Journal of *Cybersecurity and Privacy*

MDPI

Article

Statistical and Multivariate Analysis of the IoT-23 Dataset: A Comprehensive Approach to Network Traffic Pattern Discovery

Humera Ghani *, Shahram Salekzamankhani and Bal Virdee

School of Computing and Digital Media, London Metropolitan University, London N7 8DB, UK
* Correspondence: hug0051@my.londonmet.ac.uk

Abstract

The rapid expansion of Internet of Things (IoT) technologies has introduced significant challenges in understanding the complexity and structure of network traffic data, which is essential for developing effective cybersecurity solutions. This research presents a comprehensive statistical and multivariate analysis of the IoT-23 dataset to identify meaningful network traffic patterns and assess the effectiveness of various analytical methods for IoT security research. The study applies descriptive statistics, inferential analysis, and multivariate techniques, including Principal Component Analysis (PCA), DBSCAN clustering, and factor analysis (FA), to the publicly available IoT-23 dataset. Descriptive analysis reveals clear evidence of non-normal distributions: for example, the features src_bytes, dst_bytes, and src_pkts have skewness values of −4.21, −3.87, and −2.98, and kurtosis values of 38.45, 29.67, and 18.23, respectively. These values indicate highly skewed, heavy-tailed distributions with frequent outliers. Correlation analysis revealed a strong positive correlation (0.97) between orig_bytes and resp_bytes, and a strong negative correlation (−0.76) between duration and resp_bytes, while inferential statistics indicate that linear regression provides optimal modeling of data relationships. Key findings show that PCA is highly effective, capturing 99% of the dataset's variance and enabling significant dimensionality reduction. DBSCAN clustering identifies six distinct clusters, highlighting diverse network traffic behaviors within IoT environments. In contrast, FA explains only 11.63% of the variance, indicating limited suitability for this dataset. These results establish important benchmarks for future IoT cybersecurity research and demonstrate the superior effectiveness of PCA and DBSCAN for analyzing complex IoT network traffic data. The findings offer practical guidance for researchers in selecting appropriate statistical methods for IoT dataset analysis, ultimately supporting the development of more robust cybersecurity solutions.

Keywords: IoT security; network traffic; statistical analysis; traffic pattern; DBSCAN clustering; factor analysis; dataset complexity

Academic Editors: Feng Wang and Yongning Tang

Received: 25 June 2025
Revised: 8 November 2025
Accepted: 11 December 2025
Published: 16 December 2025

Citation: Ghani, H.; Salekzamankhani, S.; Virdee, B. Statistical and Multivariate Analysis of the IoT-23 Dataset: A Comprehensive Approach to Network Traffic Pattern Discovery. *J. Cybersecur. Priv.* **2025**, *5*, 112. https://doi.org/10.3390/jcp5040112

1. Introduction

The Internet of Things is one of the top emerging trends in technology with rapid growth. It is forecasted that it will grow over 16% annually until 2028 (Statista, 2024) [1]. This proliferation generates vast volumes of heterogeneous network traffic data characterized by complex communication patterns and diverse behavioral characteristics. As IoT deployments expand into critical infrastructure, understanding network traffic

patterns become essential for developing effective cybersecurity solutions and ensuring system reliability.

Despite IoT's growing importance, analyzing IoT network traffic presents significant challenges due to inherent data complexity. IoT datasets frequently exhibit non-normal distributions, high dimensionality, and class imbalance, creating substantial barriers for traditional analytical approaches. The data often contains extreme skewness and kurtosis values, indicating heavy-tailed distributions with frequent outliers that complicate pattern recognition and security threat detection.

While several IoT datasets have been released for cybersecurity research, there remains a lack of comprehensive statistical and multivariate analyses that systematically evaluate multiple analytical techniques on the same dataset. This study addresses this gap by conducting a thorough statistical and multivariate analysis of the IoT-23 dataset to: (1) characterize the statistical properties and complexity of IoT network traffic data, (2) evaluate the effectiveness of various analytical techniques for pattern discovery, and (3) establish practical benchmarks for future IoT cybersecurity research.

This research employs a comprehensive analytical framework combining descriptive statistics (mean, variance, skewness, kurtosis), inferential analysis (correlation analysis, polynomial regression), and multivariate techniques (Principal Component Analysis, DBSCAN clustering, factor analysis). The methodology enables systematic comparison of different approaches and provides insights into their relative effectiveness for IoT data analysis.

This study makes important contributions by providing detailed statistical characterization of the IoT-23 dataset, demonstrating the comparative effectiveness of different analytical methods, and establishing practical guidance for researchers. The findings reveal that PCA captures 99% of dataset variance while DBSCAN identifies six distinct traffic clusters, whereas factor analysis explains only 11.63% of variance, indicating limited applicability. These results support the development of more robust cybersecurity solutions and inform algorithm selection for IoT data analysis. The main contributions of this research are:

- Provides the exhaustive statistical analysis of the IoT-23 dataset specifically focused on IoT security applications, revealing critical insights into data complexity and distribution properties that establish foundational understanding for developing robust IoT security algorithms and intrusion detection systems.
- Conducts the comprehensive comparison of three distinct multivariate techniques (PCA, DBSCAN clustering, and factor analysis) applied to the same IoT-23 dataset, enabling informed methodological decisions for IoT security researchers in selecting optimal analytical approaches for threat detection and anomaly identification.
- Establishes performance benchmarks and methodological recommendations for IoT security research, serving as reference points for future studies in IoT cybersecurity, intrusion detection system development, and network anomaly detection in IoT environments.
- Introduces a comprehensive framework that integrates descriptive statistics, inferential measures, and multivariate techniques within a single cohesive approach, providing IoT security researchers with a complete methodological method for analyzing network traffic patterns and identifying security threats in IoT deployments.
- Provides concrete evidence and guidelines specifically for algorithm selection in IoT security and cybersecurity applications, demonstrating the relative effectiveness of different analytical approaches for IoT network traffic analysis, threat detection, and security pattern recognition in IoT ecosystems.

2. Related Works

This section reviews recent studies that have improved our understanding of IoT datasets and investigates how these studies have employed statistical methods to extract meaningful information from the data.

Wang & Hsu (2025) [2] employ descriptive statistics methods for analyzing IoT-based maritime logistics data. Metrics such as mean, median, mode, standard deviation, and variance are calculated to assess the overall performance and stability of logistics operations. Furthermore, visual representation tools like histograms and line charts were employed to show clear insights into data distribution. However, their approach is limited to basic descriptive statistics and lacks advanced multivariate analysis techniques that could reveal deeper patterns and relationships within IoT data.

Ebrahem et al. (2025) [3] presents a new feature reduction approach targeting for IoT-based cybersecurity systems. Their methodology incorporates various feature selection methods including ANOVA, Variance Threshold, Information Gain (IG), and Chi Square, which were applied to the IoT-23 dataset. The approach involves dividing universal features into several subgroups and evaluating the performance of multiple machine learning algorithms with extensive validation experiments conducted on the CICIDS2017 dataset. These researchers employed the Pearson Correlation Coefficient (PCC) technique to analyze relationships between six selected features within the CICIDS2017 dataset. The research focus primarily on feature reduction without providing comprehensive statistical characterization of the dataset's underlying distribution properties and complexity.

Elkhadir & Begdouri (2025) [4] focus on investigating the effectiveness of PCA and Kernel PCA feature extraction methods when combined with various machine learning classifiers for detecting and classifying cyberattacks in IoT systems. Their study is limited to examining only PCA-based dimensionality reduction techniques and does not provide a comprehensive statistical analysis framework that includes descriptive measures, inferential statistics, and comparative evaluation of multiple multivariate methods.

Al-Zewairi et al. (2025) [5] works on identifying unknown attacks in IoT and traditional networks. Their approach is to enhance cluster analysis by integrating the well-known DBSCAN algorithm with recent advancements in manifold learning. The study leverages DBSCAN's ability to handle complex cluster shapes and combines it with topological manifold learning techniques to improve the accuracy and robustness of cluster detection. However, their work lacks a foundational statistical characterization of the dataset and does not provide comparative analysis with other multivariate techniques to establish the relative effectiveness of their approach.

Aqil et al. (2024) [6] propose a robust and lightweight feature set, for IoT traffic, based on statistical values of payload lengths, designed to maintain stability over time and reduce the need for frequent model retraining. These researchers employed the Kolmogorov-Smirnov (K-S) test as a non-parametric statistical test to evaluate changes in feature distributions over time. The limitation of their approach is the focus on specific statistical features without conducting a comprehensive analysis of the overall dataset complexity and distribution characteristics.

Sharmila et al. (2024) [7] performed the experiments on the RT-IoT2022 dataset. The research employs statistical tests like the K-S test, skewness, kurtosis, PCC, and IG Ratio to analyze the dataset's characteristics and identify optimal features for machine learning (ML) models. While their statistical analysis is comprehensive, they focus on a different dataset (RT-IoT2022) rather than the widely used IoT-23 dataset, limiting the comparative value for IoT-23 based research.

Li et al. (2024) [8] highlight the importance of data analysis and visualization in uncovering patterns that can guide effective intrusion detection strategies for IoT traffic.

Their study investigates the 5-tuple network properties and employs statistical methods, including Pearson and Spearman correlation coefficients, to validate these properties. Furthermore, the research utilizes unsupervised learning techniques such as PCA and t-SNE to identify class imbalances, overlaps, and patterns within IoT data. The research focuses on visualization and specific network properties without providing systematic comparative evaluation of multiple analytical techniques or comprehensive statistical characterization.

Smiesko et al. (2023) [9] utilize a range of statistical techniques, including the coefficient of variation, kurtosis, skewness, autoregression, correlation, hurst exponent, and Kullback–Leibler divergence to detect DDoS attacks. The study presents detection functions that use the result from these statistical techniques during the early stages of an attack which enables quick identification. Their approach is limited in scope, focusing specifically on DDoS detection rather than providing a comprehensive analytical framework applicable to broader IoT security research.

Kim et al. (2022) [10] perform an in-depth analysis of the IoT-23 dataset, demonstrating that a high detection rate can be achieved with just 2 or 3 features, which is advantageous for resource-limited IoT processors. The study employs IG as a method for feature selection, relying on entropy measurements to evaluate the importance of different features. The authors stress the importance of aligning feature selection methods with detection algorithms, specifically pairing IG with Tree detection algorithms, as both utilize entropy-based calculations. This research is focused on feature selection and lacks comprehensive statistical characterization of the dataset's distribution properties, complexity assessment, and comparative evaluation of multiple analytical approaches.

Chakraborty et al. (2020) [11] addresses the challenges hindering IoT adoption in developing countries like Bangladesh. In their study they conducted a survey to gather data on factors influencing IoT adoption, then utilized Exploratory Factor Analysis to group the survey responses into five key factors. Next, they built a measurement model and applied Structural Equation Modeling to analyze the relationships between these factors. This research focuses on IoT adoption factors rather than network traffic analysis, making it less relevant for cybersecurity-focused IoT dataset analysis.

The comprehensive review of existing literature reveals several critical research gaps in IoT dataset analysis, each of which this study systematically addresses through a comprehensive analytical framework applied to the IoT-23 dataset. First, there is a lack of systematic comprehensive statistical characterization of IoT datasets, with most studies focusing on specific aspects rather than providing holistic understanding. This study bridges this gap by providing exhaustive descriptive analysis including mean, variance, skewness, and kurtosis measurements across all dataset records, revealing that key features exhibit extreme non-normal distributions and establishing comprehensive statistical profiles. Second, the absence of comparative evaluation of multiple analytical techniques on the same dataset limits researchers' ability to make informed methodological decisions. This research addresses this limitation by systematically comparing three distinct multivariate techniques, PCA, DBSCAN clustering, and factor analysis, on the same IoT-23 dataset, demonstrating their relative effectiveness and establishing clear performance hierarchies. Third, most studies lack detailed analysis of data distribution properties and complexity assessment, which are essential for appropriate algorithm selection. This study fills this gap by employing K-S tests confirming dataset consistency, comprehensive correlation analysis revealing important feature relationships, and polynomial regression analysis demonstrating optimal model performance characteristics. Fourth, there is insufficient provision of practical benchmarks and guidance for future IoT cybersecurity research. This research addresses this need by establishing clear performance metrics and methodologi-

cal recommendations, demonstrating effective dimensionality reduction capabilities and clustering effectiveness while providing concrete guidance for algorithm selection. Finally, limited integration of descriptive, inferential, and multivariate analysis techniques in a single comprehensive framework restricts the depth of understanding achievable from IoT datasets. This study uniquely combines descriptive statistics, inferential measures, and multivariate techniques within a unified analytical framework, providing researchers with comprehensive insights for informed algorithm selection and methodology choice in IoT cybersecurity applications.

3. Dataset

(Garcia et al., 2020) [12] The dataset employed in this study consists of 23 distinct captures of IoT traffic. Among these, 20 scenarios were derived from infected IoT devices, while the remaining 3 originated from uninfected devices. Each malicious capture was created by executing a specific malware on a Raspberry Pi. These malicious records were categorized under 12 different labels (Table 1). On the other hand, the benign captures, which represented 3 different scenarios from IoT devices, were all assigned a single label.

Table 1. Label distribution of the IoT-23 dataset.

Record Labels	Number of Records
Benign	688,812
PartOfAHorizontalPortScan	3,389,036
Okiru	1,313,012
DDoS	638,506
C&C	15,286
C&C-HeartBeat	1332
Attack	538
C&C-FileDownload	46
C&C-Torii	30
FileDownload	13
C&C-HeartBeat-FileDownload	8
Okiru-Attack	3
C&C-Mirai	1
Total	6,046,623

The IoT-23 dataset consists of 26 features and 6,046,623 records, of which 688,812 (11.39%) are benign, and 5,357,811 (88.60%) are malicious. Malicious records are categorized into 12 distinct types, with 8 of these types being rare in the dataset. The complete distribution of labels is presented in (Table 1). A comparison of this dataset with the existing datasets is shown in (Table 2).

Table 2. Comparisons of datasets.

Dataset and Year	Collection Duration	Type of Traffic	Size	No. of Records (Millions)	Labels
RT-IoT2022 (2024)	n/a	real	52.2 MB	0.12 M	13
CICIoT2023 (2023)	5 days	hybrid of real and simulated	16 GB	46.68 M	8
LITNET-2020 (2020)	10 months	real	n/a	45.49 M	13
IoT-23 (2020)	2018–2019	real	8.8 GB (small) 20 GB (full)	6.05 M (small)	10
ToN-IoT (2019)	2 months	simulated	n/a	22.7 M	10
BoT-IoT (2018)	5 days	simulated	16.7 GB	72 M	7
CICIDS2017 (2017)	5 days	simulated	51.1 GB	2.83 M	15
UNSW-NB15 (2015)	16 days	hybrid of real and simulated	100 GB	2.54 M	10

This dataset was published in January 2020 and contains labeled records of both benign and malicious IoT traffic, captured from real IoT devices. The data collection took place between 2018 and 2019 at the Stratosphere Laboratory, which is part of the Center for Artificial Intelligence at the Faculty of Electrical Engineering, Czech Technical University, Czech Republic. To facilitate this, a controlled network environment was established with unrestricted internet access. For this study, we selected the IoT-23 dataset primarily due to following reasons:

- The dataset is widely utilized within the research community; however, its statistical characteristics have yet to be studied. This presents a valuable opportunity to conduct a detailed statistical analysis, which can enhance our understanding of the dataset's distribution, variability, and underlying patterns. By sharing these insights, we aim to support research community in making more informed decisions when applying the dataset to their studies, ultimately advancing the field of IoT security
- The dataset provides extensive coverage of real IoT scenarios, capturing 20 scenarios from infected devices and 3 from uninfected devices, which allows for a thorough analysis of both normal and malicious behaviors.
- The dataset offers both small and full versions, ensuring accessibility for various research needs and facilitating comprehensive experimentation.

4. Data Preprocessing

Data preprocessing is a fundamental step in data analysis that includes data cleaning, data transformation, data reduction, and data merging. These processes collectively aim to prepare raw data for further analysis, ensuring it is in a suitable format for subsequent data processing methods. Effective data preprocessing can greatly enhance the efficiency and accuracy of subsequent processes.

The dataset contains several rare labels, including C&C-HeartBeat (1332 instances), Attack (538 instances), C&C-FileDownload (46 instances), C&C-Torii (30 instances), File-Download (13 instances), C&C-HeartBeat-FileDownload (8 instances), Okiru-Attack (3 instances), and C&C-Mirai (1 instance). These labels have been merged with other labels for a more streamlined analysis. Table 3 provides a detailed overview of which records have been merged and under which labels they now fall. This approach ensures a more efficient and effective analysis of the data, particularly when dealing with rare labels that may otherwise skew the results.

Table 3. Merged Labels.

New Label	Merged Labels	Number of Records
0	Benign	688,812 (11.39%)
1	PartOfAHorizontalPortScan Attack	3,389,574 (56.05%)
2	Okiru Okiru-Attack	1,313,015 (21.71%)
3	DDoS	638,506 (10.55%)
4	C&C C&C-HeartBeat C&C-FileDownload C&C-Torii FileDownload C&C-HeartBeat-FileDownload C&C-Mirai	16,716 (0.27%)
	Total	6,046,623 (100%)

Following features were dropped: Unnamed: 0, uid, local_orig, local_resp, missed_bytes, id.orig_p and id.resp_p as these features were not relevant to the problem at hand. Categorical feature id.resp_h consisted of IP addresses. Initially, these addresses were cleaned. Next, they were replaced with integers indicating whether they were public or private IP addresses. Other categorical features were encoded. Finally, all the features were log-transformed and scaled using the MinMaxScaler method.

5. Descriptive Measures

Wang & Hsu (2025) [2] pointed that the application of descriptive statistical analysis methods has emerged as an essential tool to enhance the comprehension and effective utilization of data. In this section, we performed the descriptive statistical analysis techniques, including mean, variance, standard deviation, skewness, kurtosis, and the Kolmogorov-Smirnov test, as employed by Moustafa & Slay (2016) [13] and Sharmila et al. (2024) [7]. These techniques are applied to two distinct datasets: the training set and the testing set. These techniques are applied to organize, summarize, and visualize the data in a clear way that enhances interpretability and helps to explore and evaluate the relationship between observations. Table 4 shows the notation, and their definitions used in this sub-section.

Table 4. Notations and their definitions.

Notation	Definition
i	It represents a number
n	It is the total number of observations
x	It represents a variable/predictor/feature
x_i	It represents each individual value of the variable x
μ	It is the mean of values
σ	It represents standard deviation
σ_x	It represents the standard deviation of variable x

Variables in a dataset often become incomparable due to differences in their units. To address this, we apply a minmax transformation, which scales features between 0 and 1. This transformation scales and translates the data without altering its distribution. The transformation is calculated by subtracting the minimum from each individual value and then dividing the result by the difference in maximum and minimum values (i.e., variable range). This process standardizes the data, expressing variable into a fixed range. This transform is defined as:

$$x_{transf} = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

5.1. Mean

Mean is the sum of all data points divided by the total number of points. It represents the average of all the values in the distribution.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

To measure the average value of the variables, the mean is calculated for each variable of both the training and test sets. Figure 1a offers a visual comparison of the mean values across all variables, enabling the identification of any substantial discrepancies between the two subsets. The close alignment of mean values for most variables in both the training and test sets suggests that the datasets are drawn from similar distributions and are thus representative of the same underlying population. This consistency is vital for the development of robust machine learning models, as it minimizes the risk of bias or data drift that could otherwise compromise model generalizability.

5.2. Variance

Variance is the average of the squared differences in each data point from the mean. It measures the spread of the distribution around the mean, indicating how far each data point deviates from it.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} \tag{3}$$

To evaluate the variability in the data, the variance is calculated. When the variance is close to the mean, it indicates that the data has less variability. However, when it deviates significantly from the mean, it suggests high variability. Figure 1b graphically represents the variance for each variable in both the training and test sets, facilitating a comparative assessment of data variability across the two subsets. The figure reveals that while some features exhibit low variance, indicating that their values are tightly clustered around the mean, others display significantly higher variance, suggesting a broader spread of values and the potential presence of outliers. The similarity in variance patterns between the training and test sets, as observed in the figure, supports the reliability of subsequent modeling efforts by confirming that both subsets capture comparable levels of data variability.

Figure 1. (**a**) Mean. (**b**) Variance. (**c**) Standard Deviation. (**d**) Skewness. (**e**) Kurtosis.

5.3. Standard Deviation

Standard deviation is the square root of the variance. It quantifies the average deviation of each data point from the mean.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \tag{4}$$

To assess the dispersion of data points from the mean and from each other, the standard deviation is computed for each variable in the training and test sets. Figure 1c provides a visual summary of the standard deviation values for all variables in both the training and test datasets, highlighting the degree to which data points typically deviate from the mean. The figure demonstrates that certain features possess notably higher standard deviations, indicating substantial variability and a greater likelihood of encountering outliers or anomalous observations. In contrast, features with low standard deviation are more stable and less likely to contribute to anomaly detection unless significant deviations occur. The overall consistency in standard deviation values between the training and test sets, as shown in the figure, further reinforces the representativeness of the data and the appropriateness of the dataset for robust model development.

5.4. Skewness

Skewness measures the asymmetry of a distribution. Positive skewness indicates a long tail on the right, suggesting more positive values, while negative skewness signifies a long tail on the left, indicating more negative values. Skewness is important as it highlights the presence of extreme values, revealing potential outliers in the data.

$$Skewness = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3} \tag{5}$$

To examine the asymmetry of the features in the training and test sets, skewness is calculated. Figure 1d reveals that most of the features exhibit extreme skewness values, indicating that they are highly skewed and not normally distributed. Except for ts, id.resp_h, duration, resp_pkts, and resp_ip_bytes, most features show negative skewness. This means the left tail of the distribution is longer or fatter, and more data points are shifted to the right side. Such skewed distributions are characteristic of datasets where outliers or anomalies are present, which is particularly relevant in the context of network anomaly detection.

5.5. Kurtosis

Kurtosis measures the thickness of a distribution's tails compared to a normal distribution. Heavy tails indicate a greater number of outliers, while light tails suggest fewer or no outliers.

$$Kurtosis = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{\sigma^4} \tag{6}$$

To measure the height and shape of the distribution, the kurtosis is calculated for each variable in the training and test sets. A positive kurtosis value indicates that the data has heavy tails and a sharp peak, suggesting more frequent extreme values than a normal distribution. Conversely, a negative kurtosis value signifies that the data has light tails and a flat peak, implying fewer extreme values than a normal distribution. Figure 1e shows that most of the variables exhibit high kurtosis values, further indicating that they are not normally distributed. This finding is important in the context of anomaly detection, as it suggests that the dataset contains frequent outliers. The similarity in kurtosis patterns between the training and test sets further validates the representativeness of the data and supports the reliability of subsequent analyses.

5.6. Kolmogorov-Smirnov Tests

The K-S test is a non-parametric statistical method used to compare probability distributions. It is employed to determine whether a sample comes from a specified theoretical distribution or to compare two samples to determine if they are drawn from the same underlying distribution. The one-sample K-S test determines whether the samples are

drawn from a specified theoretical distribution by comparing the empirical cumulative distribution function, of the sample data, with the cumulative distribution function of the reference distribution. The two-sample K-S test assesses whether two distributions are drawn from the same underlying distribution by comparing their empirical cumulative distribution functions. It calculates the maximum difference between the two functions to determine the equality of the distributions.

$$D = \sup_x |F_0(x) - F_1(x)| \tag{7}$$

where $F_0(x)$ and $F_1(x)$ are the cumulative distribution functions of the two observed distributions and $\sup_x$ is the largest absolute difference between them.

We utilized the K-S test to ascertain whether the training and test sets follow the same distribution or diverge from each other. The results of the test were determined using p-values. A p-value represents the probability of obtaining a result that is either identical to or more extreme than the observed data. The p-values we obtained are close to 1, indicating a high probability that the observed differences between the two distributions are due to chance alone. The results demonstrated remarkable consistency across all network features indicating that the distributions in both the train and test datasets are not significantly different Figure 2.

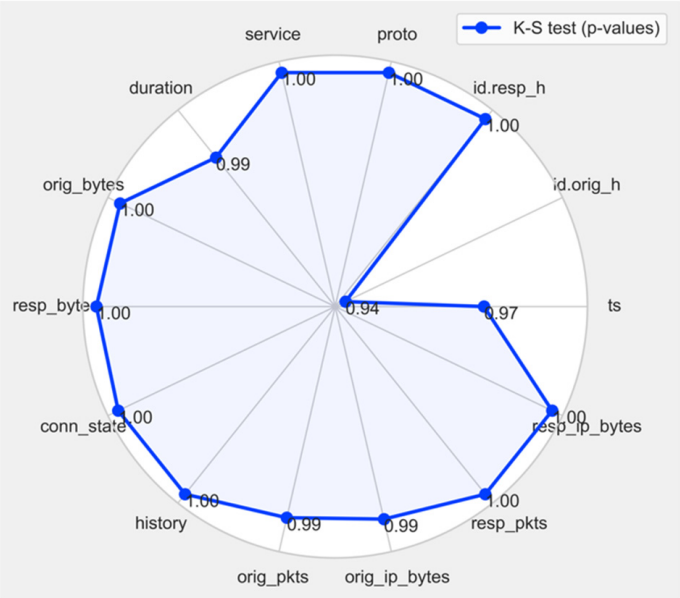

Figure 2. Kolmogorov-Smirnov test (p-values).

6. Inferential Measures

Inferential statistical measures are employed to draw conclusions from the data which helps researchers to formulate evidence-based generalization. This paper used PCC, gain ratio (GR), and polynomial regression methods for inference about the dataset.

PCC measures the linear relationship between variables, GR reduces the bias in predictors with numerous outcomes and polynomial regression models non-linearity of data using polynomial equation of varying degree.

6.1. Pearson Correlation Coefficient

Covariance measures the linear relationship between two variables. Correlation standardizes covariance by dividing it by the product of the standard deviations of the two variables. The correlation coefficient ranges between -1 and $+1$. The correlation coefficient is positive

if high values of one variable correspond with high values of the other variable. It is negative if high values of one variable correspond with low values of the other variable. If the correlation coefficient is zero, it means there is no correlation between variables. If correlation of two variables i and j is represented as σ_{ij} and their standard deviations are represented as σ_i and σ_j then PCC is defined as:

$$PCC = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{8}$$

We utilized the Pearson correlation coefficient to measure the strength and direction of the linear relationship between two variables. In Figure 3, the correlation values between any two variables in the dataset are presented. It is evident that the variables 'orig_bytes' and 'resp_bytes' have a strong positive correlation of 0.97. In contrast, the variables 'duration' and 'resp_bytes' exhibit a strong negative correlation of -0.76.

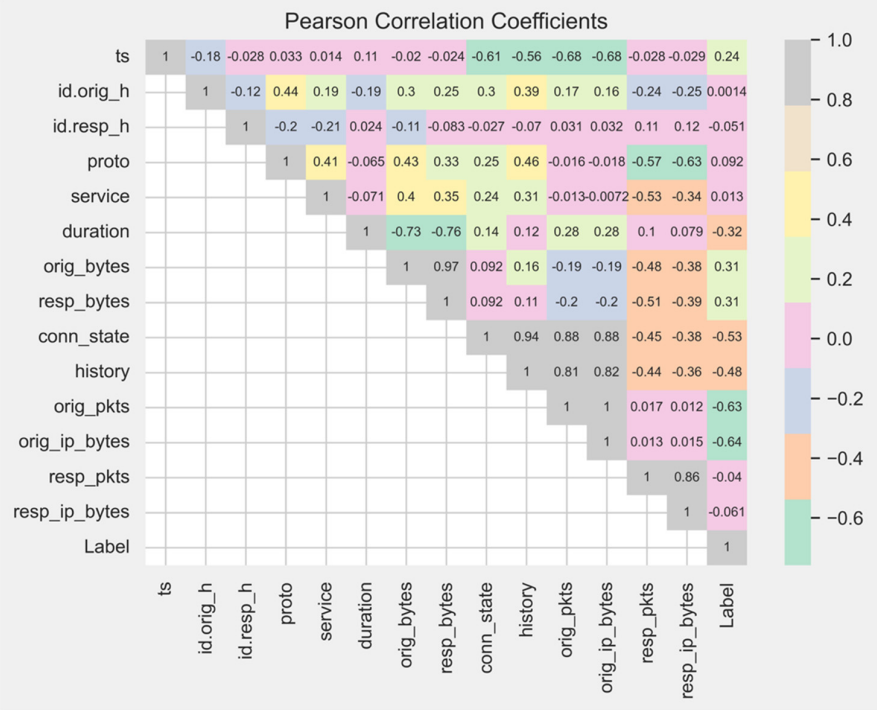

Figure 3. Pearson Correlation Cofficient.

6.2. Gain Ratio

The gain ratio is a measure used to reduce the bias towards the selection of multi-valued variable. It considers both the number and size of the partitions when choosing a variable. To define the gain ratio, first we will discuss entropy, information gain (IG), and intrinsic value (IV).

The entropy is the measure of impurity or non-homogeneity in our data which is computed as:

$$Entropy(S) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{9}$$

where p_i is the probability of a class i and n is the number of possible classes in the dataset S. Assuming we have two classes, with respect to target classes if our data is completely homogeneous (no impurity) then entropy will be 0, if it can be equally divided into two classes then entropy will be 1.

The information gain is the measure of effectiveness of a variable in classifying the data points. It is defined as the expected reduction in entropy by partitioning a dataset based on a specific attribute of a variable (or feature).

$$IG(S,\ F) = Entropy\ (S) - \sum_{v\ \in\ Values(F)} \frac{|S_v|}{|S|}\ Entropy(S_v) \tag{10}$$

where S represents the dataset, F is the feature on which the dataset is being split, $Values(F)$ is the set of values of feature F, S_v is the subset of the dataset S where feature F takes the value v, $|S_v|$ is the number of records in subset S_v, and $|S|$ is the total number of records in the dataset S, $Entropy\ (S)$ is the entropy of dataset and $Entropy\ (S_v)$ is the entropy of subset S_v.

The information gain metric favors the feature having large number of distinct values which could cause overfitting. The gain ratio metric is the refinement of information gain, it evaluates the quality of a split. It attempts to reduce the bias of information by introducing a normalizing term called the intrinsic information. The intrinsic value is the entropy of sub-dataset proportions.

$$IV\ (F) = -\sum_{v\ \in\ Values\ (A)} \frac{|S_v|}{S}\ log_2\left(\frac{|S_v|}{|S|}\right) \tag{11}$$

$$GR\ (F) = \frac{IG\ (S,\ F)}{IV\ (F)} \tag{12}$$

where $IG(S,\ F)$ is the information gain of feature F and $IV(F)$ is the intrinsic value of feature F. The gain ratio has values between 0 and 1. Higher value indicates a better feature for split while value close to 0 indicates non-informative feature.

We used the gain ratio measure to evaluate the effectiveness of a variable in splitting the data. A higher gain ratio indicates that the variable is more effective at dividing the data into distinct classes. Figure 4 shows that, on a scale of 0 to 1, the 'conn_state' variable has the highest gain ratio (0.78), while the 'id.resp_h' variable has no gain ratio (0.00). This is because the 'id.resp_h' variable contains only two distinct values, with their distribution being 98.88% and 1.11%, respectively.

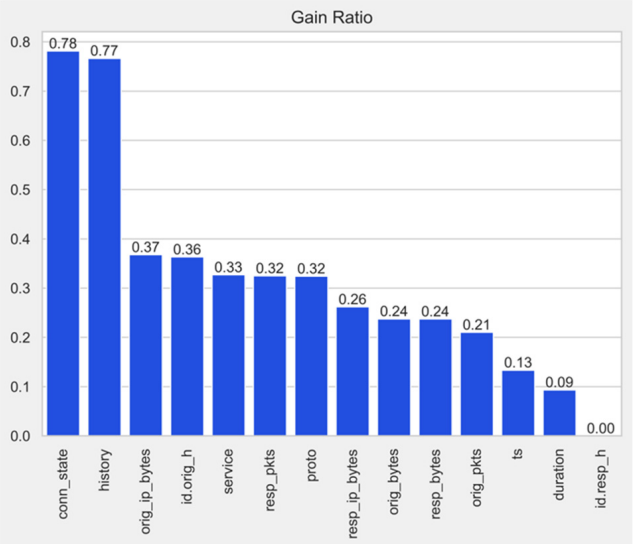

Figure 4. Result of Gain Ratio method.

6.3. Polynomial Regression

Polynomial Regression is a statistical method used to analyze and model the non-linear relationship between predictor and response variables. It is represented by the equation:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n \quad (13)$$

where β terms represent coefficients, n represent degree of the polynomial, x and y are predictor and response variables, respectively.

As we increase the degree of the polynomial, the model fits the non-linear relationship better. However, the bias-variance tradeoff must be considered. Increasing the degree of the polynomial can lead to high variance, where the model becomes overly complex and captures noise in the data, resulting in overfitting. On the other hand, reducing the degree may lead to high bias, where the model is too simplistic and fails to capture the underlying complexity of the data, resulting in underfitting. To evaluate the polynomial regression model, we used mean square error (MSE) and R^2 metrics as shown in Table 5. MSE evaluates how well a polynomial regression model fits the data. R^2 explains how well the proportion of variance in the dependent variable explained by the independent variables. We performed the test on three different degrees of freedom, i.e., 1, 2, and 3. Ridge regularization value was set to 10.0.

Table 5. Polynomial Regression Metrics.

	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Degree of Polynomial	1	1	2	2	3	3
MSE	0.28937	0.29156	0.17491	0.28612	0.15001	16.66542
R^2	0.57077	0.56760	0.74054	0.57566	0.77748	−23.71539

The linear model (degree of freedom is 1) demonstrates stable and consistent performance across both training and test datasets. With a training dataset MSE of 0.28937 and a test dataset MSE of 0.29156, the model shows minimal deviation between training and testing performance, indicating robust generalization capabilities. The R^2 values further support this stability, with the model explaining approximately 57% of the variance in both training (R^2 = 0.57077) and test (R^2 = 0.56760) sets. This consistency between training and test metrics suggests that the linear model has captured genuine patterns in the data without overfitting, making it a reliable choice for prediction purposes. The relatively small difference between training and test metrics (difference of only 0.00219 in MSE) indicates that the model maintains its predictive power well when faced with new unseen data.

The quadratic model (degree of freedom is 2) shows improved training performance but exhibits early signs of overfitting. The training MSE of 0.17491 represents a substantial improvement over the linear model's training performance, and the training R^2 of 0.74054 suggests that the model explains approximately 74% of the variance in the training data. However, when examining the test metrics, we observe some deterioration in performance. The test MSE increases to 0.28612, and the test R^2 drops to 0.57566, which is only slightly better than the linear model's test performance. The growing gap between training and test metrics (difference of 0.11121 in MSE) indicates that the additional complexity of the quadratic term might be fitting noise in the training data rather than capturing meaningful patterns that generalize well to new observations.

The cubic model (degree of freedom is 3) demonstrates severe overfitting and unstable performance, making it the least suitable choice for this dataset. While it achieves the best training metrics with an MSE of 0.15001 and an R^2 of 0.77748, its performance on the test set

is poor. The test MSE is 16.66542, representing a massive increase in prediction error, while the test R^2 is −23.71539, indicating that the model performs poorly. This deterioration in test performance (difference of 16.51541 in MSE between training and test) clearly shows that the cubic model has overfit the training data. The negative R^2 value is particularly concerning as it suggests that the model's predictions are worse than random chance, making this degree of polynomial inappropriate for the dataset understudy. In conclusion, the simple linear model is giving the best results, while complex models are showing the overfitting problem.

The linear regression model (degree of freedom = 1) reveals interesting relationships between various network traffic features and the target variable Table 6. The most influential feature is 'orig_ip_bytes' with a strong negative coefficient of −7.748449. This is closely followed by 'orig_pkts' with a positive coefficient of 7.586293. The model shows moderate positive influences from 'resp_ip_bytes' (0.375515) and 'conn_state_OTH' (0.313253). The 'ts' feature has a negative coefficient (−0.2893260). The 'conn_state_S0' shows a negative correlation (−0.284247). The 'resp_bytes' has a positive coefficient (0.272970) while 'resp_pkts' has a negative coefficient (−0.190866), indicating complex interactions in the response traffic patterns. The 'conn_state_SF' has a slight negative coefficient (−0.104724), and finally, the 'duration' shows a small positive influence (0.104279) on the predicted outcome.

Table 6. Top 10 polynomial terms (degree of freedom = 1).

Polynomial Term	Coefficient
orig_ip_bytes	−7.748449
orig_pkts	7.586293
resp_ip_bytes	0.375515
conn_state_OTH	0.313253
ts	−0.289326
conn_state_S0	−0.284247
resp_bytes	0.272970
resp_pkts	−0.190866
conn_state_SF	−0.104724
duration	0.104279

7. Multivariate Measures

Multivariate statistical analysis is an approach that performs statistical analysis on multiple variables simultaneously. Compared to univariate analysis, it examines the relationships between multiple variables. Analyzing variables separately is likely to miss key features and patterns in the data. Patterns and structures in the data are more likely to be discovered through the relationships between variables.

There are a number of approaches to performing multivariate data analysis: DBSCAN, PCA, and FA etc. The aim of these approaches is to uncover any signal in the data, despite the presence of noise, and to present the information that the data possesses.

7.1. DBSCAN

DBSCAN, an unsupervised clustering algorithm, identifies clusters of closely situated points within a dataset. Unlike some other clustering algorithms, it does not require the

number of clusters to be specified in advance. DBSCAN defines clusters as highly dense regions, while low-density regions are identified as noise. The algorithm operates around three main components: core points, border points, and noise points. Core points are those that have a minimum number of neighboring points, referred to as min_samples, within a specified distance, known as ε (epsilon). Border points, on the other hand, are within the epsilon distance of a core point but do not meet the min_samples requirement. Noise points are those that are not part of any identified cluster. The outcome of the DBSCAN algorithm depend on the selection of epsilon and min_samples parameters. In this research, these parameters were determined through experimentation with various values, followed by observation and evaluation to identify the best fit for our dataset.

The DBSCAN algorithm identified a total of six clusters in the dataset, which are labeled from 0 to 5, as shown in Figure 5. Points that are not part of any cluster are labeled as -1. To assess the quality of the clusters formed by the DBSCAN algorithm, we calculated the Silhouette score. This metric quantifies how similar a point is to its own cluster compared to other clusters. Its score ranges from -1 to $+1$. A score close to 1 indicates that a point is well-matched within its own cluster. As shown in Figure 6, the Silhouette scores for our dataset were calculated. We achieved a score of 0.8, indicating a good match within clusters, when DBSCAN identified six clusters in the dataset.

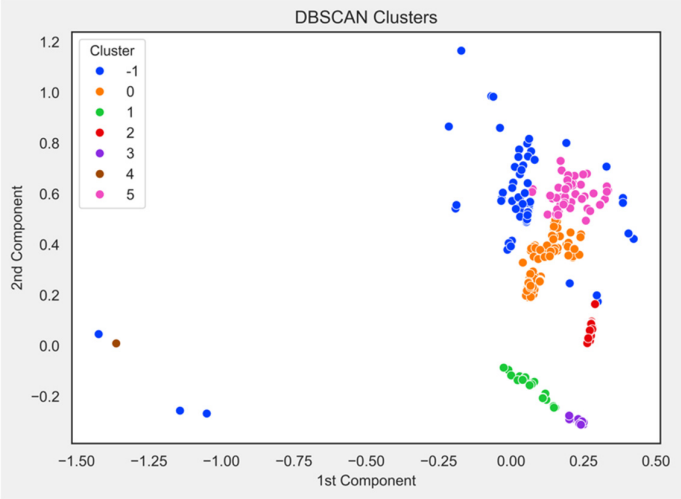

Figure 5. Clusters identified by DBSCAN.

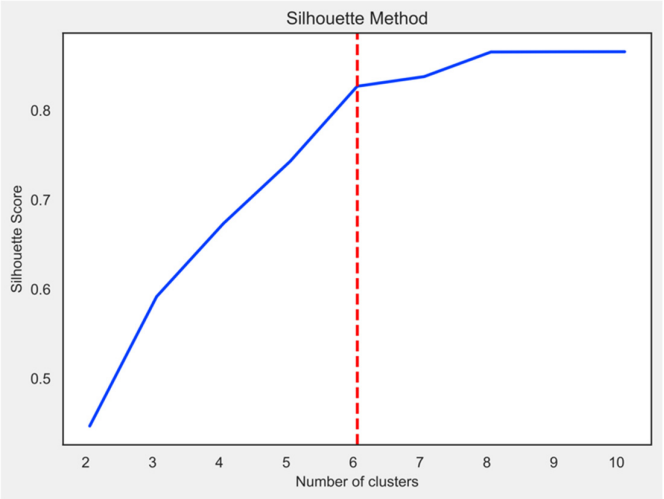

Figure 6. Silhouette score (DBSCAN).

7.2. Principal Component Analysis

PCA is a widely used dimensionality reduction technique that performs orthogonal linear transformation of correlated variables into uncorrelated features. These new features are called principal components. The PCA technique works to preserves the variance of original high-dimensional data into low-dimensional principal components. The computation of principal components involves a sequence of mathematical operations. First, the covariance matrix is computed for the input variables to understand their relationships (Equation (14)). Second, the eigenvectors and eigenvalues of the covariance matrix are calculated, this step is essential for determining the directions of maximum variance (Equation (15)). Third, these eigenvectors are arranged in order based on their corresponding eigenvalues which identifies the principal components in decreasing order of their significance.

$$cov\left(X_i,\ X_j\right) = \frac{\sum_{k=1}^{N}\left(X_{ik} - \overline{X}_i\right)\left(X_{jk} - \overline{X}_j\right)}{N-1} \tag{14}$$

X_i and X_j are the input variable and N is the number of observations.

$$(C - \lambda I)v = 0 \tag{15}$$

C is the covariance matrix, λ is representing eigen values, v is representing eigen vectors, and I is the identity matrix.

The PCA technique is employed in this research to reduce the dimensionality of a large dataset while preserving most of the original information. This simplification enhances data visualization and facilitates pattern identification. PCA does not require label information to compute its results, thereby signifying that it operates in an unsupervised manner. In our study, we transformed fourteen input variables into seven principal components. These components captured 61.05%, 24.04%, 6.99%, 3.95%, 2.22%, 0.46%, and 0.37% of the variance, respectively. Cumulatively, they accounted for 99% of the variance in the dataset. We created a visualization of the dataset using the first three components, as shown in Figure 7. This plot reveals clusters of data points with identical labels. Data points correspond to label 4, which combines seven rare labels as detailed in Table 2, are scattered throughout the plot. This distribution substantiates that these points belong to diverse label types.

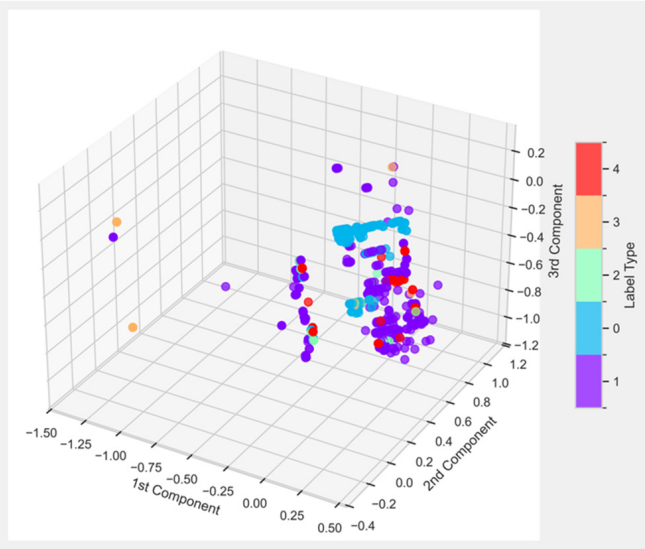

Figure 7. Principal Component Analysis (first three components).

7.3. Factor Analysis

Factor analysis is a statistical technique used for dimension reduction where numerous observed variables are represented by a smaller number of unobserved variables (i.e., factors or latent variables). Factor analysis is a latent variable model that infers factors that cannot be directly measured but are identified through correlations in observed variables. It focuses on common variance to uncover latent constructs. Main objectives of factor analysis are minimizing the data dimension, consolidating them into fewer factors, identifying underlying constructs that are not directly measurable (latent variables), and simplification of related items by grouping numerous correlated variables into factors that capture shared variance.

$$X_i = \sum_{j=1}^{m} \lambda_{ij} F_j + \epsilon_i \tag{16}$$

where X_i is the independent variable, λ_{ij} represents factor loadings, F_j represents factor, and ϵ_i is the error term.

To assess the data's suitability for factorization, we performed the Kaiser-Meyer-Olkin (KMO) test. This test measures if the correlations between the variables are sufficiently large. The KMO scores range from 0 to 1. We achieved 0.500 overall score.

For factor extraction we computed and sorted the eigenvalues of the correlation matrix of input features in descending order, visualizing them through a scree plot. The scree plot (Figure 8) illustrates the eigenvalue distribution across potential factors, with a red horizontal line representing the Kaiser Criterion (eigenvalue = 1). We implemented a factor selection process based on the Kaiser Criterion, retaining factors with eigenvalues exceeding 1.0.

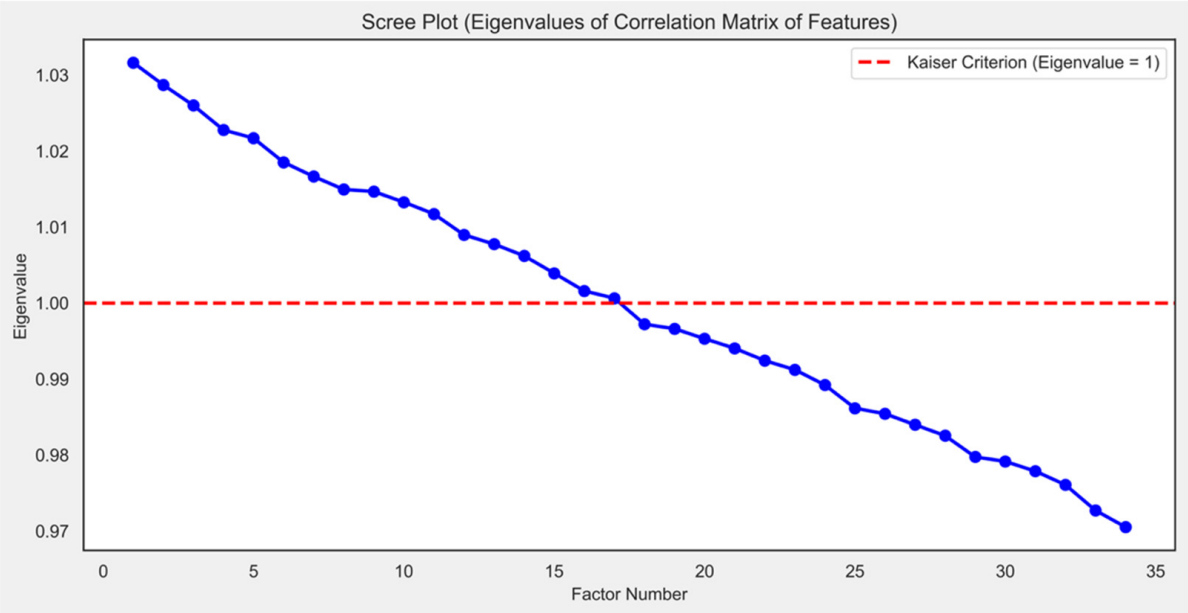

Figure 8. Scree Plot.

The factor analysis results (Table 7) reveal a weak factor structure in the dataset, characterized by very low explained variance values across all 17 factors. The first factor (Factor_1) explains only about 1.29% of the total variance. The pattern shows a gradual decline in explained variance from Factor_1 through Factor_17, with each subsequent factor contributing slightly less to the overall explanation of the data's variability. The cumulative variance explained by all 17 factors sums to 11.63%. It suggests that the factor analysis may not be the most effective method for capturing the underlying structure of the dataset.

Table 7. Factors and explained variance.

Factors	Explained Variance
Factor_1	0.012912
Factor_2	0.011695
Factor_3	0.010709
Factor_4	0.010184
Factor_5	0.009081
Factor_6	0.008745
Factor_7	0.007463
Factor_8	0.007242
Factor_9	0.006800
Factor_10	0.006024
Factor_11	0.005795
Factor_12	0.004651
Factor_13	0.004341
Factor_14	0.003569
Factor_15	0.003265
Factor_16	0.002491
Factor_17	0.001346
Total	0.116313

8. Comparison with Established Methods

In comparison to previous studies, our research offers several methodological advancements (Table 8). Unlike Moustafa & Slay (2016) [13], which focused on a limited set of descriptive and inferential measures, our study broadens the descriptive analysis by including mean, variance, and standard deviation, providing a more comprehensive view of data central tendency and variability. We also introduce polynomial regression as an inferential tool, enabling the modeling of nonlinear relationships, which is particularly beneficial for complex IoT environments. Additionally, our study incorporates advanced multivariate techniques (DBSCAN, PCA, and FA) allowing for deeper exploration of data structure and improved anomaly detection.

Compared to Damasevicius et al. (2020) [14], our study offers a more thorough descriptive statistical analysis, which is essential for understanding IoT data properties. While both studies use DBSCAN for clustering, we further enhance multivariate analysis by adding PCA and Factor Analysis, supporting more robust dimensionality reduction and latent variable discovery. The use of polynomial regression also enables us to capture complex, nonlinear patterns in IoT traffic.

Relative to Booij et al. (2021) [15], our application of polynomial regression enriches the inferential analysis by capturing nonlinear trends that correlation-based methods may overlook. Our use of DBSCAN, PCA, and Factor Analysis for multivariate analysis represents a significant methodological advancement, facilitating the discovery of complex patterns and latent structures in the data.

Table 8. Comparison with existing research.

Study	Dataset	Descriptive Measures	Inferential Measures	Multivariate Measures
(Moustafa & Slay, 2016) [13]	UNSW-NB15	Z-score, K-S test, Skewness, Kurtosis	PCC, GR	n/a
(Damasevicius et al., 2020) [14]	LITNET-2020	n/a	PCC, IGR	t-SNE, DBSCAN
(Booij et al., 2021) [15]	ToN-IoT	Z-score, K-S test, Skewness, Kurtosis	PCC, IGR	n/a
(Sharmila et al., 2024) [7]	RT-IoT2022	Z-score, K-S test, Skewness, Kurtosis	PCC, IGR	n/a
This study	IoT-23	Mean, Var, SD, Skewness, Kurtosis, K-S test	PCC, GR, Polynomial Regression	DBSCAN, PCA, Factor Analysis

When compared to Sharmila et al. (2024) [7], our study not only covers all the descriptive and inferential measures used in their work but also adds mean, variance, and standard deviation, resulting in a more complete statistical profile. The inclusion of polynomial regression allows for the exploration of nonlinear relationships, and the use of DBSCAN, PCA, and Factor Analysis provides a significant advantage in identifying clusters, reducing dimensionality, and uncovering latent variables. This comprehensive analytical framework enables our study to extract deeper insights and address more complex research questions than previous approaches.

9. Conclusions

This research successfully conducted a comprehensive statistical and multivariate analysis of the IoT-23 dataset, establishing important benchmarks for IoT security research through systematic evaluation of multiple analytical techniques. The study achieved its primary objectives of characterizing IoT network traffic data properties, comparing analytical method effectiveness, and providing practical guidance for cybersecurity applications.

The descriptive analysis confirmed that IoT network traffic exhibits pronounced non-normal distributions with extreme skewness and kurtosis values, establishing the complex nature of IoT data and the need for robust analytical approaches. The comparative evaluation revealed significant differences in technique effectiveness: PCA demonstrated superior performance for dimensionality reduction, DBSCAN proved highly effective for pattern identification, while factor analysis showed limited applicability for this dataset type. Linear regression provided optimal modeling performance compared to higher-degree polynomial alternatives.

Several limitations must be acknowledged. First, only three multivariate techniques were evaluated, excluding other potentially effective methods such as advanced clustering algorithms or deep learning approaches. Second, computational complexity and scalability considerations were not addressed, which are crucial for real-time IoT security applications.

These limitations can be addressed through multiple approaches. Methodological expansion should incorporate additional techniques including ensemble methods and deep learning-based dimensionality reduction. Computational efficiency studies should evaluate performance under varying dataset sizes and real-time constraints.

This research opens several promising avenues for investigation. Cross-dataset validation studies can establish universal applicability across different IoT environments. Hybrid

analytical approaches combining multiple techniques warrant further investigation. Deep learning integration offers potential for sophisticated pattern recognition capabilities.

This comprehensive analysis has established important methodological benchmarks and demonstrated the effectiveness of specific techniques for IoT security research. The evidence-based framework provides researchers with concrete guidance for algorithm selection and supports the development of more robust cybersecurity solutions. As IoT deployments continue expanding across critical infrastructure, these insights contribute to safer and more reliable IoT operations.

Author Contributions: Conceptualization, H.G.; methodology, H.G.; software, H.G.; validation, H.G.; formal analysis, H.G.; investigation, H.G.; resources, H.G.; data curation, H.G.; writing—original draft preparation, H.G.; writing—review and editing, S.S., B.V. and H.G.; visualization, H.G.; supervision, S.S. and B.V.; project administration, S.S., B.V. and H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in http://doi.org/10.5281/zenodo.4743746 (accessed on 3 May 2025).

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

References

1. Statista. Internet of Things Market Outlook Report. Statista. 2024. Available online: https://www.statista.com/study/109197/internet-of-things-market-outlook-report/ (accessed on 2 May 2025).
2. Wang, L.; Hsu, H.H. IoT technology in maritime logistics management: Exploration of data analysis methods. *Discov. Internet Things* **2025**, *5*, 66. [CrossRef]
3. Ebrahem, O.; Dowaji, S.; Alhammoud, S. Towards a minimum universal features set for IoT DDoS attack detection. *J. Big Data* **2025**, *12*, 88. [CrossRef]
4. Elkhadir, Z.; Begdouri, M.A. Enhancing IoT Security: A Comparative Analysis of Preprocessing Techniques and Classifier Performance on IoT23 and CIC IoT 2023 Datasets. *IAENG Int. J. Comput. Sci.* **2025**, *52*, 995–1010.
5. Al-Zewairi, M.; Almajali, S.; Ayyash, M.; Rahouti, M.; Martinez, F.; Quadar, N. Multi-Stage Enhanced Zero Trust Intrusion Detection System for Unknown Attack Detection in Internet of Things and Traditional Networks. *ACM Trans. Priv. Secur.* **2025**, *28*, 1–28. [CrossRef]
6. Aqil, N.; Zaki, F.; Afifi, F.; Hanif, H.; Kiah, M.L.M.; Anuar, N.B. Improved temporal IoT device identification using robust statistical features. *PeerJ Comput. Sci.* **2024**, *10*, e2145. [CrossRef] [PubMed]
7. Sharmila, B.S.; Nandini, B.M.; Kavitha, S.S. Performance evaluation of parametric and non-parametric machine learning models using statistical analysis for RT-IoT2022 dataset: Parametric and non-parametric machine learning models. *J. Sci. Ind. Res. (JSIR)* **2024**, *83*, 864–872.
8. Li, J.; Othman, M.S.; Chen, H.; Yusuf, L.M. Cybersecurity Insights: Analyzing IoT Data Through Statistical and Visualization Techniques. In Proceedings of the 2024 International Symposium on Parallel Computing and Distributed Systems (PCDS), Singapore, 21–22 September 2024; IEEE: New York, NY, USA, 2024; pp. 1–10.
9. Smiesko, J.; Segec, P.; Kontsek, M. Machine recognition of DDoS attacks using statistical parameters. *Mathematics* **2023**, *12*, 142. [CrossRef]
10. Kim, Y.G.; Ahmed, K.J.; Lee, M.J.; Tsukamoto, K. A Comprehensive Analysis of Machine Learning-Based Intrusion Detection System for IoT-23 Dataset. In Proceedings of the International Conference on Intelligent Networking and Collaborative Systems, Sanda-Shi, Japan, 7–9 September 2022; Springer International Publishing: Cham, Switzerland, 2022; pp. 475–486.
11. Chakraborty, S.; Khayer, N.; Ahmed, T. Assessing critical factors affecting the mass adoption of IoT in Bangladesh. In Proceedings of the International Conference on Mechanical Industrial & Energy Engineering, Khulna, Bangladesh, 19–21 December 2020; pp. 21–26.
12. Garcia, S.; Parmisano, A.; Erquiaga, M.J. IoT-23: A labeled dataset with malicious and benign IoT network traffic (Version 1.0.0) [Data set]. *Zenodo* **2020**. Available online: https://www.stratosphereips.org/datasets-iot23 (accessed on 2 May 2025).
13. Moustafa, N.; Slay, J. The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. *Inf. Secur. J. Glob. Perspect.* **2016**, *25*, 18–31. [CrossRef]

14. Damasevicius, R.; Venckauskas, A.; Grigaliunas, S.; Toldinas, J.; Morkevicius, N.; Aleliunas, T.; Smuikys, P. LITNET-2020: An annotated real-world network flow dataset for network intrusion detection. *Electronics* **2020**, *9*, 800. [CrossRef]
15. Booij, T.M.; Chiscop, I.; Meeuwissen, E.; Moustafa, N.; Den Hartog, F.T. ToN_IoT: The role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets. *IEEE Internet Things J.* **2021**, *9*, 485–496. [CrossRef]