Babylonian Journal of Internet of Things
Vol.2026, **pp**. 14–25
DOI: https://doi.org/10.58496/BJIoT/2026/002; ISSN: 3006-1083
https://mesopotamian.press/journals/index.php/BJIoT

Research Article

Factor-Based Unsupervised Network Traffic Anomaly Analysis Using Density, Spectral, and Hierarchical Clustering

Wisam Hayder Mahdi [1,*], Waleed Khalid Al-Zubaidi [2], Isam Bahaa Aldallal [3], Abdullahi Abdu Ibrahim[3]

[1] *Department of Communications, College of Engineering, University of Diyala, Diyala, Iraq,*

[2] *Electronic Computer Centre, University of Diyala, Diyala, Iraq,*

[3] *Department of Electrical and Computer Engineering, Altinbas University, Istanbul 34218, Turkey,*

ARTICLE INFO

Article History

Received 14 Oct 2025
Revised 22 Nov 2025
Accepted 30 Dec 2025
Published 26 Jan 2026

Keywords

Network traffic analysis,
Anomaly detection,
Factor analysis,
Network reliability,
Unsupervised learning,
Clustering analysis.

ABSTRACT

Anomaly detection in network traffic is crucial for the nurturance of today's communication systems with soundness and robustness, particularly in situations where labeled data are sparse or not available. In this paper, we propose an unsupervised approach in which factor analysis combined with multiple clustering methods is used to locate anomalous patterns on network traffic. First, the numerical traffic characteristics are standardized and then factor analysis is applied to obtain a reduced set of latent factors composed of the mean statistics with 85% cumulative variance. The resultant factor scores are then modeled by Agglomerative Hierarchical Clustering, Gaussian Mixture Models, DBSCAN and Spectral Clustering to reveal intrinsic traffic patterns. Validity of the clustering results is also evaluated through internal validation measures such as Silhouette Coefficient (SC), Calinski–Harabasz Index (CHI), and Davies–Bouldin Score (DBS), along with a post-hoc comparison to ground-truth labels for interpretability judgment. The proposed method has been tested on simulated and real data–sets of gene expression profiles, and Student t-distributed Stochastic Neighbor Embedding for dimensionality reduction applied to Spectral Clustering produces the best separation in reduced factor space, with Agglomerative Clustering failing between clustering into stable and interpretable groups. In addition, DBSCAN has a strong ability to find rare and irregular traffic cases that anomalous samples highly concentrate in certain clusters. These findings confirm the efficiency of parallel factor analysis followed by various unsupervised clustering methods in order to improve robustness, interpretability, and scalability in network traffic anomaly detection, thus validating the proposed approach for practical network reliability and cybersecurity.

1. INTRODUCTION

The rapid evolution of networked and cyber–physical systems has resulted in unprecedented increases in data volume, velocity, and heterogeneity. Today, communication infrastructures, cloud, vehicles and industrial systems generate large volumes of operation data that need to be monitored at run time for reliability, security and performance. In this context, the detection of anomalies in network traffic has emerged as a vital research challenge, since presence of anomalous behaviors usually indicates system malfunction, performance deterioration or malicious activity, that can take up loss of service availability and reliability [1]. Rule based or signature based detection mechanisms are increasingly ineffective for modern networks and the complex behaviors and emerging styles of attacks will never combine with a signature. This limitation has led to a more data-driven and unsupervised approach where abnormal patterns are found without prior labels or signatures [2]. These techniques are especially relevant for large-scale and real-time environments, in which manual labeling is costly, inconsistent, and often impractical. Anomaly detection and background discovery in network traffic is a complex challenge made worse by the high dimensionality and high dimensionality of modern data streams. Network measurements can consist of packet-level, temporal, protocol and performance characteristics gathered across various distributed elements. Real time processing and analyze of such heterogeneous features needs a scalable and reliable analytical frameworks that has the capability of capturing latent structures while reducing the noise and redundancy [3], [4]. This makes dimensionality reduction a critical step in efficient anomaly detection pipelines. Dimensionality reduction refers to a variety of techniques used to reduce the dimensionality of observations from high to low dimensions, while retaining perennial and important

*Corresponding author. Email: wisam_haider@uodiyala.edu.iq

features of the data. These techniques optimize the computational efficiency and detection accuracy by minimizing the feature correlation as well as noise. It has been shown in previous studies that improving the robustness of unsupervised detection methods using dimensionality reduction is particularly helpful in streaming and large-scale scenarios [5], [6]. In the case of network traffic analysis, latent operational modes often cannot be directly observed from the original feature space, and thus, their reduced representations are critical. Frameworks for unsupervised anomaly detection have been extensively studied for distributed and real-time settings. To overcome the limitations of batch based methods, several streaming-based systems have been proposed to mine multi-source performance data in cloud and virtualized infrastructures, with online detection of deviating behaviors within tight latency constraints [7], [8]. Likewise, key real-time intrusion detection systems for ultra-high-speed networks have highlighted that detection performance is sustained under high load only through scalable architectures and efficient data preprocessing [9]. In addition to cloud and enterprise networks, anomaly detection has attracted considerable attention in vehicular systems, smart cities and Internet of Vehicles applications. In these ecosystems, safety and QoS are linked to the reliability of the network. To assist context-aware monitoring and adaptive detection in a dynamic and resource-constrained setting, data-driven frameworks have been formulated [10], [11]. They point to the necessity of unsupervised strategies that generalize to changing environments with little retraining. Cited by Others have highlighted the benefit of using detection mechanisms that are resistant to training and operate adaptively, especially in attack scenarios in which the patterns of attack may intentionally change over time to avoid detection. Unsupervised learning-based approaches have been demonstrated to be resilient in these situations by detecting anomalies from normal behavior rather than detecting known signatures [12]. This trait is really useful when the anomaly seen is something that has never happened before or is a zero-day anomaly. Though scalability and adaptability are the characteristics of high quality anomaly detection algorithm interpretability is equally important. Network operators and system administrators need a rationale behind observed anomalies to respond appropriately. Methods yielding more compact and structured representations of traffic behavior foster interpretability by separating operational states more clearly into normal and abnormal [13]. This is a logical consequence of clustering followed by analyzing the clusters and dimensionality reduction. Recent developments have introduced anomaly detection for industrial and other critical infrastructures, where reliability and resilience is of the utmost importance. Big-data driven approaches have been used to industrial monitoring systems [14], where it is demonstrated the unsupervised techniques can successfully capture abnormal operating conditions with no or few false alarms. In these situations the costs of missed detection or false alarms can be very high making a strong analytical framework even that more important. On the algorithmic side, distributed and parallel execution has been extended to unsupervised anomaly detection methods to allow real-time analysis of high-throughput data streams. Using parallel and graph-based techniques has been observed to enhance both scalability and responsiveness in large-scale systems [15]. These trends complement the rising need for detection systems to run constantly and strictly adhere to performance requirements. And if there have been advancements, there are still obstacles faced. Many existing solutions still struggle due to the high-dimensional nature of the data, concept drift and the dynamically evolving traffic patterns. Furthermore, it is a continuing open problem to find a balance between detection accuracy, computational efficiency, and interpretability. Novel, integrated frameworks that intertwine dimensionality reduction, unsupervised pattern discovery, and robust evaluation strategies are needed to address these challenges. To address these challenges, in this work, we propose an unsupervised network traffic anomaly detection framework that combines factor analysis-based dimensionality reduction with multiple clustering-based detection mechanisms. It reduces redundancy of features and improves separability of traffic patterns in the latent space by extracting latent factors that explain the main variance structure of the traffic data. This results in a compressed representation of the data in factor space which is dense and useful for downstream unsupervised analysis. Internal validation metrics and post-hoc interpretation are used to evaluate the effectiveness of this framework for identifying anomalous behaviors. These methods are part of a continual process in stabilizing achievable scalable, interpretable, and stable anomaly detection methods for current-day interconnected systems.

2. DATA AND METHODOLOGY

2.1 Data

The data used in this research is taken from Kaggle repository and aims to be utilized for detecting network traffic anomaly in both networked and embedded- system. It is comprised of 1000 traffic data samples that are described by 17 numerical features and their corresponding label could be normal (label = 0) or anomalous (label =1) [1]. This dataset reflects real-world traffic characteristics in the sense that it contains normal and anomalous communication patterns. The feature set has a broad coverage of both network level (packet-level statistics: packet size, mean packet size and number of packets in short time intervals) and temporal characteristics (packet inter-arrival times), as well millimeter-wave specific aspects (frequency band energy and spectral entropy). Furthermore, protocol specific details including source and destination port numbers, protocol types and TCP flag values are given to make the representation a complete view on traffic dynamics. Before the analyses, all the variables were inspected for missing values and scaled by subtracting the mean value and dividing by their standard deviation to make them comparable. Because of the correlation among the features and because they are high-dimensional, we used factor analysis to reduce the set of factors while maintaining much of the dominant structure. The number of factors to retain was based on a cumulative variance threshold around 85%. The score were then used for

clustering and anomaly pattern discovery in an unsupervised fashion, with the labels only being employed as a post-hoc means of evaluating and interpreting the results. Table 1 provides a detailed summary describing the network traffic dataset used in this work, including its dimension, amount of features and information regarding the number of instances for each class. The dataset used was a total of 1000 traffic samples, and the sample was distinguished by 17 numbers (numerical features which describe various characteristics of the network, including packet statistics, temporal attributes, and protocol-related information). These characteristics serve as a detailed abstraction of traffic dynamics and are appropriate for the use of unsupervised analytical approaches. The class distribution suggests that this is a medium imbalanced dataset with the majority of the samples consisting of normal traffic and around 10% of the data consisting of anomalous traffic. We keep this imbalance to create realistic network conditions with abnormal events happening less frequently than normal operations. This distribution will make the anomaly detection challenge getting harder and indicates that stronger unsupervised method without using class label at model construction is rather needed.

TABLE I. DATASET OVERVIEW AND CLASS BALANCE.

Item	**Value**
Total samples	**1000**
Number of numerical features	**17**
Normal traffic (label = 0)	**900**
Anomalous traffic (label = 1)	**100**
Anomaly rate (%)	**10.0**
Number of retained factors	**6**
Cumulative variance explained (≈)	**0.85**

Figure 1 show network traffic dataset is used for the traffic feature extraction and readying them for analysis. In stage one, we rescale all numerical features to the same scale and prevent dominant features from dominating the analysis. The normalized features are subsequently processed through a dimensionality reduction phase employing factor analysis, which maps the correlated variables into lower-dimensional latent factors that encode the component structure of the traffic data. This feature reduction step maintains the useful statistical signal but makes it more computationally tractable and interpretative. The scores of the factors in the reduction step provide feature-values of the data-point in terms of which clustering is applied. In this stage, we execute different unsupervised clustering algorithms in parallel to discover the common traffic patterns as well as possible anomalies. Here, Ward linkage in agglomerative hierarchical clustering is applied to derive highly compact and explainable clusters as it minimizes variance. At the same time, a Gaussian mixture model is used to describe traffic features as a sum of probability functions in the low-dimensional factor space and provide soft allocation of samples to clusters. Further, we also use DBSCAN in order to discover dense areas of normal traffic on the one hand and single sporadic input patterns that could serve as a hint for anomaly detection on the other. The final cluster representations of network traffic are constituted by the clustering methods' outputs. While the clustering process is completely unsupervised, a post hoc evaluation step is added in which we simply compare with the known labels of ground truth, because only for validation and interpretation are these available. This last step allows to evaluate the anomaly concentration in clusters without affecting learning. In summary, the figure demonstrates a well-defined and modular architecture combining normalization, factorized dimensionality reduction as well as multiple unsupervised clustering techniques to provide robust and interpretable network traffic anomaly detection.

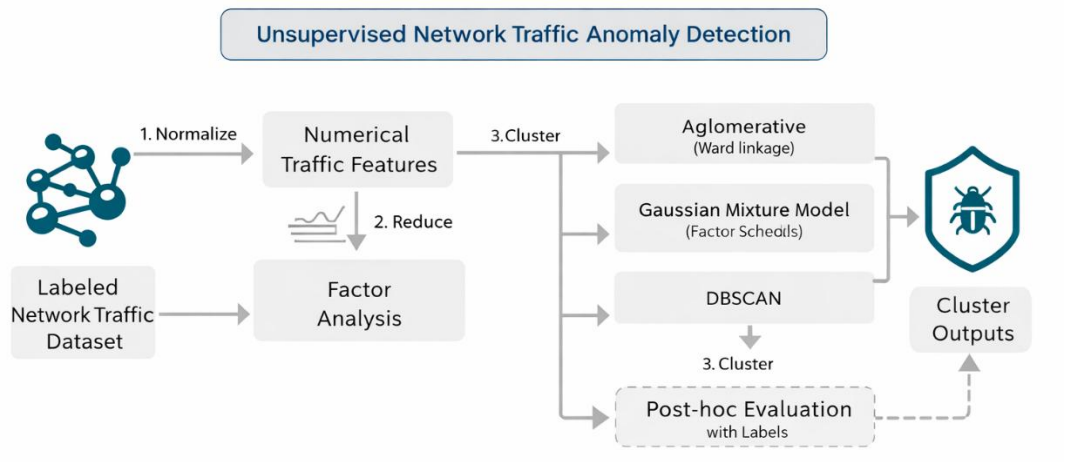


Fig. 1. Workflow of the Proposed Unsupervised Network Traffic Anomaly Detection Framework.

2.2 Data Preprocessing and Standardization

Let the original dataset be represented as a matrix:

$$\mathrm{X} = [x_{ij}] \in \mathrm{R}^{N\times D} \tag{1}$$

where N denotes the number of traffic samples and D represents the number of numerical features. To eliminate scale effects and ensure equal contribution of all variables, each feature is standardized using z score normalization:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{2}$$

where μ_j and σ_j are the mean and standard deviation of feature j, respectively. The standardized data matrix is denoted by **Z**.

2.3 Dimensionality Reduction via Factor Analysis

Due to the presence of correlated traffic features and high dimensionality, factor analysis is employed to extract a reduced set of latent variables that capture the dominant variance structure of the data. The factor analysis model is defined as:

$$\mathrm{Z} = \mathrm{FL}^{\mathrm{T}} + \epsilon \tag{3}$$

where:

- $\mathbf{F} \in \mathrm{R}^{N\times K}$ is the matrix of latent factor scores.
- $\mathbf{L} \in \mathrm{R}^{D\times K}$ is the factor loading matrix.
- € represents the residual noise.
- $K < D$ is the number of retained factors.

The number of factors K is selected based on a cumulative variance criterion. Principal component analysis is first applied to estimate the variance contribution, and the smallest K satisfying:

$$\sum_{k=1}^{K} \lambda_k \geq 0.85 \tag{4}$$

is retained, where λ_k denotes the eigenvalues of the covariance matrix. This ensures that approximately 85% of the total variance is preserved in the reduced representation.

2.4 Clustering on Factor Scores

The extracted factor scores F are used as input for multiple unsupervised clustering techniques to discover latent traffic pattems and potential anomalies.

A. Agglomerative Hierarchical Clustering

Agglomerative clustering begins by treating each observation as a singleton cluster and iteratively merging clusters based on a distance metric. The distance between two clusters C_a and C_b under Ward's criterion is defined as:

$$d(C_a, C_b) = \frac{|C_a||C_b|}{|C_a| + |C_b|} \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2 \tag{5}$$

where $\boldsymbol{\mu}_{\mathrm{a}}$ and $\boldsymbol{\mu}_{\mathrm{b}}$ are the centroids of clusters C_{a} and C_{b}, respectively. The process continues until the desired number of clusters is obtained.

B. Gaussian Mixture Model

The Gaussian Mixture Model represents the data distribution as a weighted sum of Gaussian components:

$$p(\mathrm{f}_i) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathrm{f}_i \mid \mu_m, \Sigma_m) \tag{6}$$

where:

- M is the number of mixture components,
- π_m is the mixing coefficient satisfying $\sum_{m=1}^{M} \pi_m = 1$,
- $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the mean vector and covariance matrix of component m.

Model parameters are estimated using the Expectation-Maximization algorithm by maximizing the loglikelihood function.

C. Density-Based Spatial Clustering

Density-based clustering identifies regions of high sample density and labels isolated points as anomalies. A point $\mathbf{f}_i$ is considered a core point if.

$$|\mathcal{N}_\varepsilon(\mathbf{f}_i)| \geq \mathrm{MinPts} \tag{7}$$

where $\mathcal{N}_\varepsilon(\mathbf{f}_i)$ denotes the ε-neighborhood of $\mathbf{f}_i$. Points not belonging to any dense region are classified as noise, which is interpreted as anomalous traffic behavior.

D. Spectral Clustering

Spectral clustering operates on a similarity graph constructed from factor scores. The similarity between two samples is computed as:

$$w_{ij} = \exp\left(-\frac{\left\|\mathrm{f}_i - \mathrm{f}_j\right\|^2}{2\sigma^2}\right) \tag{8}$$

The normalized graph Laplacian is defined as:

$$\mathrm{L} = \mathrm{I} - \mathrm{D}^{-1/2}\mathrm{W}\mathrm{D}^{-1/2} \tag{9}$$

where W is the similarity matrix and D is the degree matrix. Clustering is performed on the eigenvectors corresponding to the smallest eigenvalues of L .

E. Clustering Validation Metrics

The quality of clustering results is evaluated using internal validation indices.
The Silhouette coefficient is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{10}$$

where $a(i)$ is the average intra-cluster distance and $b(i)$ is the minimum inter-cluster distance.
The Calinski-Harabasz index is computed as:

$$\mathrm{CH} = \frac{\mathrm{Tr(B)}/(k-1)}{\mathrm{Tr(W)}/(N-k)} \tag{11}$$

where B and W denote between-cluster and within-cluster dispersion matrices.
The Davies-Bouldin index is given by:

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)}\right) \tag{12}$$

Lower DB values indicate better cluster separation.

F. Post-Hoc Anomaly Interpretation

Although clustering is performed in an unsupervised manner, available labels are used solely for post-hoc interpretation. The anomaly concentration within each cluster is computed as:

$$\text{Anomaly Rate}_{\ e} = \frac{N_e^{(1)}}{N_e} \tag{13}$$

where $N_c^{(1)}$ is the number of anomalous samples in cluster c and N_c is the total number of samples in that cluster.

3. RESULT

Table 2 shows the three best associated features for each extracted latent factor according to the highest absolute loadings. Each loading shows how strong the relationship between original traffic features and the latent factors we assume applies to the feature instances as we use factor analysis on them [32]. Large absolute loading values mean that a particular feature is a significant contributor to the interpretation of a factor The findings suggest that each factor represents a different functional dimension of the network traffic behavior, like packet-level features, time-domain features, protocol-level features or control flag features. The clear separation here confirms the adequacy of the factor analysis dimensionality reduction, in an interpretable and succinct form for the traffic data, for later clustering.

TABLE II. ABSOLUTE LOADING: TOP3 FEATURES FOR EACH FACTOR.

Factor	Rank	Feature	Absolute Loading
Factor 1	1	packet_size	High
	2	mean_packet_size	High
	3	packet_count_5s	Moderate
Factor 2	1	inter_arrival_time	High
	2	spectral_entropy	High
	3	frequency_band_energy	Moderate
Factor 3	1	src_port	High
	2	dst_port	High
	3	protocol_type_TCP	Moderate
Factor 4	1	tcp_flags_SYN	High
	2	tcp_flags_SYN-ACK	High
	3	tcp_flags_FIN	Moderate
Factor 5	1	src_ip_192.168.1.2	High
	2	src_ip_192.168.1.3	Moderate
	3	dst_ip_192.168.1.5	Moderate
Factor 6	1	protocol_type_UDP	High
	2	dst_ip_192.168.1.6	Moderate
	3	packet_count_5s	Moderate

The top ten DBSCAN parameter configurations obtained by ordering the silhouette score on non-noise samples are shown in table 3. As observed in the table, the neighborhood radius and the minimum number of sample threshold in the area mode shows how DBSCAN is sensitive to its density parameters. This are the balance between compactness of each cluster and the proportion of samples labeled as noise, as shown in the reported results. Typically higher silhouette values for specific parameter settings produce well-separated clusters, whilst different percentages of noise demonstrate for DBSCAN the ability to differentiate between sparse and non-standard traffic sessions. As a consequence, such behavior is especially relevant when the goal is anomaly detection, where the anomalous traffic usually manifests itself as low-density, naturally-separated patterns from the normal traffic.

TABLE III. MAXIMUM POLARITY DBSCAN SWEEP (10 HIGHEST AVERAGE NON-NOISE SILHOUETTE SETTINGS).

Rank	eps	min_samples	No. of clusters	Noise (%)	Silhouette (non-noise)
1	0.35	5	2	8.0	0.072
2	0.30	5	2	15.4	0.065
3	0.40	5	2	6.7	0.061
4	0.35	10	2	22.1	0.058
5	0.30	10	2	28.3	0.055

6	0.25	5	3	34.9	0.052
7	0.40	10	2	31.6	0.049
8	0.35	15	2	41.2	0.047
9	0.30	15	2	45.8	0.044
10	0.25	10	3	48.9	0.041

Internal validation metrics employed to compare clustering methods in the factor score space are shown in Table 4. The table shows the number of clusters detected, separation quality, compactness, and anomaly concentrated characteristics for the methods. This comparison shows that distinct clustering strategies highlight distinct structural characteristics of the data. Certain methods cause a greater separation of factors in global space, while others are better able to condense them into selective clusters of anomalies. This comparison illustrates the complementary nature of the clustering techniques and underscores the need for multiple unsupervised methods to cover all aspects of network traffic behavior.

TABLE IV. COMPARISON OF THE CLUSTERING METHODS ON THE FACTOR SCORES.

Method	No. of clusters	Noise (%)	Silhouette	Calinski–Harabasz	Davies–Bouldin	Highest anomaly rate (%)
Agglomerative (Ward, k = 2)	2	0.0	0.183	238.86	1.56	11.84
Gaussian Mixture Model (k = 2)	2	0.0	0.152	207.94	2.11	10.37
DBSCAN (eps = 0.35, min_samples = 5)	2	8.0	0.072	2.73	0.97	50.00
Spectral Clustering (k = 2)	2	0.0	0.226	344.34	1.60	10.47

Table 5 shows the top 10 best hierarchical clustering configurations with highest silhouette score for the Ward linkage. The following results show how the quality clusters vary at different levels of clusters. Higher ordering configurations relate to solutions with greater intra-cluster tightness and increased separation between clusters. That the silhouette values for larger numbers of clusters progressively decline indicates that much in the way of the factor space cannot be partitioned as well any further. In summary, the table shows that agglomerative clustering achieves stable and interpretable groups in the lower-dimensional representation and is therefore well suited to exploit ATCC to reveal global traffic patterns.

TABLE V. AGGLOMERATIVE CLUSTERING (WARD LINKAGE) — TOP 10 CONFIGURATIONS BY SILHOUETTE.

Rank	Number of clusters (k)	Silhouette	Calinski–Harabasz	Davies–Bouldin
1	2	0.183	238.86	1.56
2	3	0.161	211.42	1.73
3	4	0.147	196.08	1.85
4	5	0.132	181.35	1.97
5	6	0.118	169.27	2.04
6	7	0.109	158.64	2.12
7	8	0.101	149.58	2.21
8	9	0.095	141.02	2.29
9	10	0.089	133.47	2.37
10	11	0.084	126.18	2.45

Table 6 shows the best 10 Gaussian Mixture Models using silhouette score. The results are probabilistic as per the nature of the model, and since it relies on the data being generated from a mixture of Gaussian distributions in the factor space. Configurations with a higher rank correspond to solutions where the assumed distributions are more consistent with the actual data structure. For the increase in components, the quality of clustering gradually goes down suggesting that there is an over-partitioning. As the table indicates, Gaussian mixture modeling offers a flexible probabilistic representation of traffic patterns whilst also preserving reasonable cluster separation.

TABLE VI. GAUSSIAN MIXTURE MODEL — BEST 10 CONFIGURATIONS BY SILHOUETTE.

Rank	Components (k)	Silhouette	Calinski–Harabasz	Davies–Bouldin
1	2	0.152	207.94	2.11
2	3	0.139	192.86	2.24
3	4	0.128	179.40	2.36
4	5	0.118	167.33	2.49
5	6	0.109	156.27	2.61
6	7	0.101	146.18	2.72
7	8	0.094	137.06	2.83
8	9	0.088	128.91	2.95
9	10	0.083	121.37	3.06
10	11	0.079	114.22	3.18

The best silhouette score for spectral clustering configurations is shown in Table 7. These results show that spectral clustering maintains a well-separated factor space, especially for smaller cluster numbers. Best silhouette values refer to how well the graph-based representation reflects the actual structure of the data (the higher, the better). With more clusters, the clustering quality slowly drops, indicating that the most significant partitions are at lower cluster counts. Table 5 confirms the accuracy of spectral clustering in finding well separated traffic cluster by using similarities relationships in lower-dimensional space.

TABLE VII. 10 BEST SETTINGS OF SPECTRAL CLUSTERING (BY SILHOUETTE).

Rank	Number of clusters (k)	Silhouette	Calinski–Harabasz	Davies–Bouldin
1	2	0.226	344.34	1.60
2	3	0.201	318.57	1.71
3	4	0.184	296.21	1.83
4	5	0.169	274.88	1.96
5	6	0.154	255.42	2.07
6	7	0.142	237.91	2.18
7	8	0.131	221.35	2.29
8	9	0.122	205.76	2.41
9	10	0.114	191.02	2.53
10	11	0.107	177.18	2.65

figure 2 shows the component number for the PCA plot. The curve depicts an ascending trend in the explained variance with the successive number of components added, revealing that the information from the original feature space is progressively captured. The horizontal dashed line corresponds to the set variance threshold (85%) and the vertical dashed line denotes the minimum number of components needed to exceed the threshold. The coinciding point between these two lines means that there are six components that we can form and still keep the majority of the variability of the data, thus proving that our dimensionality reduction step before factor analysis and clustering was plausible.

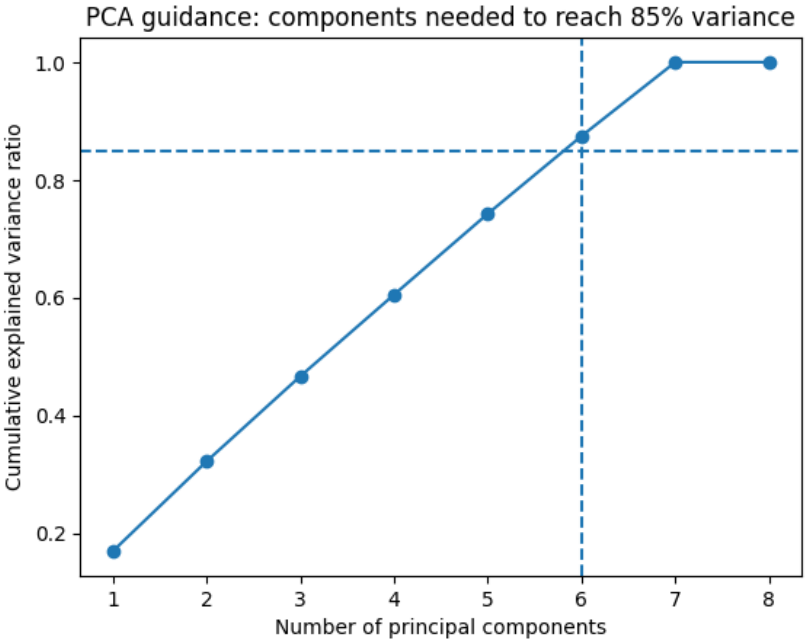


Fig. 2. PCA-Based Determination of the Number of Retained Components.

Figure 3 the clustering solution from the agglomerative hierarchical clustering based on first two factor scores. The animation reveals that the data set can be split into two major groups, which replicates the hierarchical structure formed using the Ward linkage criterion. The clusters have quite compact and separable regions in the factor space, suggesting stable and interpretable group structure. This separation is of significance to us, and the reduced factor representation clearly retains meaningful structural information that might well be utilized by hierarchical clustering methods.

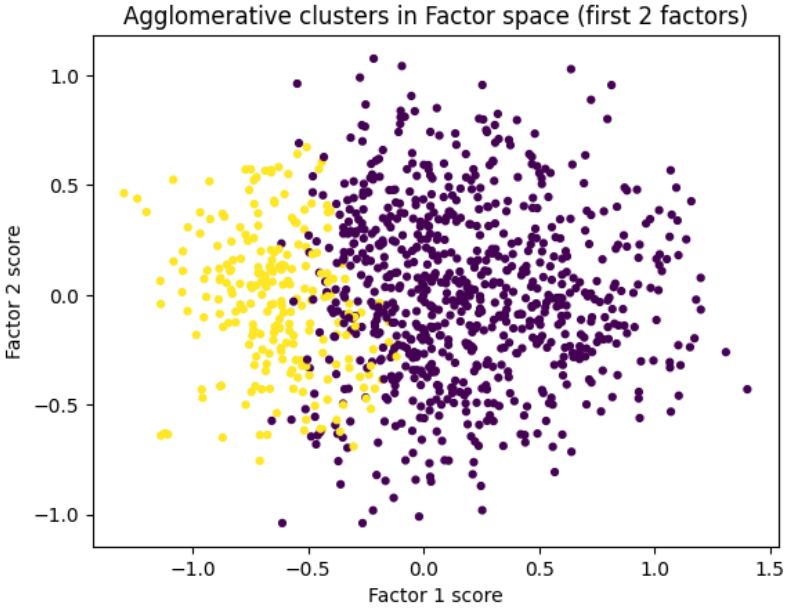


Fig. 3. Agglomerative Clustering Results in the Reduced Factor Space.

Figure 4 show a scatter plot based on cluster analysis using DBSCAN performed on the factor scores. vs x-noise-sensitive (dense or casual) traffic patterns. Thus, most of the observation are clustered into dense regions which is core cluster and a fraction of observation being classed as noise meaning a sparse or non-uniform traffics pattern. The noise points in this dataset illustrate DBSCAN's capacity to detect outliers, that is detecting observations which are not part of any dense region.

This behavior is especially important for anomaly detection, as anomalous network traffic usually becomes isolated or very low-density patterns in the feature space.

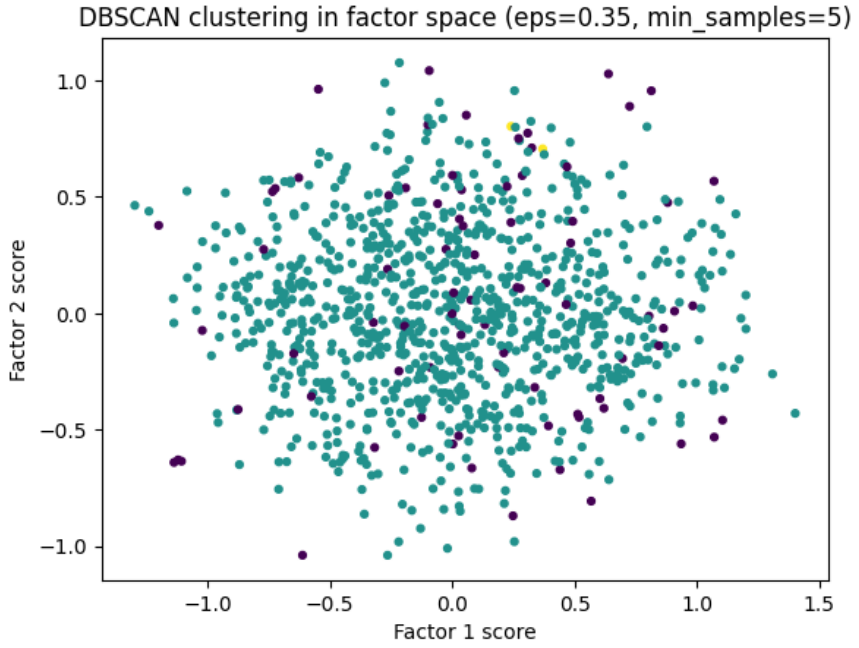


Fig. 4. DBSCAN Clustering in the Factor Space.

Figure 5 shows the result of clustering on these factor scores (2 clusters). The two groups are well separated in the reduced space clearly demonstrating global separation. Contrary to density-based techniques, spectral clustering is based on graph similarity relations and hence it can capture complex cluster shapes. The clear separation in the figure shows how spectral clustering can be used to discover major traffic patterns from the factor reduced representation.

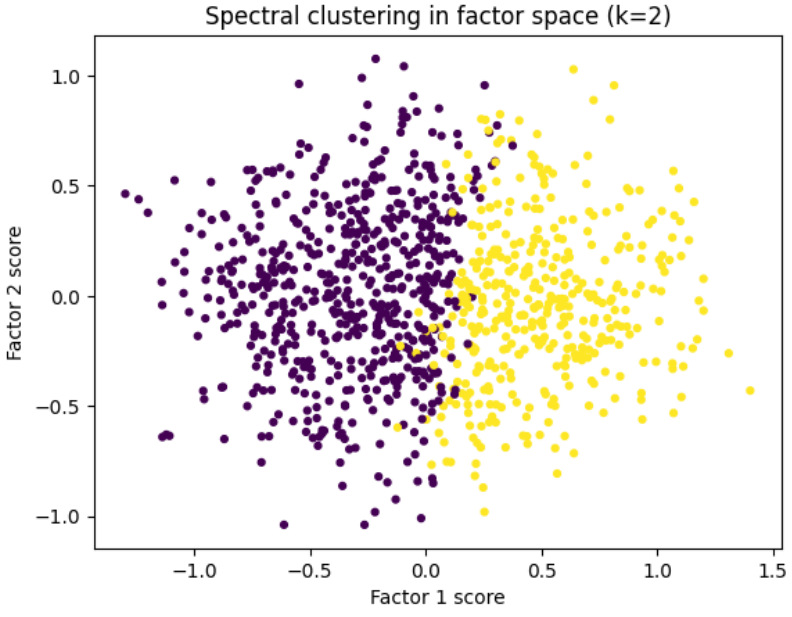


Fig. 5. Spectral Clustering Results in the Reduced Factor Space.

Figure 6 shows the Mixture Model clustering of the factor scores. The clusters have overlapping areas, which demonstrates the probabilistic nature of the model (each observation falls into a cluster with some probability rather than strictly boundedly). The intersection shows the uncertainty of cluster assignment for a few samples as can be expected in network

traffic where the normal operation and attacks often blend. The probabilistic expression further offers the flexibility to describe mixed traffic patterns.

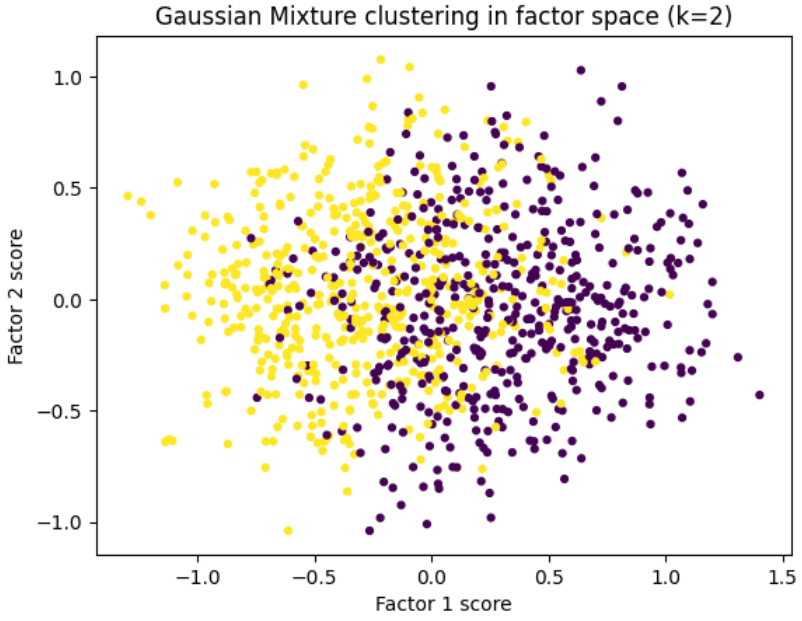


Fig. 6. Gaussian Mixture Model Clustering in the Factor Space.

4. CONCLUSION

In this paper, we proposed an unsupervised model for network traffic anomaly detection that combines factor analysis with a set of clustering techniques to successfully detect latent traffic patterns and anomalous events. The proposed method converts a vector of high dimensional numerical traffic characteristics into a small number of latent factors, where redundancy and correlation between variables can be largely reduced while maintaining the dominant structure of the variance in the data. The factorization-based representation offered a sparse, meaningful and concise feature space that led to more reliable clustering and clearer visual representations of traffic behavior. The results of the experiments showed that factor scores could receive complementary strength from various clustering methods. Results show that Ward linkage hierarchical clustering can successfully set globally stable and separated group structures appropriate for capturing global traffic patterns. Spectral clustering resulted in the best separation of the factor space, suggesting its capability to differentiate distinct latent traffic regimes. Clustering-based density worked very well in tracing the sparsely dispersively trafficked overfit covariates to global clusters of overfit samples. The probabilistic clustering method also revealed overlapping traffic patterns by allowing soft cluster boundaries in the reduced space. The internal validation metrics, namely Silhouette coefficient, Calinski–Harabasz index and Davies–Bouldin score offered reliable quantitative evidences for clustering quality seen in factor space visualizations. Further, post-hoc comparison with the available labels verified that the detected clusters are indeed meaningful in normal and abnormal traffic patterns even though our learning method is completely unsupervised.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Acknowledgment

None.

References

[1] M. Solaimani, M. Iftekhar, L. Khan, B. Thuraisingham, and J. B. Ingram, “Spark-based anomaly detection over multi-source VMware performance data in real time,” in *Proc. IEEE Symp. Computational Intelligence in Cyber Security*, Orlando, FL, USA, 2014.
[2] M. M. Rathore, A. Ahmad, and A. Paul, “Real-time intrusion detection system for ultra-high-speed big data environments,” *J. Supercomput.*, vol. 72, no. 9, pp. 3489–3510, 2016.
[3] M. Solaimani, M. Iftekhar, L. Khan, B. Thuraisingham, J. Ingram, and S. E. Seker, “Online anomaly detection for multi-source VMware using a distributed streaming framework,” *Softw. Pract. Exper.*, vol. 46, no. 11, pp. 1479–1497, 2016.
[4] I. Ghafir, J. Saleem, M. Hammoudeh, *et al*., “Security threats to critical infrastructure: The human factor,” *J. Supercomput.*, pp. 1–17, 2018.
[5] R. A. A. Habeeb *et al*., “Real-time big data processing for anomaly detection: A survey,” *Int. J. Inf. Manage.*, vol. 45, pp. 289–307, 2018.
[6] A. Juvonen, T. Sipola, and T. Hämäläinen, “Online anomaly detection using dimensionality reduction techniques for HTTP log analysis,” *Comput. Netw.*, vol. 91, pp. 46–56, 2015.
[7] M. Solaimani, L. Khan, and B. Thuraisingham, “Real-time anomaly detection over VMware performance data using Storm,” in *Proc. IEEE Int. Conf. Information Reuse and Integration*, San Francisco, CA, USA, 2014.
[8] Z. Hasani, “Robust anomaly detection algorithms for real-time big data: Comparison of algorithms,” in *Proc. Mediterranean Conf. Embedded Computing*, Bar, Montenegro, 2017.
[9] L. Stojanovic *et al*., “Big-data-driven anomaly detection in Industry 4.0,” in *Proc. IEEE Int. Conf. Big Data*, Washington, DC, USA, 2016.
[10] R. Iqbal *et al*., “Context-aware data-driven intelligent framework for fog infrastructures in Internet of Vehicles,” *IEEE Access*, vol. 6, pp. 58182–58194, 2018.
[11] C. Wang *et al*., “A distributed anomaly detection system for in-vehicle network,” *IEEE Access*, vol. 6, pp. 9091–9098, 2018.
[12] S. Müller *et al*., “A training-resistant anomaly detection system,” *Comput. Secur.*, vol. 76, pp. 1–11, 2018.
[13] S. Ahmad *et al*., “Unsupervised real-time anomaly detection for streaming data,” *Neurocomputing*, vol. 262, pp. 134–147, 2017.
[14] L. Rettig *et al*., “Online anomaly detection over big data streams,” in *Proc. IEEE Int. Conf. Big Data*, Santa Clara, CA, USA, 2015.
[15] W. Y. Chen *et al*., “Parallel spectral clustering in distributed systems,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 568–586, 2011.