packets	0.6800
22.	Total Length of Fwd Packets and Subflow Bwd Packets	0.6800
23.	Active Std and Idle Mean	0.6798
24.	Total Length of Bwd Packets and Bwd Header Length	0.6794
25.	Idle Mean and Idle Std	0.6784
26.	Flow Packets/s and Flow IAT Mean	-0.9584
27.	Flow Duration and Flow Packets/s	-0.8854

Table 5: The cross-selected and highly correlated pairs from Spearman and Kendall methods

No.	Cross-selected feature pair	Kendall coefficient	Spearman coefficient
1.	Flow Duration - Flow IAT Mean	0.8442	0.9579
2.	Total Fwd Packets - Subflow Fwd Packets	1	1
3.	Total Length of Bwd Packets - Bwd Packet Length Mean	0.8396	0.9374
4.	Fwd Packet Length Mean - Avg Fwd Segment Size	1	1
5.	Total Backward Packets - Total Length of Bwd Packets	0.7478	0.8621
6.	Bwd Packet Length Mean - Avg Bwd Segment Size	1.0000	1.0000
7.	Total Backward Packets - Total Length of Fwd Packets	0.6800	0.8012
8.	Bwd Header Length - Subflow Bwd Packets	0.9129	0.9627
9.	Total Fwd Packets - Fwd Header Length	0.8726	0.9341
10.	Total Length of Bwd Packets - Bwd Header Length	0.6794	0.8227
11.	Total Backward Packets - Bwd Header Length	0.9129	0.9627
12.	Total Length of Fwd Packets - Subflow Bwd Packets	0.6800	0.8012
13.	Total Length of Bwd Packets - Avg Bwd Segment Size	0.8396	0.9374
14.	Subflow Fwd Packets - act_data_pkt_fwd	0.8704	0.9119
15.	Fwd Header Length - Subflow Fwd Packets	0.8726	0.9341
16.	Fwd Header Length - act_data_pkt_fwd	0.7357	0.8258
17.	Total Fwd Packets - act_data_pkt_fwd	0.8704	0.9119
18.	Total Length of Bwd Packets - Subflow Bwd Packets	0.7478	0.8621
19.	Active Mean - Idle Mean	0.9013	0.9859
20.	Total Backward Packets - Subflow Bwd Packets	1	1
21.	Active Std - Idle Std	0.9546	0.9954
22.	Flow Duration - Flow Packets/s	-0.8854	-0.9749
23.	Flow Packets/s - Flow IAT Mean	-0.9584	-0.9951

4.2 Correlation and Clustering-Based Feature Selection

Dendrogram clustering integrated both Spearman and Kendall correlation analyses to estimate the robustness of the proposed selection method. The clustering method helps to provide a visualized global structure that validates the features' relationship. Furthermore, the clustering analysis reveals the dataset features' unique characteristics or different behavior. Figure 4 and likely Figure 5 display the dendrogram hierarchical clustering generated from the Spearman and Kendall correlation matrix. This clustering

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

visualizes groups of features based on their similarity. For example, features with more similarity are joined together by shorter vertical lines from the merge point. In contrast, the height of the merge point represents the dissimilarity between feature clusters. For example, 'Total Fwd Packets' and 'Subflow Fwd Packets' in both correlations are connected by a short line represents a high similarity. Several key observations have emerged from the analysis, including:

- **The blue line (Highest level):** The blue line in both clustering figures indicates the highest hierarchical cluster in the dendrogram at a broad level. The datasets have two main groups of features that have weak correlations with each other.
- **The green line (Intermediate-level cluster):** This line represents a sub-cluster within the large hierarchy so that such features are closely related but still have high hierarchy value, as appeared in Figure 3 with features Flow Packets/s, Flow Bytes/s, Down/Up Ratio.
- **The orange line (Lowest level, strongest correlation):** Such group represents the highly correlated features that can be represented by a single representative feature without losing information. For instance, Total Fwd Packets, Subflow Fwd Packets, and act_data_pkt_fwd, which also have lower distance values, suggest that these features should be examined for redundancy.
- **The red line (Outliers):** This group represents features that are less correlated with other features of the dataset, meaning such features provide unique information, for example, SYN Flag Count and ACK Flag Count. These features are likely to indicate attack-related behaviors. Therefore, such features should be carefully analyzed because they are related to identifying DDoS attacks.
- **Feature groupings and redundancy:** The clustering results confirm the pairs correlation and cross-pair selection results obtained from Spearman-Kendall analysis such as Total Backward Packets and Total Length of Fwd Packets (spr_cof=0.8621 and ken_cof=0.6800), and Total Length of Fwd Packets and Subflow Bwd Packets (spr_cof=0.8012 and ken_cof=0.6800). Such pairs are closely grouped in both dendrogram figures. Hence, such clustering further validates their strong interdependence and redundancy.

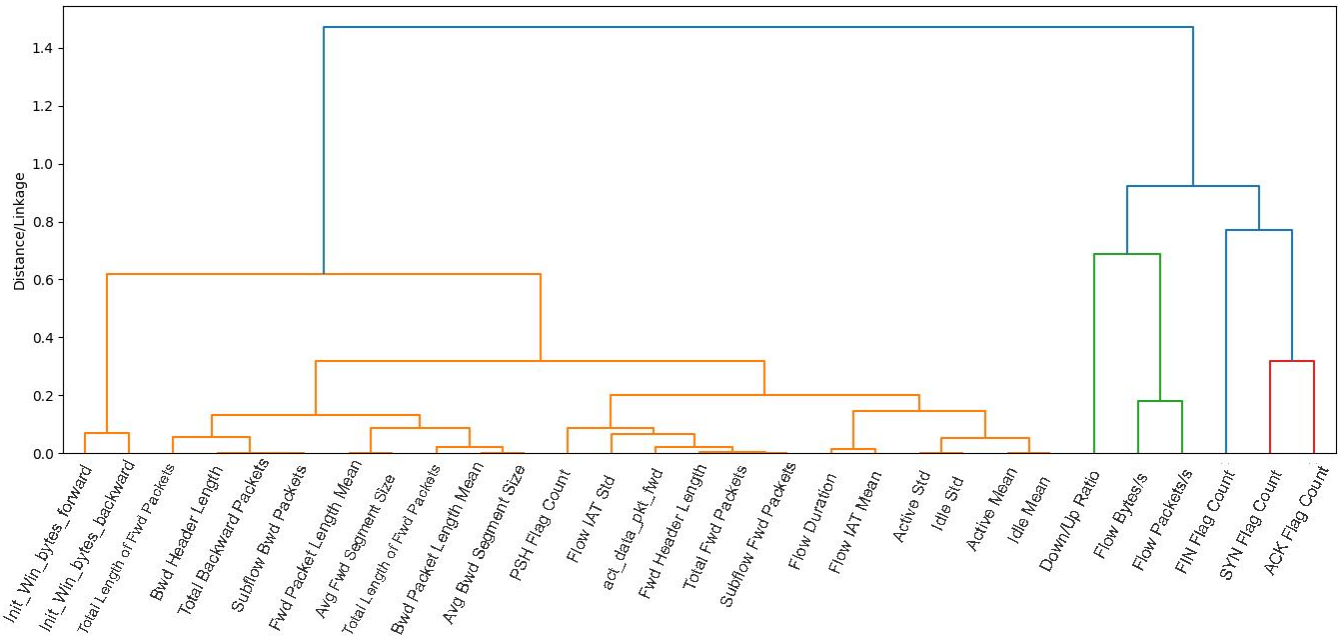

Fig. 4. Dendrogram clustering for the Spearman correlation.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

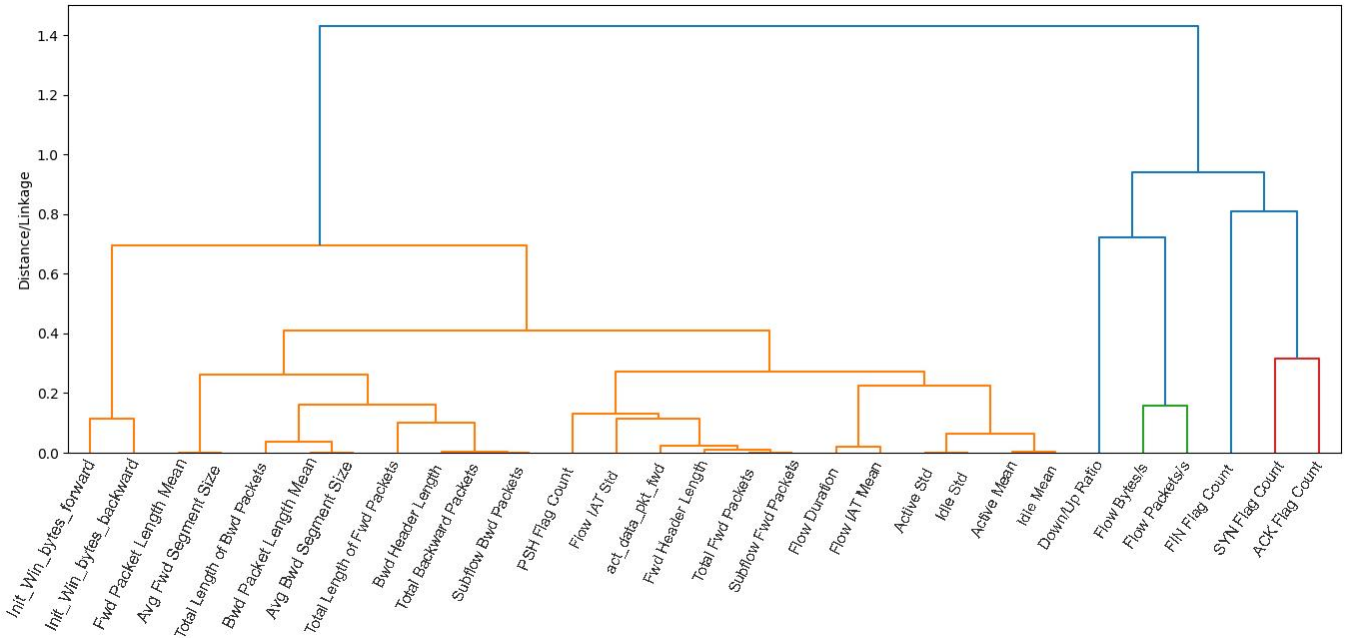

Fig. 5. Dendrogram clustering for the Kendall correlation.

In both dendrograms (Kendall and Spearman), longer branch lengths indicate that the correlation is more sensitive to nonlinear dependencies. Moreover, the dendrograms of both cases reveal that certain features (pairs) appear together in one but not the other, which confirms differences in how Kendall and Spearman measure similarity.

4.2 Implications for Feature Selection

In this section, the correlated pairs will be analyzed and indicate the redundant and final retained features. Based on the aforementioned analysis, the retained features are (Subflow Fwd packets, Fwd Packet Length Mean, Bwd Packet Length Mean, active mean, active Std, Flow IAT Mean, act_data_pkt_fwd, Total Fwd Packets, Total Length of Fwd Packets, Flow Packets/s, SYN Flag Count and ACK Flag Count). The removed highly correlated features help to represent only the core network behaviors, removing others that duplicate the same information.

4.3. Potential Applications for the Selected Features

The results of the correlation and clustering analysis provide valuable insights that can significantly aid network analysts in several terms in understanding different network behaviors besides optimizing network monitoring, security, and performance, as follows:

1. Understanding Network Behaviors: Understanding the network traffic features helps to identify the dependencies and hence make informed decisions regarding feature selection for network security and performance monitoring. The following features might be included in any network model designed for such a task:

- **Flow IAT Mean:** such a feature helps in understanding network traffic behavior so network analysts can relate it to the network environment and traffic pattern.
- **Active Mean:** measures the average transmission periods and shows whether the traffic follows a steady varies between bursts and inactivity.

- **Active Std:** This feature captures only the high values of abnormal traffic patterns made mainly by network congestion**.**
- **Total Fwd Packets:** This feature provides information about the overall volume of the transmission packets, helping in analyzing sent data patterns.
- **Total Length of Fwd Packets:** such a feature provides different information from Total Fwd Packets, to distinguish between small and large flows based on the number of carried packets.
- **Flow Packets/s:** This feature helps in identifying traffic transmission pace (low rate or high rate).

2. Optimizing Intrusion Detection: Generally, IDS analyzes the network traffic to identify the attack signatures and other suspicious behaviors. Detecting attacks such as DDoS attacks, and brute force requires analyzing flows features, such as:

- **Flow IAT Mean**: Used to indicate suspicious activities such as low or high consecutive packets.
- **Flow Packets/s**: used to identify malicious traffic such as DDoS and flood attacks.
- **Active Std**: A high value (spike) indicates a malicious activity pattern; a low value could indicate other stealthy attacks.
- **Total Fwd Packets**: A large volume of sent packets could indicate aggressive kinds of attacks such as port scanning.
- **SYN Flag Count and ACK Flag Count**: In a TCP connection, sending high SYNs (synchronize) volumes without receiving the same amount of ACKs (acknowledgment) indicates SYN flood attacks.

3. Network Performance Monitoring: Maintaining high network availability, low latency, and high throughput is critical to fulfilling the service-level agreements (SLAs). Therefore, network analysts should track different features to optimize the network performance and detect potential failures before impacting the service level.

- **Flow Duration:** Indicates how long- or short-lived flows remain in the network. Such information contributes to network congestion**.**
- **Flow IAT Mean:** To show the information about the time gap between the arrived packets, to help in identifying latency issues.
- **Flow Packets/s:** It allows real-time evolution of the network traffic.
- **Total Length of Fwd Packets:** It is important for monitoring the consumed bandwidth of the network.
- **Active Std:** It highlights the user activity, and load balancing problems, hence helps in network resource allocation.
- **SYN Flag Count and ACK Flag Count:** Help to identify connection failures due to service issues.

4. Future-Proofing the Network: Future-proofing strategy ensures that the network remains scalable, robust, and adaptable to future events such as attacks and new traffic patterns (sudden increasing loads), by analyzing the following features:

- **Flow IAT Mean:** It is used to predict the long-term pattern of the network traffic; such a feature is important for designing an efficient traffic management strategy.
- **Flow Packets/s:** This feature provides anticipation about the network throughput rates; it can be related to the future scalability needs of the network applications' demands.
- **Active Std:** This feature identifies the maximum observed irregular active time patterns; this behavior is crucial for optimizing resource allocation.
- **Total Length of Fwd Packets:** Monitoring this feature provides information about the maximum volume of sent data, helping in planning the maximum network capacity for future demands.

- **SYN Flag Count and ACK Flag Count:** Help to understand the evolving of connection behaviors, an increasing number of SYN messages over time indicates higher connection requests, thus, the session mechanism must be improved.
- **Bwd Packet Length Mean:** It helps to understand the reverse traffic pattern behavior and make sure the bidirectional flows are always balanced; hence, the congestion is minimized.

These models can be utilized for network monitoring or analysis using machine learning algorithms or heuristic solutions. However, by focusing on strongly related features and eliminating redundancies, the network analyst can build more effective and accurate models. Furthermore, understanding flow characteristics and different network behaviors can provide deeper insights into network designing, planning, and enabling proactive defending strategies.

4.4. Comparative Analysis of Feature Selection Approaches

Feature selection is a significant aspect of network management and security optimization. Traditional methods used for correlation analysis reveal network feature relationships but struggle to address the requirements of current networking paradigms. This section compares the proposed model with existing correlation-based methodologies to highlight its advantages.

Table 6: A comparison with existing works.

Aspect	**Proposed Work**	**Li et al, [10] and Pires and Mascarenhas [11]**
Feature Selection	Pairwise cross-correlation based on Spearman, Kendall, and dendrogram clustering	Pearson and Spearman
Analysis Approach	Hybrid approach optimization using correlation and clustering	Exploratory analysis and correlation-based
Objective	Provide interpretable measures of the network feature selection and removal that allow network experts to enhance traffic flow management, optimize resources, and improve security through feature reduction.	Understand relationships between network features and cyber-attacks.
Scalability and adaptability	It is more scalable to cover different aspects of modern network environments, such as (IoT, cloud, and edge), since the network traffic is highly nonlinear, making it adapt through rank-based correlation and clustering.	It is less scalable since the Pearson correlation method assumes a linear relationship between the tested features, making it less effective in handling dynamic traffic patterns.
Correlation threshold Strategy	Dynamic: 90^{th} percentile of correlation distribution.	Fixed thresholds.

As Table 6 shows, related works such as Li et al. [10] and Pires and Mascarenhas [11] leveraged Pearson and Spearman correlations; however, their approaches fail to catch non-linear relationships or outliers, which the proposed approach addresses by incorporating Kendall correlation. Furthermore, the dynamic threshold that determined as the 90^{th} percentile of the correlation values for both Spearman and Kendall methods, such that this mechanism ensures that only the highly correlated features specific to the dataset distribution are selected. Relying on a predefined fixed threshold (e.g., 0.7 as suggested in [10]) may not be a robust methodology to obtain the most correlated features and can be less effective across diverse datasets when correlation strength varies significantly. Therefore, the proposed approach offers a robust correlation selection approach that suits all dataset types and can choose the most relevant features and hence reducing the risks of including or omitting other relevant ones. In conclusion, the proposed approach is better for

analyzing the features of evolving network architectures because of its adaptive nonlinear correlation and adaptive clustering. Moreover, it provides clear feature visualizations that can help network experts validate feature relationships, hence improving real-time decision-making in various network-related areas.

4.5. Limitations and Potential Future Works

The proposed centric feature selection and reduction may result in information loss if too many features are removed, especially in security performance which might increase the false positive/false negative rates. Therefore, it is suggested to evaluate the proposed farmwork effectiveness using machine learning-based classification models such as Random Forest, Deep Learning, and XGBoost on intrusion detection and traffic classification. Furthermore, deploying the proposed centric feature solution in a live network environment such as software-defined networking (SDN) systems, cloud or IoT networks, to reveal the framework's practical efficiency. The performance of the proposed framework needs to be compared against deep learning-based methods such as reinforcement learning and feature attention mechanisms.

5. CONCLUSIONS

This paper proposes a centralized feature framework for optimizing network traffic management, network performance monitoring, resource utilization, future-proofing the network, and security enhancement through advanced pairwise cross-correlation based on Spearman and Kendall, with the help of dendrogram-based clustering. The framework automatically analyzes the analysis of features and eliminates the redundant ones, thus keeping only the informative ones and improving the interpretability, allowing for a more structured and efficient analysis of network behavior. Comparing it with existing solutions, the proposed framework provides a more robust nonlinear approach for feature selection, making it adaptable through rank-based correlation and clustering and well-suited for modern, dynamic network environments. However, this work still needs further assessment regarding live network management and security systems. This framework provides a viable foundation for feature optimization and reduction in network analysis with potential applications in understanding network behaviors, optimizing intrusion detection, network performance monitoring, and future-proofing the network.

REFERENCES

[1] A. Azab, M. Khasawneh, S. Alrabaee, K.-K. R. Choo, and M. Sarsour, “Network traffic classification: Techniques, datasets, and challenges,” *Digital Communications and Networks*, vol. 10, no. 3, pp. 676–692, Jun. 2024, doi: 10.1016/j.dcan.2022.09.009.

[2] F. Iglesias and T. Zseby, “Analysis of network traffic features for anomaly detection,” *Mach Learn*, vol. 101, no. 1–3, pp. 59–84, Oct. 2015, doi: 10.1007/s10994-014-5473-9.

[3] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization:,” in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018, pp. 108–116. doi: 10.5220/0006639801080116.

[4] A. Nazir and R. A. Khan, “A novel combinatorial optimization based feature selection method for network intrusion detection,” *Computers & Security*, vol. 102, p. 102164, Mar. 2021, doi: 10.1016/j.cose.2020.102164.

[5] H. Nguyen, K. Franke, and S. Petrovic, “Improving Effectiveness of Intrusion Detection by Correlation Feature Selection,” in *2010 International Conference on Availability, Reliability and Security*, Krakow, Poland: IEEE, Feb. 2010, pp. 17–24. doi: 10.1109/ARES.2010.70.

[6] M. A. Bouke, A. Abdullah, J. Frnda, K. Cengiz, and B. Salah, "BukaGini: A Stability-Aware Gini Index Feature Selection Algorithm for Robust Model Performance," *IEEE Access*, vol. 11, pp. 59386–59396, 2023, doi: 10.1109/ACCESS.2023.3284975.
[7] Y. Zhang, H. Zhang, and B. Zhang, "An Effective Ensemble Automatic Feature Selection Method for Network Intrusion Detection," *Information*, vol. 13, no. 7, p. 314, Jun. 2022, doi: 10.3390/info13070314.
[8] E. Jaw and X. Wang, "Feature Selection and Ensemble-Based Intrusion Detection System: An Efficient and Comprehensive Approach," *Symmetry*, vol. 13, no. 10, p. 1764, Sep. 2021, doi: 10.3390/sym13101764.
[9] P. Chen, F. Li, and C. Wu, "Research on Intrusion Detection Method Based on Pearson Correlation Coefficient Feature Selection Algorithm," *J. Phys.: Conf. Ser.*, vol. 1757, no. 1, p. 012054, Jan. 2021, doi: 10.1088/1742-6596/1757/1/012054.
[10] J. Li, M. S. Othman, H. Chen, and L. M. Yusuf, "Cybersecurity Insights: Analyzing IoT Data Through Statistical and Visualization Techniques," in *2024 International Symposium on Parallel Computing and Distributed Systems (PCDS)*, Singapore, Singapore: IEEE, Sep. 2024, pp. 1–10. doi: 10.1109/PCDS61776.2024.10743769.
[11] S. Pires and C. Mascarenhas, "Cyber Threat Analysis Using Pearson and Spearman Correlation Via Exploratory Data Analysis," in *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India: IEEE, May 2023, pp. 257–262. doi: 10.1109/ICSCCC58608.2023.10176973.
[12] C. Xiao, J. Ye, R. M. Esteves, and C. Rong, "Using Spearman's correlation coefficients for exploratory data analysis on big dataset," *Concurrency and Computation*, vol. 28, no. 14, pp. 3866–3878, Sep. 2016, doi: 10.1002/cpe.3745.
[13] J. Hauke and T. Kossowski, "Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data," *Quaestiones Geographicae*, vol. 30, no. 2, pp. 87–93, Jun. 2011, doi: 10.2478/v10117-011-0021-1.
[14] J. Tekli, "An overview of cluster-based image search result organization: background, techniques, and ongoing challenges," *Knowl Inf Syst*, vol. 64, no. 3, pp. 589–642, Mar. 2022, doi: 10.1007/s10115-021-01650-9.