ILKOM Jurnal Ilmiah Vol. 17, No. 3, December 2025, pp. 323-335
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article **Open Access (CC–BY-SA)**

K-Means and K-Medoid in Clustering Analysis of Network Congestion Level

Herdianti Darwis[a,1,*]; Purnawansyah[a,2]; Alfi Syahrin Umalekhoa[a,3]; Adam Adnan [a,4]; Yulita Salim[b,5]; Fitriyani Umar[a,6]; Roesman Ridwan Raja[c,7]; Muh. Aqil Fajar AR.[a,8]
[a] *Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM 5, 90231, Makassar, Indonesia*
[b] *Universiti Kuala Lumpur, 1016 Jl. Sultan Ismail, Bandar Wawasan, 50250, Kuala Lumpur, Malaysia*
[c] *Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, 820-8502, Fukuoka, Japan*
[1]*herdianti.darwis@umi.ac.id;* [2]*purnawansyah@umi.ac.id;* [3]*alfiumalekhoa@gmail.com,* [4]*adamadnan.iclabs@umi.ac.id;*
[5]*yulita.salim@s.unikl.edu.my;* [6]*fitriyani.umar@umi.ac.id;*[7]*raja.roesman-ridwan757@mail.kyutech.jp;* [8]*13020200294@umi.ac.id*

Article history: Received February 10, 2024; Revised May 07, 2024; Accepted December 04, 2025; Available online December 21, 2025

Abstract

This research investigates the application of clustering techniques to network congestion data at Universitas Muslim Indonesia, employing a hybrid metric approach based on packet loss and delay. The study utilized two algorithms, K-Means and K-Medoid, applied in a semi-supervised scenario to group 255,147 network data points into 3, 4, and 5 clusters, considering 10 principal variables. During the pre-processing phase, data cleansing was conducted to address missing values, followed by normalization to standardize the scale of numerical variables, thereby preparing the data for the clustering process. Model validation was performed using four cluster evaluation methods: Gap Statistic, Davies-Bouldin Index, and Elbow Method. The evaluation results indicate that both algorithms were capable of forming valid and reliable clusters. However, the K-Means algorithm demonstrated superior performance compared to K-Medoid, particularly when utilizing three Quality of Service variables: throughput, packet loss, and delay. In this configuration, K-Means yielded more stable clusters, a clearer separation between clusters, and a more structured visualization. Consequently, K-Means is considered more optimal for classifying network congestion levels and presents an effective approach for network data segmentation.

Keywords: Network Congestion Clusters; Davies Bouldin Index; Elbow method; Gap-Statistic; K-Means; QoS Parameters

Introduction

Similar to traffic jams in transportation systems caused by accidents, repairs, or other inevitable factors, network communications is also often experience congestion. This problem often arises due to bandwidth allocation errors, where the number of data packets sent exceeds the available network bandwidth capacity. In computer networks, bottlenecks can be triggered by high node density, many-to-many data transmission architectures, and packet collisions, resulting in reduced packet delivery rates, reduced throughput, increased packet loss and delays, decreased energy efficiency, and connection blockages. These factors always decrease the Quality of Network Service (QoS) and Quality of Experience (QoE) of users AAQ [1]–[6].

Some strategies, such as regulation, identification, maintenance, and administration, are often used to address bottlenecks at all layers of the network [7]. In the Open System Interconnection (OSI) model, the transport layer, a fundamental element in Internet communications architecture, assumes an important role. This layer, one of seven in the OSI model, oversees end-to-end data transmission in the network [8], [9].

Typically, Congestion Control (CC) is executed in two scenarios: preventing congestion first, being recognized as an open-loop control system, and overcoming bottlenecks that arise, identified as a closed-loop control system. Both situations are managed through two main mechanisms: Transmission Control Protocol (TCP) regulation and queue administration, specifically Active Queue Management (AQM) applied to routers [10], [11].

Basically, the effectiveness of network protocols depends on the congestion control (CC) algorithm used. Various algorithms have been designed to adjust metrics according to specific requirements. For example, delay-loss based congestion control (CC) algorithms combine packet loss and delay into a single congestion metric to optimize network performance [12]–[14].

http://dx.doi.org/10.33096/ilkom.v17i3.2083.323-335

Universitas Muslim Indonesia (UMI) is a private higher education institution located in Makassar City, founded on June 23 1954, UMI is also marked by its proven academic achievements as the only private university in eastern Indonesia to obtain A level institutional accreditation. In the field of education, UMI uses the internet network as one of the supporting facilities and infrastructure. UMI has 13 faculties, where each faculty uses a wireless LAN network as a facility for students [15]. This study aims to perform a clustering analysis to facilitate the labeling and modeling of network congestion data within the network infrastructure of Universitas Muslim Indonesia. It employs hybrid metrics—combining packet loss and delay-based congestion—which are not adequately addressed by the TIPHON standard, as it categorizes network congestion using either packet loss or delay independently. K-Means, previously utilized for clustering analysis in related studies [16], has been adopted in this research to enhance the classification process. The study specifically investigates scenarios involving varying numbers of input features, namely 10, 6, and 3. Furthermore, the clustering results are thoroughly validated to ensure that the data used in the analysis is both valid and reliable.

Method

A. Network Congestion

Network congestion refers to a condition in which data transmission slows down due to an error in the allocation of bandwidth resources, where the number of data packets sent exceeds the available bandwidth capacity. The presence of congestion on the network can result in a decrease in QoS service quality and QoE user experience. QoS itself includes the ability to provide resources and differentiation on network services, measured through parameters such as packet loss, delay, and throughput.

1. Packet loss is a parameter that describes a condition that shows the number of data packets loss due to collusion or congestion in the network. The unit used is (%). The purpose of measuring packet loss is to see the reliability of the method used in sending packets in the event of network-congestion.
2. Delay is the time it takes for data to travel the distance from origin to destination. The purpose of delay measurement is to find out how long it takes for one packet to get from source to destination.
3. Throughput can be defined as the amount of time unity data sent in a network, from one network point to another. The purpose of measuring throughput is to determine whether the network relies on forwarding packets that come to their destination.
4. Bandwidth refers to the capacity of a communication channel to transmit data over a period of time, generally measured in bits per second (bps) or multiples such as Kbps, Mbps, or Gbps. This is a crucial element in determining the speed and efficiency of data transmission in a network.

Based on Telecommunications and Internet Protocol Harmonization over Network (TIPHON), the QoS parameter assessment standard issued by the European Telecommunications Standards Institute (ETSI), packet loss and delay standards are set as shown in **Table 1** [17].

Table 1. TIPHON Standard

Parameter	Grade	QoS Category
Packet loss	0%	Excellent
	3%	Good
	15%	Medium
	25%	low
Delay	<150ms	Excellent
	150ms-300ms	Good
	300ms-450ms	low

B. Data Mining

Data mining is a set of techniques for finding previously unknown knowledge in large databases. The patterns found can be used to help make a decision. Data mining is also defined as the process that uses static, mathematical, artificial intelligence, and machine learning techniques to ecstasy and identify useful information and knowledge derived from large databases or data warehouses [18].

C. Clustering

Clutering is one of the multivariate techniques of interdependence. Therefore, in clustering analysis between independent variables and dependent variables, there is no distinction between one another [19]. Clustering analysis is a process used as a method of combining observations into groups, so that each homogeneous set will have certain characteristics [20]. Thus the observations in each group are the same as other observations in the same group. While each set should be different from other sets. Thus an observation in one set should be different from an observation in another set [21].

D. Dataset

The source of data used in this study is primary data collected from the Indonesian Muslim University network using Wireshark in 10 working days from 06.00 AM to 07.00 PM. The data collected as many as 633 captures have been aggregated into a dataset with approximately 652247 records (milliseconds) and carried out a pre-processing by removing noise and missing values so that the remaining 255147 records. The dataset comprises ten principal variables (X_{it}) that characterize network traffic behavior over time. $\mathrm{X_{1t}}$ denotes the total number of packets transmitted, whereas $\mathrm{X_{2t}}$ reflects the proportion of packets lost during communication. Network efficiency is further described through throughput ($\mathrm{X_{3t}}$) and end-to-end delay ($\mathrm{X_{4t}}$). The magnitude of transmitted information is represented by the file size in bytes ($\mathrm{X_{5t}}$) and the corresponding data size in bytes ($\mathrm{X_{6t}}$) Moreover, the dataset includes the data byte rate ($\mathrm{X_{7t}}$) and data bit rate ($\mathrm{X_{8t}}$), which quantify the flow of information per second. Finally, $\mathrm{X_{9t}}$ captures the mean packet size, while $\mathrm{X_{10t}}$ specifies the average packet transmission rate. Collectively, these attributes provide a detailed depiction of temporal fluctuations in network traffic performance.

E. Research Design

The research design carried out was clusterization analysis using the K-Means method with cluster validation using the Gap-Statistic Elbow Method and Davies bouldin Index. The analysis flow chart is shown in **Figure 1**.

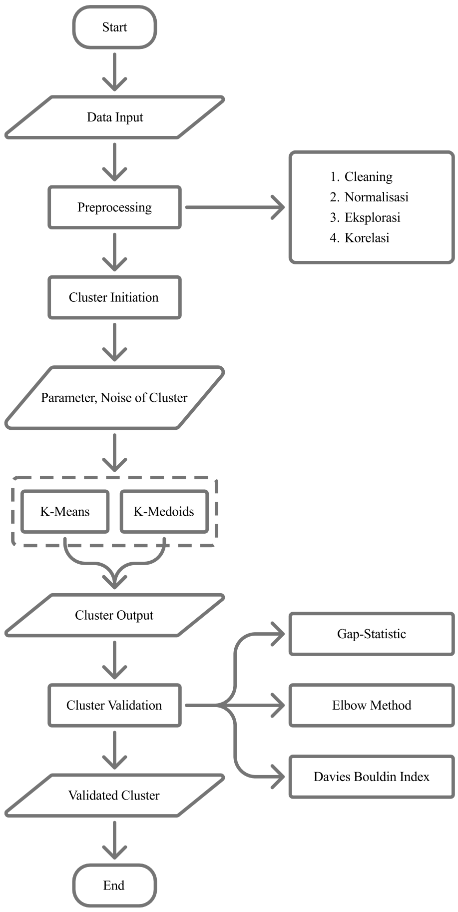


Figure 1. Research Design

To analyze network congestion patterns at the Universitas Muslim Indonesia, this study involved a series of stages. This stage includes various processes, starting with the input of a prepared data set which includes number of packet,

packet loss, throughput, delay, file size, siza data, data byte rate, data bit rate, average packet size, average packet rate. Furthermore, the preprocessing stage is carried out, followed by cluster initiation, determination of parameter values and number of clusters, conducting clustering analysis using the K-means algorithm, validating clustering results using Gap-statistic, elbow method and davies bouldin index, the last stage is labeling data records. A visual representation of the stages of research is depicted in **Figure 1**.

F. K-Means Algorithm

The K-Means algorithm is one of the clustering algorithms that is included in the Unsupervised Learning group which is used to group data into several groups with a partition system [22]. Unsupervised learning is a data mining algorithm to look for patterns from a variable (attribute), the variable (attribute) that is the target/ label/ class is not specified (none) [23]. The K-Means algorithm method seeks to group existing data into several groups where the data contained in one group has the same characteristics as one another but has different characteristics from data in other groups. Thus, this method can be used to minimize variation between data contained in a cluster and maximize variation with data contained in other clusters. The purpose of clustering is to group objects into k clusters. Clustering using the K-Means method is generally done with the following basic algorithm [24].

1. Determining the k number of clusters
2. Initializing the cluster center k randomly. Cluster centers
3. Allocating all object data on the nearest cluster. The similarity of two objects can be determined based on the distance between the two objects. While the similarity of a data in a cluster is determined by the distance between the data and the centroid. At this stage, calculating the distance of each data to the centroid can use Euclidean distance with equation 1.

$$d(x_1, x_{2,}) = \sqrt{\sum_{i=1}^{p} (x_{2i-} x_{1i})^2} \tag{1}$$

Where $d(x_1, x_2)$ is Euclidean distance, x_{2i} is i^{th} test data on the 2nd variable.

x_{1i} : data samples are on the first variable.

p : dimensionality of independent variable data

4. Recalculating the distance between the centroid and the current data. Centroid is the average value of all objects/data in the class. However, in the next centroid determination, the median can be used if possible.
5. Repeat steps 2-4 until the centroid value does not change.

G. K-Medoids Algorithm

The K-Medoids algorithm, a data mining clustering technique, is employed for cluster formation [25]. This algorithm forms clusters by calculating the shortest distance between data points. Initially, K-Medoids determines the central point for each cluster, and subsequent cluster formation relies on the proximity of data points to these central points, grouping them based on similarity or shared characteristics [26].

The distance matrix relative to the medoids is calculated using the Euclidean Distance formula (2), which involves computing the distance of each data point to the two predetermined medoids.

$$D(i, C_j) = \sqrt{\sum_{k=1}^{n} (X_i, k \ \ -C_{j,k})^2 \ )} \tag{2}$$

The cluster centres are recomputed according to the Formula (3) :

$$C_j = \frac{\sum_{i=1}^{n} u_{ij}^m . x_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{3}$$

H. Cluster Evaluation

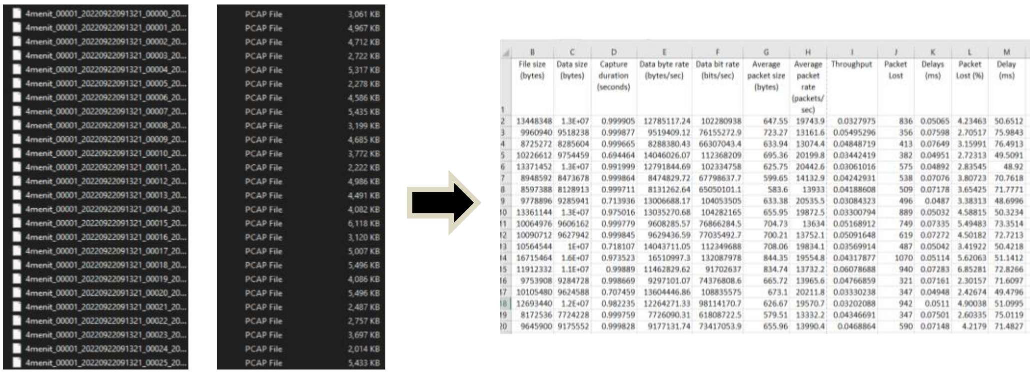
Cluster evaluation is the process of assessing the quality and suitability of sharing data into groups (clusters), with the aim of measuring the extent to which the cluster represents the actual structure in the data, as follows:

1. Gap-Statistic is a cluster evaluation method that compares cluster metrics in actual data with metrics in random data to determine the optimal number of clusters in a dataset, helping to overcome the challenges of determining cluster counts without relying on prior knowledge [27].
2. Elbow Method is a clustering analysis approach used to select the optimal number of clusters in a dataset [28]. The selection is done by identifying the most significant angle or drop between the first and second cluster values in the graph, where the number of clusters at that point is considered the optimal because it provides maximum representation of the structure in the data [21].
3. Davies Bouldin Index is a cluster evaluation method used to evaluate clusters in general based on quantity and proximity between cluster members [29]. This index shows a positive correlation for "in-class" situations and a negative correlation for "inter-class" situations. The calculation of the Davies Bouldin Index value is based on the ratio between the i^{th} cluster and the j^{th} cluster. The smaller the Davies Bouldin Index value, the better the resulting cluster [30].

Results and Discussion

A. Data Collection Results

This step is done using Wireshark with PCAP (Packet Capture), an Application Programming Interface (API) is used to capture network traffic, which is then stored in a packet capture file. Data transformation is done from PCAP files into excel files that are extracted by coding using python. Where there are Number of Packet, File size (bytes), Data size (bytes), Capture duration (seconds), Data byte rate (bytes / sec), Data bit rate (bits/sec), Average packet size (bytes), Average packet rate (packets / sec), Throughput, Packet Loss, Delays (mc), Packet Loss (%), Delay (ms).

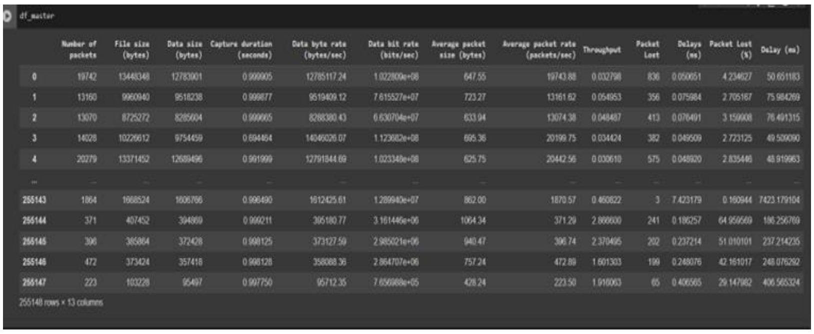


File size (bytes)	Data size (bytes)	Capture duration (seconds)	Data byte rate (bytes/sec)	Data bit rate (bits/sec)	Average packet size (bytes)	Average packet rate (packets/sec)	Throughput	Packet Lost	Delays (ms)	Packet Lost (%)	Delay (ms)
13448348	1.3E+07	0.999905	12785117.24	102280938	647.55	19743.9	0.0327975	836	0.05065	4.23463	50.6512
9960940	9518238	0.999877	9519409.12	76155272.9	723.27	13161.6	0.05495296	356	0.07598	2.70517	75.9843
8725272	8285604	0.999665	8288380.43	66307043.4	633.94	13074.4	0.04848719	413	0.07649	3.15991	76.4913
10226612	9754459	0.694464	14046026.07	112368209	695.36	20199.8	0.03442419	382	0.04951	2.72313	49.5091
13371452	1.3E+07	0.991999	12791844.69	102334758	625.75	20442.6	0.03061016	575	0.04892	2.83545	48.92
8948592	8473678	0.999864	8474829.72	67798637.7	599.65	14132.9	0.04242931	538	0.07076	3.80723	70.7618
8597388	8128913	0.999711	8131262.64	65050101.1	583.6	13933	0.04188608	509	0.07178	3.65425	71.7771
9778896	9285941	0.713936	13006688.17	104053505	633.38	20535.5	0.03084323	496	0.0487	3.38313	48.6996
13361144	1.3E+07	0.975016	13035270.68	104282165	655.95	19872.5	0.03300794	889	0.05032	4.58815	50.3234
10064976	9606162	0.999779	9608285.57	76866284.5	704.73	13634	0.05168912	749	0.07335	5.49483	73.3514
10090712	9627942	0.999845	9629436.59	77035492.7	700.21	13752.1	0.05091648	619	0.07272	4.50182	72.7213
10564544	1E+07	0.718107	14043711.05	112349688	708.06	19834.1	0.03569914	487	0.05042	3.41922	50.4218
16715464	1.6E+07	0.973523	16510997.3	132087978	844.35	19554.8	0.04317877	1070	0.05114	5.62063	51.1412
11912332	1.1E+07	0.99889	11462829.62	91702637	834.74	13732.2	0.06078688	940	0.07283	6.85281	72.8266
9753908	9284728	0.998669	9297101.07	74376808.6	665.72	13965.6	0.04766859	321	0.07161	2.30157	71.6097
10105480	9624588	0.707459	13604446.86	108835575	673.1	20211.8	0.03330238	347	0.04948	2.42674	49.4796
12693440	1.2E+07	0.982235	12264271.33	98114170.7	626.67	19570.7	0.03202088	942	0.0511	4.90038	51.0995
8172536	7724228	0.999759	7726090.31	61808722.5	579.51	13332.2	0.04346691	347	0.07501	2.60335	75.0119
9645900	9175552	0.999828	9177131.74	73417053.9	655.96	13990.4	0.0468864	590	0.07148	4.2179	71.4827

Figure 2. Data Transformation Wireshark to Excel

B. Data Preparation

Before conducting K-Means analysis, data preparation is first carried out which is to clean, compile, and format the data so that it can be analyzed more effectively as shown in **Figure 3**.



df_master

	Number of packets	File size (bytes)	Data size (bytes)	Capture duration (seconds)	Data byte rate (bytes/sec)	Data bit rate (bits/sec)	Average packet size (bytes)	Average packet rate (packets/sec)	Throughput	Packet Lost	Delays (ms)	Packet Lost (%)	Delay (ms)
0	19742	13448348	12783001	0.999905	12785117.24	1.022809e+08	647.55	19743.88	0.032798	836	0.050651	4.234627	50.651183
1	13160	9960940	9518238	0.999877	9519409.12	7.615527e+07	723.27	13161.62	0.054953	356	0.075984	2.705167	75.984269
2	13070	8725272	8285604	0.999665	8288380.43	6.630704e+07	633.94	13074.38	0.048487	413	0.076491	3.159908	76.491315
3	14028	10226612	9754459	0.694464	14046026.07	1.123682e+08	695.36	20199.75	0.034424	382	0.049509	2.723125	49.509090
4	20279	13371452	12689496	0.991999	12791844.69	1.023348e+08	625.75	20442.56	0.030610	575	0.048920	2.835446	48.919963
...	...	...	...	...	...	...	...	...	...	...	...	...	...
255143	1864	1668524	1606766	0.996490	1612425.61	1.289940e+07	862.00	1870.57	0.460822	3	7.423179	0.160944	7423.179104
255144	371	407452	394869	0.999211	395180.77	3.161446e+06	1064.34	371.29	2.866600	241	0.186257	64.959569	186.256769
255145	396	385864	372428	0.998125	373127.59	2.985021e+06	940.47	396.74	2.370495	202	0.237214	51.010101	237.214235
255146	472	373424	357418	0.998128	358088.36	2.864707e+06	757.24	472.89	1.601303	199	0.248076	42.161017	248.076292
255147	223	103228	95497	0.997750	95712.35	7.656988e+05	428.24	223.50	1.916063	65	0.406565	29.147982	406.565324

255148 rows × 13 columns

Figure 3. Data Preparation

C. *Exploratory Data Analysis*

Table 2. Descriptive Statistic

Measure	Number Of Packet	Throughput	Packet Loss (%)	Delay (ms)
count	255148.000000	255148.000000	255148.000000	255148.000000
mean	12575.172523	0.066869	11.081240	179.101606
std	4164.102445	0.528798	9.169622	3299.737035
min	11.000000	0.006252	0.000000	0.000000
25%	10561.750000	0.042339	3.752086	57.320469
50%	13327.000000	0.053029	8.623442	71.979518
75%	15266.000000	0.063652	16.772250	83.946977
max	50352.000000	113.936625	100.000000	662133.000000

In **Figure 4**, the correlation between variables is also calculated which is then used as a fundamental consideration for the right analysis at the following stage. High correlation between variables result in multicollinearity which has an impact on the efficiency of the model which is unstable and difficult to determine so that it requires long computations. In addition, multicollinearity will also result in the risk of overfitting and inaccurate testing. **Figure 4** shows the correlation value between variables where some variables have a very high correlation value (close) to other variables which shows that the selection and simplification of variables is needed so that overfitting does not occur.

The first step of the exploratory data analysis (EDA) process involves descriptive statistical calculations calculating the mean, standard of division, minimum, maximum, and quartile values on each variable to understand the overall distribution of data. The calculation results from the EDA stage have been obtained and cover several aspects as in **Table 2**.

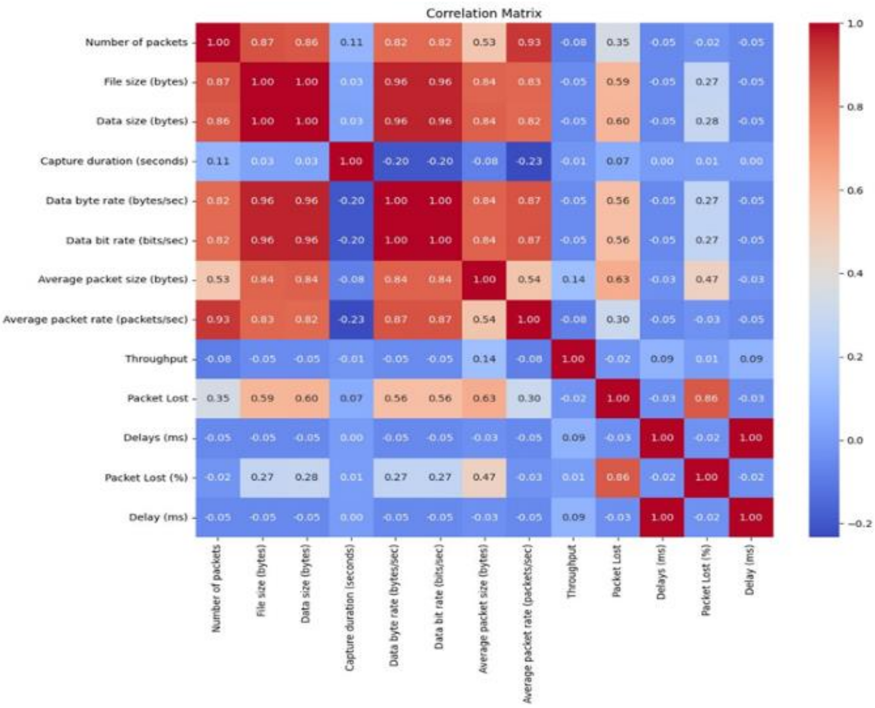


Figure 4. Correlation between variables

Figure 5 shows the distribution or distribution of data for each variabel used

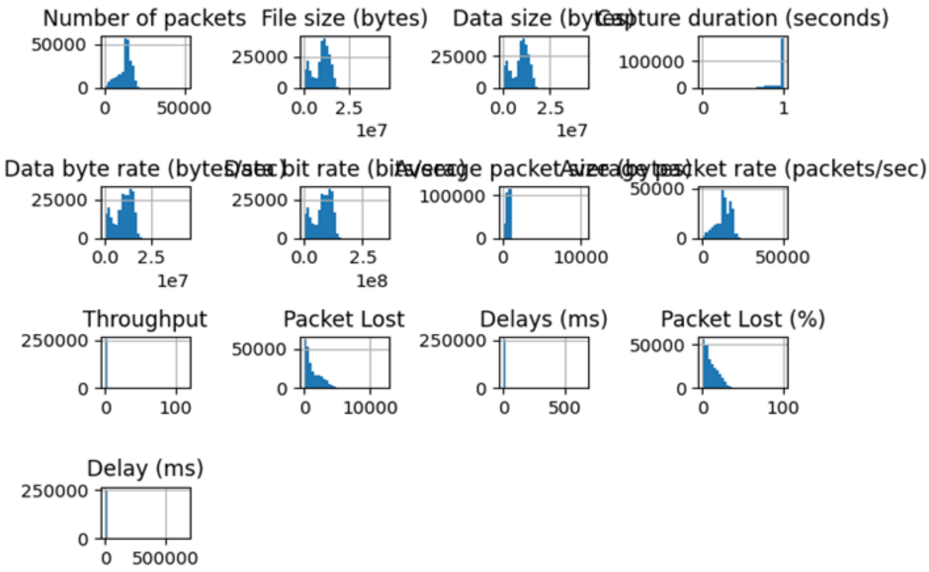
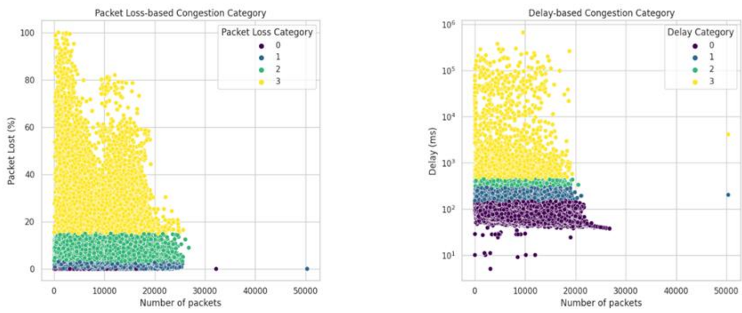


Figure 5. Exploratory Data Analysis

Figure 6 Shows the grouping of congestion types based on the TIPHON Standard. Where scatter plots are used displays the relationship between the two variables "Number of Packets" and "Packet loss (%)". As well as "Delay" and "Number of Packet".



(a) (b)

Figure 6. Congestion Level 4 categories based on TIPHON standards (a) Packet loss, (b) Delay

D. K-Means Analysis

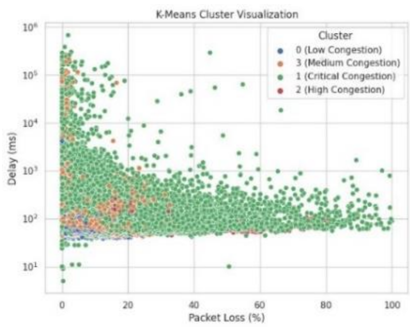
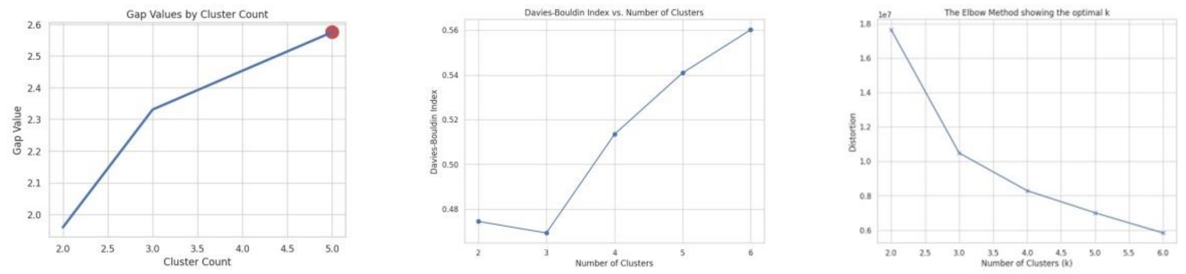


Figure 7. Congestion Level 10 variables with 4 clusters using K-Means based on Packet-Loss & Delay

At this stage, Capture duration, Average packet rate, and Delays were dropped as variables. The analysis was then conducted using the remaining 10 variables: Number of Packets, File size, Data size, Data byte rate, Data bit rate, Average packet size, Throughput, Packet Loss, and delay. These variables were analyzed within four proposed clusters: Low, Medium, High, and Critical. However, as illustrated in **Figure 7**, the clustering pattern was not discernible when combining Packet Loss and Delay.



(a) (b) (c)

Figure 8. Validate the number of clusters with 10 input variables using (a) Gap-Statistic, (b) Davies-Bouldin Index and (c) Elbow Method

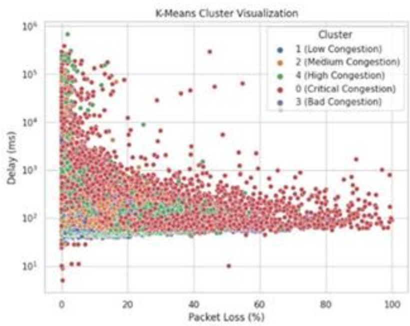


Figure 9. Congestion Level 10 variables with 5 clusters using K-Means based on Packet Loss & delay

The evaluation results, utilizing the Gap Statistic, Davies-Bouldin Index, and Elbow Method, suggest that with the application of 10 variables, the optimal number of clusters is 5, as recommended by the Gap Statistic evaluation, and supported by the Davies-Bouldin Index and Elbow Method, as illustrated in **Figure 8**.

Following experimentation with 5 clusters, the results did not reveal a consistent pattern in the combination of packet-loss and delay, as illustrated in **Figure 9**.

In the subsequent experiment, six variables were examined, categorized into three distinct clusters.

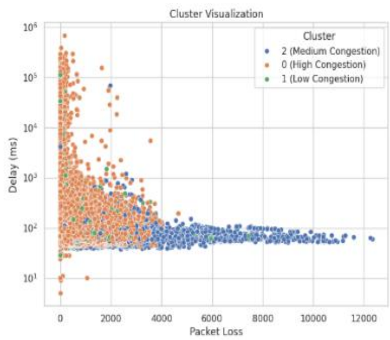
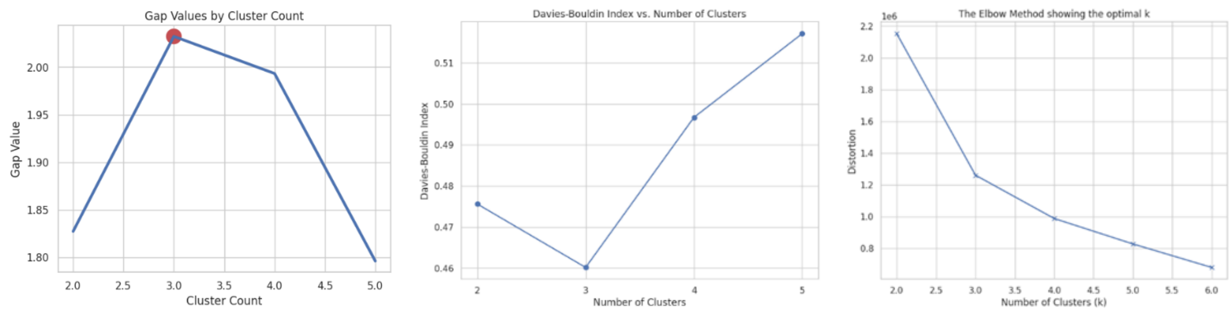


Figure 10. Congestion Level 6 variable with 3 clusters using K-Means based on Packet Loss & Delay

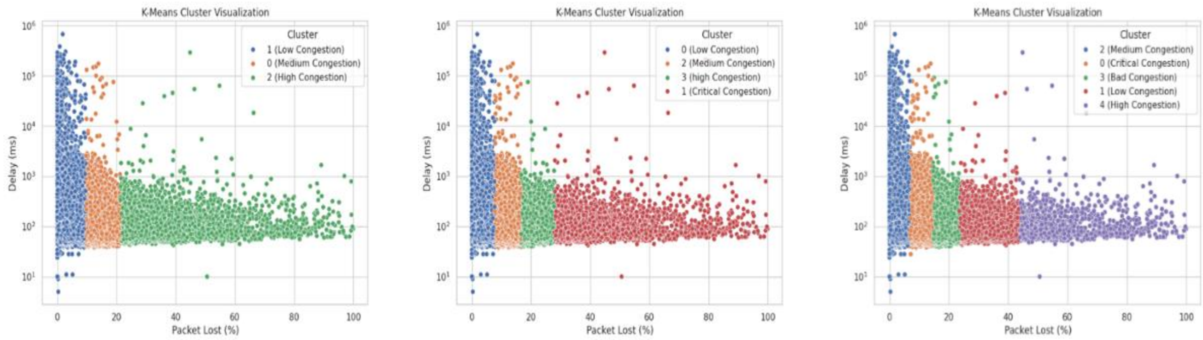
As depicted in **Figure 10**, the resultant output maintains an inconsistent configuration, with the visualization highlighting anomalies within the structure. Consequently, the cluster evaluation phase was reiterated utilizing six variables—the Gap Statistic, Davies-Bouldin Index, and Elbow Method—collectively proposing a three-cluster arrangement, as demonstrated in **Figure 11**.



(a) (b) (c)

Figure 11. Validation the number of clusters with 6 input variables using (a) Gap-Statistic, (b) Davies-Bouldin Index and (c) Elbow Method

The evaluation of clusters using Gap Statistic, Davies-Bouldin Index, and Elbow Method led to tests employing three clusters with three QoS variables: throughput, packet loss, and delay. In this testing phase, the combined patterns of packet loss and delay were clearly observable. Subsequently, tests were conducted using four and five clusters with the same three QoS variables, yielding results that were clearly and systematically observable, as depicted in **Figure 12**.



(a) (b) (c)

Figure 12. Congestion Level 3 variable with 3, 4 and 5 clusters using K-Means based on (a) Packet-Loss & Delay 3 clusters, (b) Packet-Loss & Delay 4 clusters and (c) Packet-Loss & Delay 5 clusters

The validation of K-Means modeling was conducted using the Gap-Statistic, Davies-Bouldin Index, and Elbow method. The results indicate that utilizing three Quality of Service variables throughput, packet loss, and delay, allows for the formation of 3, 4, and 5 clusters that are both valid and reliable.

E. K-Medoids Analysis

To provide a more comprehensive comparison, a similar approach was conducted using the K-Medoid algorithm. Initial testing also used four clusters, as proposed by the research, as shown in **Figure 13**. In the visualization of clustering results based on the combination of Packet Loss and Delay variables, the separation pattern between clusters was not clearly visible. Therefore, the evaluation of the optimal number of clusters was performed using the Gap Statistic, Davies-Bouldin Index, and Elbow Method. Based on the evaluation results, using 10 variables, the recommended number of clusters is five. Although the Gap Statistic showed the highest value at six clusters, both the Davies-Bouldin Index and the Elbow Method indicated that five clusters were a more optimal choice, as shown in **Figure 14**.

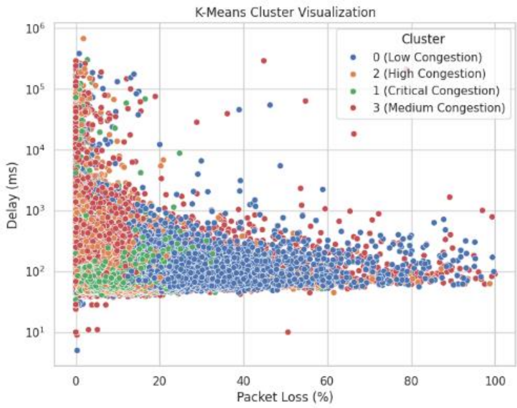
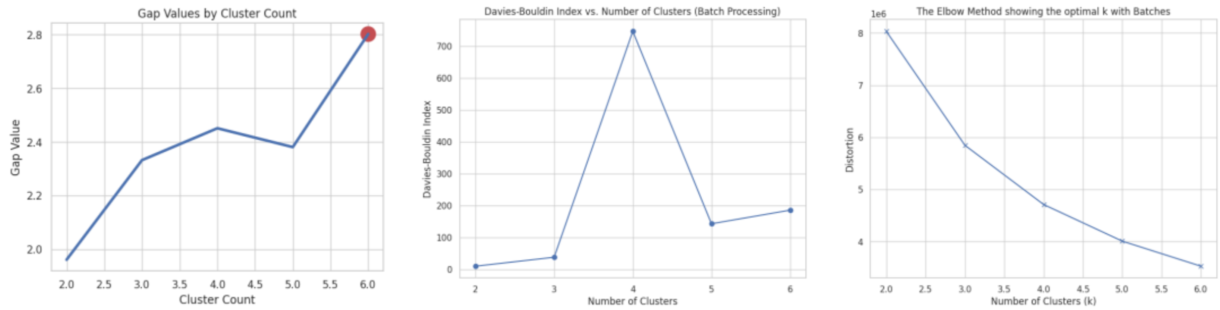


Figure 13. Congestion Level 10 variables with 4 clusters using K-Medoids based on Packet-Loss & Delay



(a) (b) (c)

Figure 14. Validation the number of clusters with 10 input variables using (a) Gap-Statistic, (b) Davies-Bouldin Index and (c) Elbow Method

Following a re-evaluation using five clusters, the resulting visualization is depicted in **Figure 15**. Despite adjusting the number of clusters based on prior evaluations, a clear and organized pattern of separation between clusters in the Packet Loss and Delay variable combination remains elusive. This suggests that, notwithstanding the adjustment in cluster quantity, the combination of these two variables is insufficient for producing a visually informative cluster separation.

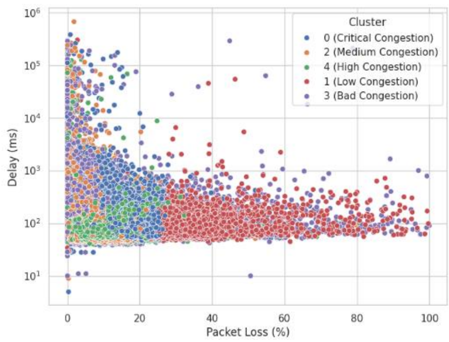


Figure 15. Congestion Level 10 variables with 5 clusters using K-Medoids based on Packet Loss & delay

In subsequent testing, six variables were considered during the clustering process. Based on the validation results obtained using the Gap Statistic, Davies-Bouldin Index, and Elbow Method, it was determined that the optimal number of clusters is three, as illustrated in **Figure 16**.

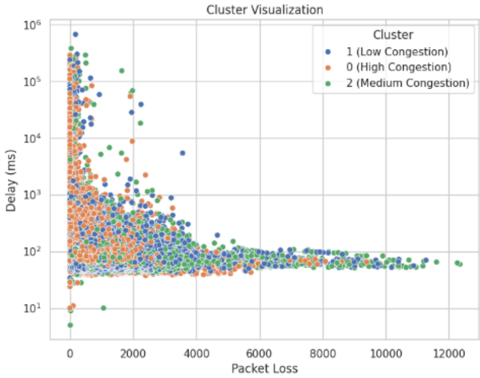
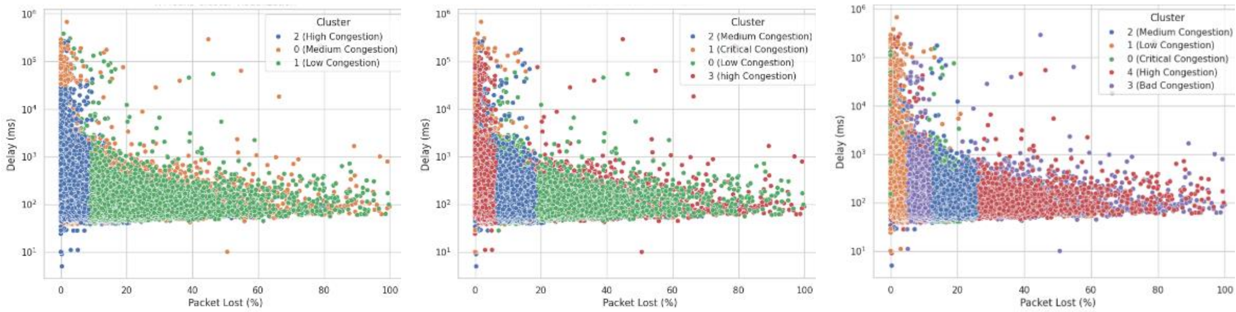


Figure 16. Congestion Level 6 variabel dengan 3 cluster menggunakan K-Medoid berdasarkan Packet-Loss & Delay

The evaluation also incorporated three QoS variables: throughput, packet loss, and delay. The clustering process yielded three distinct clusters, as depicted in **Figure 17**. Visual analysis reveals that the separation patterns between clusters, based on the combination of packet loss and delay variables, appear more distinct and organized compared to previous experiments.



(a) (b) (c)

Figure 17. Congestion Level 3 variable with 3, 4 and 5 clusters using K-Medoid based on (a) Packet-Loss & Delay 3 clusters, (b) Packet-Loss & Delay 4 clusters and (c) Packet-Loss & Delay 5 clusters

The To assess the consistency and potential for more detailed segmentation, further testing was conducted using 4 and 5 clusters, as illustrated in Figure 17. Despite the increased number of clusters, the visualization of the clustering results indicates that the separation between clusters remains orderly and consistent, which reinforces the reliability of the K-Medoid method in grouping the data.

In summary, validation results from the Gap Statistic, Davies-Bouldin Index, and Elbow Method that employing these three QoS variables allows the formation of 3 to 5 clusters via the K-Medoid method. This approach yields valid, reliable, and representative data segmentation, effectively depicting network congestion levels in a more structured manner. Nevertheless, in comparison to K-Medoid, the K-Means method demonstrates more stable cluster patterns, clearer separation between clusters, and more structured and informative data visualization.

Conclusion

Based on the research conducted using the K-Means and K-Medoid methods, it can be concluded that the clustering process for network congestion levels was successfully implemented. The process effectively grouped data through various tests on the number of variables and clusters. The modeling of both algorithms was validated using Gap Statistic, Davies-Bouldin Index, and Elbow Method. The results showed that using three QoS variables—throughput, packet loss, and delay—both methods were able to form valid and reliable clusters. However, K-Means demonstrated superior performance in forming stable clusters and more structured patterns in the visualization compared to K-Medoid.

References

[1] D. Sunitha, K. R. Balmuri, R. P. de Prado, P. B. Divakarachari, R. Vijayarangan, and K. L. Hemalatha, "Congestion centric multi-objective reptile search algorithm-based clustering and routing in cognitive radio

sensor network," *Trans. Emerg. Telecommun. Technol.*, vol. 34, no. 11, p. e4629, 2023, doi: https://doi.org/10.1002/ett.4629.

[2] S. Bauer, D. Clark, and W. Lehr, "The Evolution of Internet Congestion," *37th Res. Conf. Commun. Inf. Internet Policy*, vol. 64, no. August 2009, pp. 1–34, 2009, doi: 10.1016/j.addbeh.2015.05.016

[3] C. Ma, J. P. Sheu, and C. X. Hsu, "A game theory based congestion control protocol for wireless personal area networks," *J. Sensors*, vol. 2016, 2016, doi: 10.1155/2016/6168535.

[4] M. Pundir and J. K. Sandhu, "A Systematic Review of Quality of Service in Wireless Sensor Networks using Machine Learning: Recent Trend and Future Vision," *J. Netw. Comput. Appl.*, vol. 188, no. April, p. 103084, 2021, doi: 10.1016/j.jnca.2021.103084.

[5] C. Sergiou, P. Antoniou, and V. Vassiliou, "A comprehensive survey of congestion control protocols in wireless sensor networks," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 4, pp. 1839–1859, 2014, doi: 10.1109/COMST.2014.2320071.

[6] M. Gholipour *et al.*, "Author ' s Accepted Manuscript Hop-by-Hop Congestion Avoidance in Wireless Machine Reference : To appear in : Neurocomputing Hop-by-Hop Congestion Avoidance in Wireless Sensor Networks Based on Genetic Support Vector Machine," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 4, pp. 1839–1859, 2016, doi: 10.1109/COMST.2014.2320071.

[7] K. Wang, C. Ma, Y. Qiao, X. Lu, W. Hao, and S. Dong, "A hybrid deep learning model with 1DCNN-LSTM-Attention networks for short-term traffic flow prediction," *Phys. A Stat. Mech. its Appl.*, vol. 583, p. 126293, 2021, doi: https://doi.org/10.1016/j.physa.2021.126293.

[8] M. K. Kodmelwar, R. Kumar, J. Singh, R. Kumar, T. Ramachandran, and G. Ramya, "An Advanced Encryption Protocols for Wireless Network Security in Web Applications," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–5. doi: 10.1109/ICCCNT61001.2024.10725848.

[9] J. Y. Colin, H. Mathieu, and M. Nakechbandi, "A proposal for an open logistics interconnection reference model for a Physical Internet," *Proc. 3rd IEEE Int. Conf. Logist. Oper. Manag. GOL 2016*, 2016, doi: 10.1109/GOL.2016.7731719.

[10] A. Piotrowska, "On cross-layer interactions for congestion control in the internet," *Appl. Sci.*, vol. 11, no. 17, 2021, doi: 10.3390/app11177808.

[11] A. A. Rezaee and F. Pasandideh, "A Fuzzy Congestion Control Protocol Based on Active Queue Management in Wireless Sensor Networks with Medical Applications," *Wirel. Pers. Commun.*, vol. 98, no. 1, pp. 815–842, 2018, doi: 10.1007/s11277-017-4896-6.

[12] J. Lorincz, Z. Klarin, and J. Ožegović, "A comprehensive overview of tcp congestion control in 5g networks: Research challenges and future perspectives," *Sensors*, vol. 21, no. 13, 2021, doi: 10.3390/s21134510.

[13] B. Turkovic, F. A. Kuipers, and S. Uhlig, "Fifty Shades of Congestion Control: A Performance and Interactions Evaluation," *arXiv:1903.03852*, 2019, doi: https://doi.org/10.3390/s21134510.

[14] B. Zhou, D. He, and Z. Sun, "Traffic Modeling and Prediction using ARIMA/GARCH Model," in *Modeling and Simulation Tools for Emerging Telecommunication Networks*, A. Nejat Ince and E. Topuz, Eds., Boston, MA: Springer US, 2006, pp. 101–121. doi: https://doi.org/10.1007/0-387-34167-6_5.

[15] E. I. Alwi and L. B. Ilmawan, "Analisis Kinerja QoS (Quality of Service) Jaringan WLAN Ukhuwahnet Pada Universitas Muslim Indonesia," *Semin. Nas.*, pp. 55–60, 2019, doi: https://doi.org/10.29407/inotek.v3i1.513.

[16] P. Purnawansyah, "Congestion Predictive Modelling on Network Dataset Using Ensemble Deep Learning," *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1597–1613, Dec. 2024, doi: 10.47738/jads.v5i4.333.

[17] I. Nurhaida and I. Ichsan, "Congestion Control Pada Jaringan Komputer Berbasis Multi Protocol Label Switching (Mpls)," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 11, no. 1, pp. 77–88, 2020, doi: 10.24176/simet.v11i1.3671.

[18] X. Zhou, W. Liang, K. I.-K. Wang, and L. T. Yang, "Deep Correlation Mining Based on Hierarchical Hybrid Networks for Heterogeneous Big Data Recommendations," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 171–178, 2021, doi: 10.1109/TCSS.2020.2987846.

[19] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Inf. Process. Manag.*, vol. 57, no. 2, p. 102034, 2020, doi: https://doi.org/10.1016/j.ipm.2019.04.002.

[20] W. Balzano, A. Murano, L. Sorrentino, and S. Stranieri, "Behavioral Clustering: A New Approach for Traffic Congestion Evaluation," L. Barolli, F. Amato, F. Moscato, T. Enokido, and M. Takizawa, Eds., Cham: Springer International Publishing, 2020, pp. 1418–1427. doi: 10.1007/978-3-030-44038-1_129.

[21] A. Et-Taleby, M. Boussetta, and M. Benslimane, "Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image," *Int. J. Photoenergy*, vol. 2020, 2020, doi: 10.1155/2020/6617597.

[22] K. Kandali, L. Bennis, and H. Bennis, "A New Hybrid Routing Protocol Using a Modified K-Means Clustering Algorithm and Continuous Hopfield Network for VANET," *IEEE Access*, vol. 9, pp. 47169–47183, 2021, doi: 10.1109/ACCESS.2021.3068074.

[23] Y. Sinambela, S. Herman, A. Takwim, and S. R. Widianto, "a Study of Comparing Conceptual and Performance of K-Means and Fuzzy C Means Algorithms (Clustering Method of Data Mining) of Consumer Segmentation," *J. Ris. Inform.*, vol. 2, no. 2, pp. 49–54, 2020, doi: 10.34288/jri.v2i2.116.

[24] T. Widiyaningtyas, M. I. W. Prabowo, and M. A. M. Pratama, "Implementation of K-means clustering method to distribution of high school teachers," *2017 4th Int. Conf. Electr. Eng. Comput. Sci. Informatics*, no. September, pp. 1–6, 2017, doi: 10.1109/EECSI.2017.8239083.

[25] A. Sobrinho Campolina Martins, L. Ramos de Araujo, and D. Rosana Ribeiro Penido, "K-Medoids clustering applications for high-dimensionality multiphase probabilistic power flow," *Int. J. Electr. Power Energy Syst.*, vol. 157, no. February, 2024, doi: 10.1016/j.ijepes.2024.109861.

[26] S. Sen, L. Sahoo, K. Tiwary, V. Simic, and T. Senapati, "Wireless Sensor Network Lifetime Extension via K-Medoids and MCDM Techniques in Uncertain Environment," *Appl. Sci.*, vol. 13, no. 5, 2023, doi: 10.3390/app13053196.

[27] D. Thi Thu Nguyen, L. V Chernenkaya, and H. Van Le, "A Model for Classification Based on the K-Means Method Using the Gap-Statistic Algorithm," J. C. Bansal, H. Sharma, and A. Chakravorty, Eds., Singapore: Springer Nature Singapore, 2023, pp. 351–363. doi: 10.1007/978-981-99-2468-4_27.

[28] I. D. Sudirman, I. Rahmatillah, and O. O. Sharif, "An Examination of Plastic Bag Consumption by Age, Gender, and Spending Using K-Means and Elbow Method," in *2023 International Conference on Informatics Engineering, Science & Technology (INCITEST)*, 2023, pp. 1–6. doi: 10.1109/INCITEST59455.2023.10396903.

[29] T. T. D. Nguyen, L. T. T. Nguyen, Q.-T. Bui, U. Yun, and B. Vo, "An efficient topological-based clustering method on spatial data in network space," *Expert Syst. Appl.*, vol. 215, p. 119395, 2023, doi: https://doi.org/10.1016/j.eswa.2022.119395.

[30] J. Manullang, P. Sirait, and Andri, "Aplication of Validity Index in K Means and Fuzzy C Means," *J. Mantik*, vol. 4, no. 2, pp. 1430–1438, 2020, doi: https://doi.org/10.35335/mantik.Vol4.2020.958.pp1430-1438