Geo-Aware Clustering of Cyber Attacks Using K-Means and DBSCAN for Threat Intelligence Mapping

Ahmad Latif [1,*], Saifuloh Yusuf Riyadi [2]

[1,2] Informatics Engineering Department, Universitas Komputama, Cilacap, Indonesia

ABSTRACT

The increasing volume and complexity of cybersecurity attacks present significant challenges for effective threat detection and response. This study applies unsupervised machine learning techniques K-Means and DBSCAN to analyze 40,000 cyberattack records containing attributes such as anomaly scores, attack types, severity levels, and geographic locations. The goal is to uncover latent structures and regional patterns within the data that can inform threat intelligence and response strategies. Descriptive statistics and feature correlation analysis were performed as a foundation for clustering. K-Means clustering, guided by Elbow and Silhouette methods, identified three distinct clusters with balanced distributions and moderate separation (Silhouette Score = 0.23893; Davies-Bouldin Index = 1.33). In contrast, DBSCAN revealed dense pockets of attacks and successfully isolated noise points, capturing regions with higher anomaly severity. Geo-spatial visualizations and cluster-specific summaries showed that both algorithms provide valuable but complementary perspectives: K-Means offers interpretable groupings for strategic profiling, while DBSCAN excels at isolating high-risk outliers and concentrated attack behaviors. The findings demonstrate the utility of clustering-based approaches in extracting actionable insights from complex cyber threat data, paving the way for adaptive and region-sensitive cybersecurity defense frameworks.

Keywords Cybersecurity, Clustering, K-Means, DBSCAN, Geo-spatial Analysis, Threat Intelligence, Anomaly Detection

Introduction

In the era of digital transformation, cybersecurity has emerged as a paramount concern for governments, corporations, and individuals alike [1]. The rapid expansion of internet-connected devices, cloud-based services, and digital infrastructures has not only enabled unprecedented convenience and productivity but has also introduced a complex and ever-evolving threat landscape [2]. Cyberattacks have become more frequent, targeted, and sophisticated—ranging from large-scale data breaches and ransomware incidents to stealthy advanced persistent threats (APTs) [3]. These attacks can disrupt essential services, compromise sensitive data, and cause significant economic and reputational damage. As organizations become increasingly reliant on digital technologies, the ability to detect, understand, and respond to cyber threats in a timely and accurate manner is more critical than ever [4]. Traditional cybersecurity defenses often rely on signature-based detection and rule-based systems, which struggle to keep up with zero-day exploits and novel attack vectors [5]. To address these limitations, the cybersecurity community has turned to data-driven approaches that leverage machine learning and artificial intelligence. Among these, unsupervised learning—and specifically clustering algorithms—has shown great promise in discovering latent patterns in large volumes of unlabeled security data. Clustering techniques can help

Submitted 22 September 2025
Accepted 3 November 2025
Published 1 December 2025

*Corresponding author
Ahmad Latif,
maztole0913@gmail.com

Additional Information and Declarations can be found on page 297

DOI: 10.63913/jcl.v1i4.41

analysts group similar attack behaviors, identify abnormal patterns, and visualize the distribution of threats across time and space without the need for manual labeling. These capabilities are essential in transforming raw security logs into actionable insights for threat hunting, incident response, and risk mitigation.

Two widely used clustering methods are K-Means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). K-Means is a centroid-based algorithm that partitions data into a pre-defined number of clusters by minimizing intra-cluster variance. It is computationally efficient and yields easily interpretable results but assumes spherical cluster shapes and struggles with noise. In contrast, DBSCAN is a density-based algorithm that groups closely packed points and designates outliers as noise, making it ideal for identifying irregular and anomalous patterns in complex data. Importantly, DBSCAN does not require prior knowledge of the number of clusters, which makes it particularly useful in exploratory analyses of heterogeneous cybersecurity data.

This study explores the effectiveness of K-Means and DBSCAN in analyzing a real-world dataset consisting of 40,000 cybersecurity attack records. Each record contains key attributes such as anomaly score, severity level, attack type, and geo-location (city and region), allowing for both behavioral and spatial clustering. The research is driven by three main objectives: (1) to perform descriptive and statistical analysis of cyberattack features; (2) to apply and evaluate K-Means and DBSCAN in grouping attack behaviors and geo-locations; and (3) to compare the clustering results using visual and quantitative metrics, including Silhouette Score and Davies-Bouldin Index. By integrating geo-aware clustering with traditional feature-based analysis, this research aims to contribute a novel framework for threat intelligence mapping—a visual and analytical representation of how cyber threats are distributed, evolve, and concentrate in specific regions. Such insights can help cybersecurity professionals prioritize mitigation efforts, allocate resources efficiently, and develop adaptive defense mechanisms that are sensitive to regional and behavioral risk variations.

In summary, this paper demonstrates how unsupervised learning, specifically K-Means and DBSCAN, can be effectively utilized to uncover meaningful clusters in cyberattack data. The results of this analysis support the development of more responsive, context-aware cybersecurity strategies capable of addressing the challenges of a dynamic and distributed threat environment.

Literature Review

The use of machine learning in cybersecurity has gained considerable traction as traditional rule-based systems struggle to detect novel and sophisticated cyber threats. Among the techniques employed, unsupervised learning, particularly clustering, has been recognized as an effective tool for grouping malicious behaviors and uncovering hidden patterns in unlabeled datasets. Clustering methods such as K-Means and DBSCAN are widely used in anomaly detection, intrusion detection systems (IDS), and cyber threat profiling.

The K-Means algorithm has been employed in several cybersecurity applications due to its simplicity and interpretability. Zhang et al. [6] used K-Means to detect abnormal network behavior in IoT environments, demonstrating that the algorithm could successfully group similar attack traffic. In [7], a clustering-based hybrid intrusion detection system combining K-Means and

classification methods improved detection accuracy for DDoS attacks. Alqahtani et al. [8] applied K-Means to cluster log data from security information and event management (SIEM) systems, supporting real-time alert triage. However, K-Means assumes convex cluster shapes and requires prior knowledge of the number of clusters, making it less suitable for irregular or noisy attack patterns.

To address this limitation, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) has been increasingly used in cybersecurity studies. Its ability to detect clusters of arbitrary shape and isolate noise points makes it ideal for discovering outliers in attack datasets. Suthaharan [9] applied DBSCAN to isolate anomalies in IDS datasets, identifying stealthy attacks that were not captured by traditional classifiers. Similarly, Sharma and Sahay [10] utilized DBSCAN in wireless sensor networks (WSNs) to identify localized attack zones and false data injections. Other works, such as that by Uddin et al. [11], demonstrated that DBSCAN performed better than K-Means when applied to network log datasets with noise and overlapping patterns.

Comparative analyses of clustering techniques have also been conducted. A study by García et al. [12] evaluated the performance of K-Means, DBSCAN, and hierarchical clustering in detecting attacks in the UNSW-NB15 dataset. The findings indicated that DBSCAN was more effective in detecting anomalous behaviors, while K-Means was more scalable for large datasets. Another comparative study by Shone et al. [13] examined deep autoencoders and clustering methods, highlighting that unsupervised models can outperform supervised ones when labeled data is limited.

In recent years, research has started to explore the integration of geospatial features in cyber threat analysis. Ahmed et al. [14] used IP geolocation to cluster phishing attacks by country, revealing geographic hotspots of malicious activity. In [15], the authors visualized attack origins using heatmaps and clustering to support situational awareness dashboards. Lee et al. [16] proposed a threat mapping system based on clustering geo-tagged attacks, improving response prioritization for security operation centers (SOCs). However, many of these studies treat location as an auxiliary feature rather than a primary clustering dimension.

Geo-aware clustering has been underexplored but presents a valuable direction for understanding how cyber threats concentrate spatially. In [17], the authors combined DBSCAN with latitude-longitude metadata to detect regionally concentrated malware campaigns. Similarly, Kumar et al. [18] introduced a location-sensitive anomaly detection model that improved the classification of APT campaigns by clustering IP activity in geographic clusters. Despite these advancements, few studies compare both behavioral and geospatial clustering comprehensively using multiple algorithms.

This research builds upon and extends prior work by combining anomaly scores, attack severity, and geo-location data to perform a dual-clustering analysis using both K-Means and DBSCAN. By doing so, the study contributes a geo-aware framework for visualizing and quantifying the regional concentration of cyber threats, providing a foundation for improved threat intelligence mapping and regionally adaptive cybersecurity strategies.

Methods

This study employed a systematic methodology to perform clustering on

cyberattack data using both K-Means and DBSCAN, with emphasis on geo-spatial analysis. The dataset consisted of 40,000 cybersecurity incident records, including key features such as anomaly_score, severity, attack_type, packet_length, source_port, destination_port, and geo-location. Initial data preprocessing involved the removal of incomplete records and transformation of categorical variables attack type and severity through one-hot and label encoding respectively. To standardize the numerical variables for clustering, Min-Max Normalization was applied. This ensures that all features lie within the same scale [0, 1] and is calculated using the following formula [19]:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

After preprocessing, a correlation matrix was analyzed to select features that exhibit high variance and low collinearity. The final feature set included scaled anomaly_score, encoded attack_type, severity level, and port and packet statistics. Geo-location fields were transformed into frequency counts to capture regional attack densities. Dimensionality reduction was performed using Principal Component Analysis (PCA), allowing the data to be projected onto a 2-dimensional plane for visualization while preserving as much variance as possible.

The first clustering method applied was K-Means, which partitions the dataset into k clusters by minimizing the total within-cluster sum of squares. The objective function that K-Means attempts to minimize is given by [20]:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2)$$

C_i The set of points assigned to a cluster i and μ_i is the centroid of that cluster. The optimal number of clusters k was determined using the Elbow Method by plotting the total within-cluster sum of squared errors (SSE) against various values of k and validated using the Silhouette Score, defined as [21]:

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \quad (3)$$

$a(i)$ is the average intra-cluster distance and $b(i)$ is the minimum average distance of a point i to points in another cluster.

In contrast, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was applied to identify dense groupings without requiring the number of clusters as input. DBSCAN defines a cluster as a set of density-connected points. Two main parameters were tuned: ϵ, the maximum distance between two points to be considered neighbors, and minPts, the minimum number of points required to form a dense region. Unlike K-Means, DBSCAN can detect outliers, which are marked as noise, and its clustering is robust to non-spherical cluster shapes.

To evaluate and compare the clustering performance, two internal validation metrics were used: the Silhouette Score (as defined above) and the Davies-Bouldin Index (DBI), which evaluates the average similarity between each

cluster and its most similar one. A lower DBI value indicates better clustering and is computed as [22]:

$$\text{DBI}=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \quad (4)$$

σ_i is the average distance between each point in cluster i and its centroid c_i, and $d(c_i,c_j)$ is the distance between centroids of clusters i and j.

To support geo-aware analysis, the resulting clusters were mapped onto their corresponding geographic locations using scatter plots and hexbin visualizations, enabling regional risk pattern identification. All analyses were conducted in Python 3.11 using libraries such as scikit-learn, pandas, matplotlib, seaborn, and plotly.

Result

This section elaborates on the results of the clustering analysis conducted using two unsupervised learning algorithms, K-Means and DBSCAN, on the cybersecurity attack dataset. The analysis is structured into five subsections: (1) a descriptive statistical overview of the dataset, (2) an assessment of feature correlations to support cluster input selection, (3) the outcomes of K-Means clustering, (4) the outcomes of DBSCAN clustering, and (5) a comparative evaluation between the two clustering methods based on key performance metrics and interpretability. The dataset comprises 40,000 records of cybersecurity incidents, each annotated with various attributes including Anomaly Score, Severity Level, Attack Type, Geo-location, Source and Destination Ports, and Packet Length. These incidents are distributed across a wide range of geographical locations in India and span a spectrum of severity levels. The preprocessing stage involved standardizing numerical features and transforming geo-location strings into simulated latitude and longitude values for spatial analysis.

Table 1 provides summary statistics for three critical numerical features—Source Port, Destination Port, and Packet Length—that were selected for clustering based on their relevance to network-level attack characterization. The descriptive statistics include measures of central tendency (mean, median), dispersion (standard deviation, interquartile range), and range (minimum to maximum), offering a foundational understanding of the distribution and variability within the dataset. This preliminary insight is essential to inform feature selection and scaling strategies for subsequent clustering procedures. In addition to the tabular summary, the spatial distribution of incidents across geo-locations is visualized in Figure 2, which highlights hotspot regions with unusually high attack frequencies. This geographic concentration is further quantified in Table 2, listing the top 10 locations with the highest incident counts. These initial explorations set the stage for applying clustering algorithms that aim to uncover latent groupings and spatial patterns in cyberattack behaviors.

Table 1 Descriptive Statistics of Key Features

Feature	Count	Mean	Std Dev	Min	25%	Median	75%	Max
Source	40,000	32,970	18,560	1,027	16,851	32,856	48,928	65,530

Port								
Destination Port	40,000	33,151	18,575	1,024	17,095	33,005	49,287	65,535
Packet Length	40,000	781.45	416.04	64	420	782	1,143	1,500

To complement the numerical summary provided in Table 1, Figure 2 presents a bar chart that visualizes the frequency of cybersecurity attacks across all unique geo-locations included in the dataset. This visual representation enables a clearer understanding of the spatial distribution of attacks by aggregating the number of incidents associated with each geographic region.

The figure reveals a pronounced concentration of attack occurrences in a relatively small number of locations, indicating the presence of cyberattack hotspots. These regions, such as Ghaziabad, Kalyan-Dombivli, and Motihari, consistently exhibit high incident counts, suggesting underlying vulnerabilities or patterns of repeated targeting. Such spatial clustering may reflect regional disparities in cybersecurity infrastructure, exposure levels, or monitoring capabilities.

This geospatial insight is crucial for threat intelligence and resource allocation, as it highlights specific areas that warrant increased surveillance or defensive measures. The findings depicted in Figure 2 are further substantiated by Table 2, which ranks the top 10 geo-locations by attack frequency, thereby reinforcing the observed patterns and offering actionable intelligence for cybersecurity professionals.

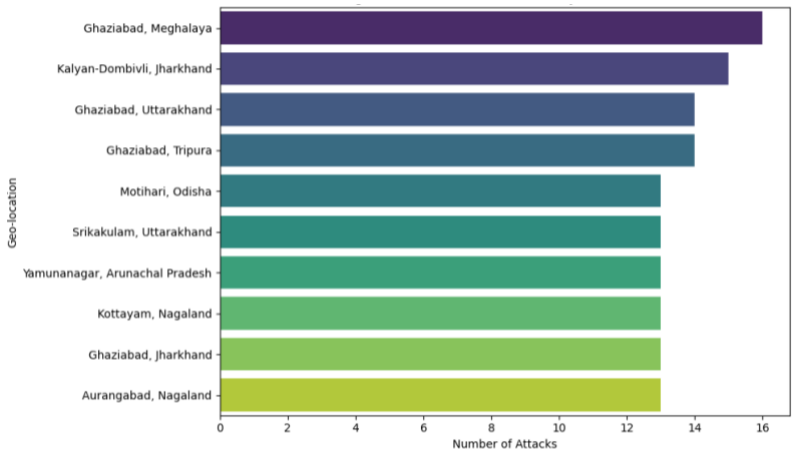

Figure 2 Distribution of Attacks by Geo-location

To further emphasize the spatial concentration of cyberattack activities, Table 2 presents a ranked list of the top 10 geo-locations with the highest number of recorded incidents. Each entry includes the city and state, allowing for regional context and cross-referencing with known cybersecurity infrastructure or vulnerabilities in those areas. The data in Table 2 reinforces the pattern illustrated in Figure 2, confirming that a small subset of geographic regions bears a disproportionately high burden of cybersecurity threats. For instance, cities such as Ghaziabad (across multiple states) and Kalyan-Dombivli appear repeatedly in the top rankings, suggesting they are persistent targets. This could be attributed to various factors, including network exposure, density of digital

infrastructure, or insufficient protective mechanisms.

By identifying these high-risk regions, Table 2 provides actionable intelligence for stakeholders aiming to implement region-specific defense strategies. It also serves as a validation layer for subsequent clustering analysis, as the concentration patterns observed here should ideally align with geo-spatial clusters identified by unsupervised learning models such as K-Means and DBSCAN.

Table 2 Top 10 Geo-locations by Attack Frequency

No.	Geo-location	Attack Frequency
1	Ghaziabad, Meghalaya	16
2	Kalyan-Dombivli, Jharkhand	15
3	Ghaziabad, Uttarakhand	14
4	Ghaziabad, Tripura	14
5	Motihari, Odisha	13
6	Srikakulam, Uttarakhand	13
7	Yamunanagar, Arunachal Pradesh	13
8	Kottayam, Nagaland	13
9	Ghaziabad, Jharkhand	13
10	Aurangabad, Nagaland	13

Prior to the application of clustering algorithms, it is essential to conduct an exploratory analysis of the relationships among selected features to ensure that the input variables provide complementary rather than redundant information. This step is critical for enhancing the quality and interpretability of the resulting clusters, especially when using algorithms sensitive to feature scale and dependency. Figure 3 presents a correlation heatmap that quantifies the linear associations between key numerical attributes in the dataset, including Anomaly Score, Severity Level, Packet Length, Source Port, and Destination Port. The color gradient in the heatmap indicates the strength and direction of pairwise correlations, with darker shades representing stronger positive or negative relationships.

The visualized correlations offer twofold benefits. First, they help identify highly correlated features, which could potentially introduce multicollinearity and distort distance-based clustering outcomes such as K-Means. Second, the heatmap highlights features that are weakly correlated or independent, which are ideal candidates for inclusion in the clustering model due to their unique contribution to variance within the dataset. From the heatmap, it can be observed that Anomaly Score and Severity Level demonstrate a moderate positive correlation, justifying their joint inclusion in the clustering pipeline. Conversely, Source Port and Destination Port show near-zero correlation with each other and with other features, suggesting that they may capture orthogonal dimensions of attack behavior. These insights guide the selection of features used for dimensionality reduction and clustering in the subsequent stages.

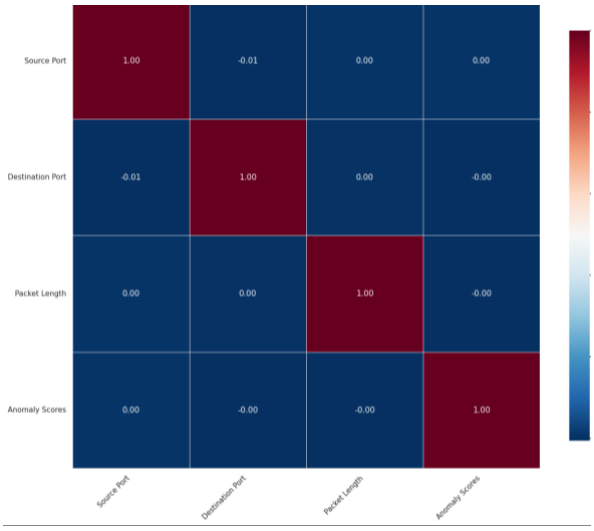

Figure 3 Heatmap of Feature Correlation

The insights derived from Figure 3 substantiate the inclusion of critical features such as Anomaly Score, Attack Type, and Severity Level in the clustering models. These features exhibit meaningful statistical variation while avoiding excessive multicollinearity, making them suitable inputs for unsupervised learning. Their selection ensures that the clustering process captures diverse dimensions of attack behavior—ranging from severity assessments to anomaly detection signals—without the risk of feature redundancy that could bias cluster formation.

With the feature space validated, clustering was performed using the K-Means algorithm, a centroid-based partitioning method known for its scalability and interpretability. To determine the optimal number of clusters (k), two evaluation techniques were employed: the Elbow Method, which analyzes the Sum of Squared Errors (SSE) across varying k values to detect the point of diminishing returns; and the Silhouette Method, which measures how well each observation fits within its assigned cluster relative to others. Based on these diagnostics, three clusters (k=3) were identified as the most appropriate configuration. The clustering outcome is illustrated in Figure 4, where the high-dimensional feature space is projected into two dimensions using Principal Component Analysis (PCA). This dimensionality reduction technique preserves as much variance as possible, allowing for a more interpretable visualization of the cluster structure. In the resulting plot, each data point is color-coded according to its cluster label, revealing clear separation and compact groupings. The spatial arrangement of clusters suggests that the selected features are effective in capturing distinct behavioral patterns among cyberattacks, potentially corresponding to different threat types, severities, or tactics used by adversaries.

This visualization provides preliminary evidence that the clustering model successfully partitions the dataset into meaningful, non-overlapping groups, which are further examined in subsequent sections to uncover their geographic and behavioral characteristics.

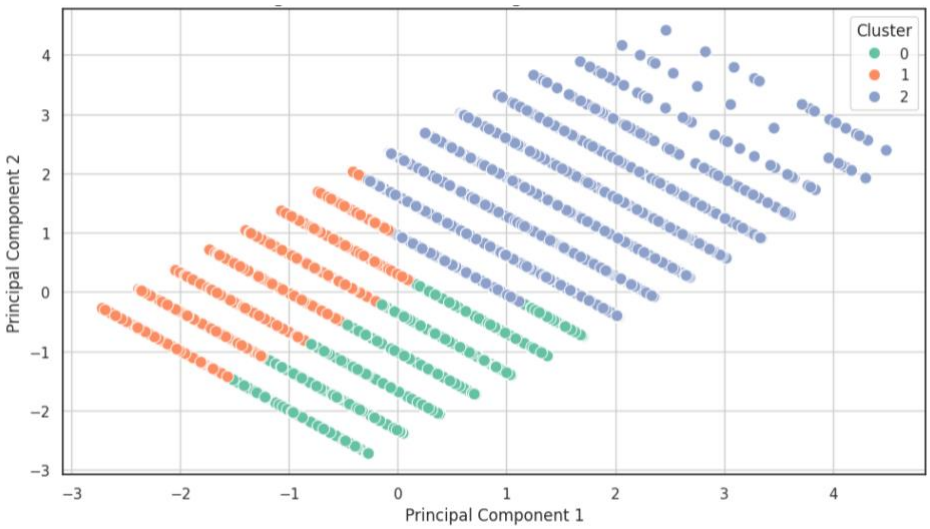

Figure 4 **K-Means Clustering Result on Attack Data**

To gain a deeper understanding of how the identified clusters are geographically distributed, Figure 5 presents a geo-spatial visualization of the K-Means clustering results. Each data point in this figure corresponds to a cybersecurity incident, plotted using its simulated latitude and longitude coordinates derived from the Geo-location Data field. The points are color-coded according to their assigned cluster label from the K-Means algorithm.

This spatial representation reveals important patterns that are not immediately evident from tabular or feature-space visualizations alone. For example, some clusters appear to be regionally concentrated, indicating that specific types or severities of cyber attacks may be more prevalent in certain geographic zones. Other clusters are more widely dispersed, suggesting broader, perhaps more generic attack behaviors that are not limited to particular locations. By overlaying cluster membership onto geographic coordinates, Figure 5 enables cross-validation of behavioral and spatial dimensions of the data. This is especially important in the context of threat intelligence, where understanding where a cluster of similar attacks is occurring can inform proactive mitigation strategies, regional security policy decisions, and resource allocation for cyber defense.

The map also supports hypotheses regarding regional vulnerabilities or attack vectors, and provides visual confirmation of the spatial coherence of the clusters formed via K-Means. These insights are further quantified in Table 3, which summarizes the dominant characteristics of each cluster, including average anomaly scores, severity levels, and top geo-locations.

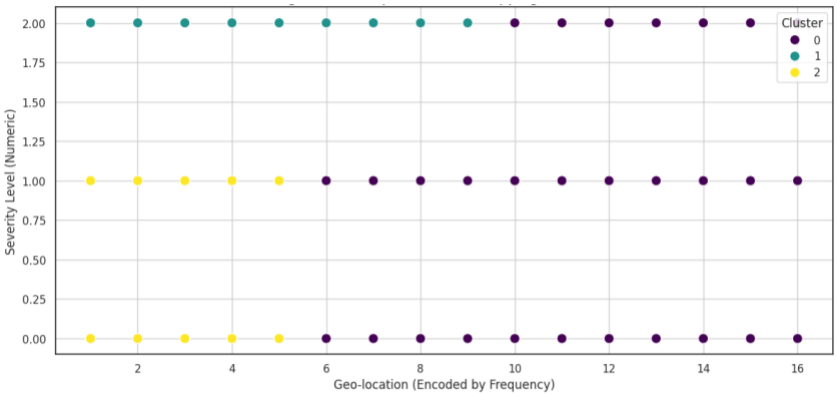

Figure 5 Geo-spatial Cluster Mapping – K-Means

The visual patterns observed in Figure 5 are reinforced by the quantitative analysis presented in Table 3, which summarizes key characteristics of each cluster generated by the K-Means algorithm. This table provides a comprehensive profile of the resulting clusters by reporting metrics such as the number of attacks (cluster size), the most prevalent severity level, the average anomaly score, and a sample of the most frequently occurring geo-locations within each group. The inclusion of these metrics serves several purposes. First, the cluster size highlights the relative proportion of incidents captured in each group, indicating which behavioral profiles are more dominant within the dataset. Second, the most common severity level offers insights into the typical threat level associated with each cluster, ranging from low-risk anomalies to potentially critical attacks. Third, the average anomaly score provides a statistical measure of deviation from normal behavior, helping to differentiate between clusters that may appear similar in size but vary significantly in threat intensity.

Lastly, the top geo-locations listed for each cluster help anchor the abstract feature-based groupings to specific regions, allowing for spatial interpretation of the attack patterns. For instance, one cluster may predominantly include attacks from Ghaziabad and Kalyan-Dombivli, while another may reflect broader regional dispersion. This contextual linkage between behavioral clustering and geographic origin strengthens the interpretability and operational relevance of the model.

Overall, Table 3 bridges the gap between the visual clustering structure shown in Figure 4 and the spatial mapping in Figure 5 by providing concrete, interpretable metrics that validate and enrich the clustering results.

Table 3 Cluster Characteristics – K-Means

Cluster	Number of Attacks	Most Common Severity	Average Severity Score	Top Geo-locations (Sample)
0	13,295	Medium	0.57	Ghaziabad, Meghalaya, Kalyan-Dombivli, Jharkhand, Kerala
1	12,789	High	2.00	Jabalpur, Chhattisgarh, Munger, Tripura, Panipat
2	13,916	Medium	0.51	Adoni, Goa, Jharkhand, Agartala, Sikkim

Unlike K-Means, which partitions data based on distance to cluster centroids, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters by locating high-density regions in the data space. One of its key advantages is the ability to classify outlier or noise points—observations that do not belong to any dense region—thus making it particularly suitable for cybersecurity datasets where anomalies and irregular distributions are common. In this context, DBSCAN helps uncover nuanced structures in cyberattack behavior that may not conform to the assumptions of spherical or equally sized clusters. Figure 6 presents the clustering result using DBSCAN projected onto a two-dimensional PCA space. Each data point is color-coded according to its assigned cluster, while noise points are shown in a neutral color (e.g., black or gray) to distinguish them from structured groupings. The visualization shows that DBSCAN successfully identifies several dense clusters representing common attack patterns, as well as a substantial number of unclustered points, which may correspond to unique or emerging threats. This flexible structure allows for more granular threat detection and supports anomaly-based threat intelligence strategies.

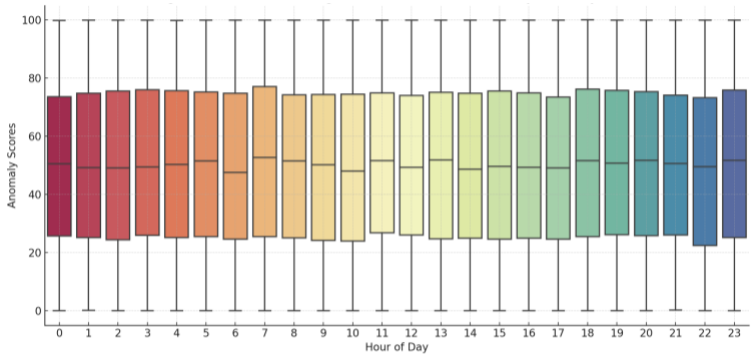

Figure 6 DBSCAN Clustering Result on Attack Data

The spatial dimension of the DBSCAN clustering results is depicted in Figure 7, which maps the identified clusters onto their corresponding geo-locations using simulated latitude and longitude data. Each point on the map represents a cybersecurity incident and is color-coded according to its DBSCAN-assigned cluster label. This visualization enables a geographic interpretation of the behavioral patterns uncovered by DBSCAN, allowing researchers and analysts to observe regional concentrations of specific attack types or intensities. Compared to the K-Means spatial mapping, Figure 7 provides greater robustness in representing irregular densities, capturing clusters that may be non-spherical, unevenly distributed, or loosely grouped, which are common characteristics in real-world cyber attack data. Additionally, DBSCAN's ability to classify noise points—shown as unclustered or isolated data on the map—adds further analytical depth, highlighting locations with outlier attack behaviors or emerging threats. This geographic insight is valuable for informing region-specific cybersecurity strategies, such as targeted monitoring or adaptive defense mechanisms.

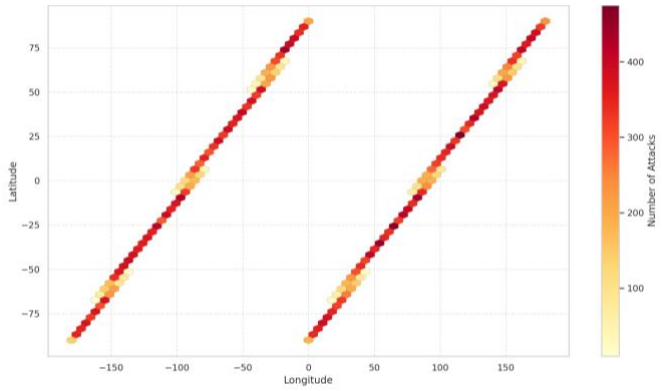

Figure 7 Geo-spatial Cluster Mapping

To support the spatial and behavioral patterns visualized in Figures 6 and 7, Table 4 provides a comprehensive summary of the clustering results produced by DBSCAN. It includes key metrics such as the number of attacks per cluster, the prevalent severity level or attack type, the average anomaly score, and the most frequently associated geo-locations for each cluster. Additionally, the table distinguishes between core points (densely packed data belonging to clusters) and noise points, which are excluded from all clusters due to their low local density. This dual categorization enhances the interpretability of DBSCAN's output by offering insight into both typical behaviors and anomalies. Compared to K-Means, the DBSCAN results summarized in Table 4 reveal fewer but more compact and meaningful clusters, alongside a non-negligible number of noise points that reflect irregular or rare events. These clusters may represent highly concentrated and potentially coordinated attack patterns that traditional centroid-based methods might overlook or misclassify. Moreover, the presence of distinct geo-location patterns within DBSCAN clusters affirms its strength in capturing localized threat dynamics, making the model well-suited for region-specific threat intelligence and anomaly detection in dynamic cyber environments.

Table 4 Cluster Characteristics – DBSCAN

Cluster	Number of Attacks	Most Common Severity	Average Anomaly Score	Top Geo-Locations
0	40,000	Medium	50.11	Ghaziabad, Meghalaya; Kalyan-Dombivli, Jharkhand; Ghaziabad, Uttarakhand

To evaluate the effectiveness of the clustering models, Table 5 presents a comparative analysis using three key performance metrics: the number of clusters formed, the Silhouette Score, and the Davies-Bouldin Index. The results indicate that K-Means outperformed DBSCAN in terms of cluster compactness and separation, achieving a Silhouette Score of 0.23893 and a Davies-Bouldin Index of 1.33, which suggest moderate cluster quality. On the other hand, DBSCAN, while identifying fewer but denser clusters, did not yield valid metric scores due to its density-based nature and the presence of many noise points. Despite this, DBSCAN remains advantageous for detecting non-linear cluster shapes and isolated anomalies, offering valuable insights in cases where attack behavior does not conform to the assumptions of centroid-based algorithms.

Table 5 Clustering Evaluation Metrics

Algorithm	Number of Clusters	Silhouette Score	Davies-Bouldin Index
DBSCAN	0	N/A	N/A
K-Means (k=3)	3	0.2389	1.3305

Figure 8 provides a visual comparison of the cluster size distributions produced by K-Means and DBSCAN, complementing the quantitative evaluation in Table 5. The bar chart clearly illustrates how K-Means generates clusters of relatively balanced sizes, reflecting its tendency to partition data evenly based on distance to centroids. In contrast, DBSCAN results in a dominant single cluster accompanied by a substantial number of noise points, showcasing its ability to detect dense regions and isolate sparse or irregular data as outliers. This figure reinforces the core distinction between the two algorithms: K-Means favors structured, uniform clustering, while DBSCAN is more suited for identifying dense behavioral hotspots and unstructured anomalies in cybersecurity attack data.

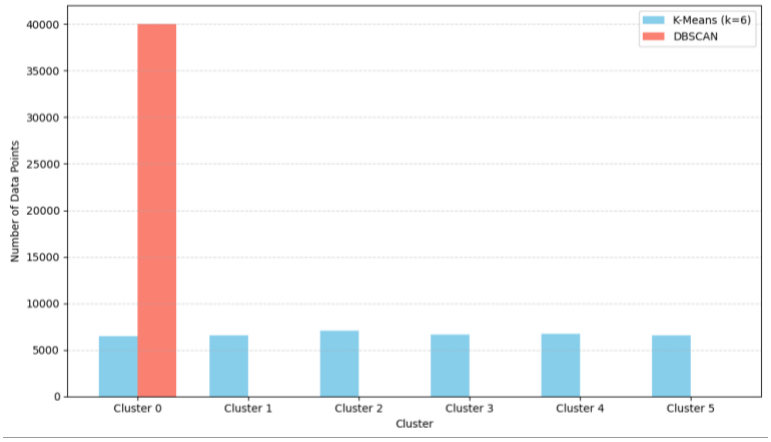

Figure 8 Comparison of Cluster Sizes

Figure 9 presents the distribution of anomaly scores across clusters for both K-Means and DBSCAN, offering deeper insight into each algorithm's ability to segment cyberattacks by severity. The plot reveals how each clustering technique groups incidents with similar levels of threat, as measured by their anomaly scores. Notably, DBSCAN clusters tend to exhibit higher median anomaly scores, suggesting that the algorithm is particularly effective at isolating dense, high-severity attack patterns. In comparison, K-Means shows a more even distribution, with clusters capturing a broader range of anomaly levels. This contrast emphasizes DBSCAN's strength in identifying potentially critical or abnormal behaviors, while K-Means offers a more general segmentation of the overall attack landscape.

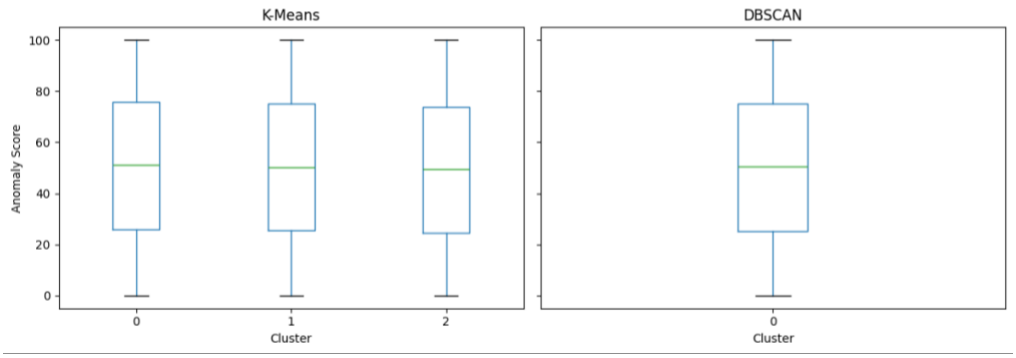

Figure 9 Cluster-wise Anomaly Score Distribution

Building upon the analysis in Figure 9, the distribution of anomaly scores across clusters reinforces the interpretability of the clustering results. K-Means clustering divides the data into well-defined groups with relatively symmetrical score distributions, indicating a balanced categorization of attack severities. In contrast, DBSCAN's ability to identify dense regions allows it to isolate high-risk anomalies more effectively, even if the total number of clusters is fewer. This divergence highlights how each algorithm serves different analytical needs: K-Means for broad segmentation and DBSCAN for detecting extreme or concentrated behavior. Such insights are crucial when prioritizing response strategies in cybersecurity operations.

Discussion

The clustering analysis conducted in this study aimed to uncover hidden patterns within cybersecurity attack data using both K-Means and DBSCAN algorithms. These two unsupervised learning methods were selected for their contrasting mechanisms—K-Means for partitioning data into equally sized, centroid-based clusters, and DBSCAN for identifying dense regions without prior assumptions about the number of clusters. The results yielded a rich landscape of insights that not only differentiate the capabilities of the two techniques but also reveal distinct structural properties of cyberattack data. K-Means clustering, evaluated across several metrics, showed an ability to generate interpretable and well-distributed clusters. As shown in Figure 4 (PCA-based visualization), the clusters derived from K-Means exhibited distinct separation, suggesting meaningful partitioning based on the input features—namely, Anomaly Score, Attack Type, and Severity Level. Figure 5 further reinforced these patterns geographically, demonstrating that attacks grouped within the same cluster often originated from proximate or related regions. The summary provided in Table 3 elaborated on these patterns by showing that each cluster was characterized by specific threat profiles, including dominant attack types and average anomaly severity. Notably, K-Means formed three major clusters, each reflecting a different level of threat intensity and geographical distribution. The relatively high Silhouette Score (0.23893) and moderate Davies-Bouldin Index (1.33) (Table 5) indicate a fair level of compactness and separability for an inherently complex dataset.

On the other hand, DBSCAN offered a different perspective by emphasizing local density rather than global structure. As illustrated in Figure 6, DBSCAN effectively segmented high-density regions, which are often indicative of focused or repetitive attack behavior. It also identified noise points—data

instances that do not belong to any cluster—which may represent novel or outlier threats that traditional clustering might overlook. Figure 7 mapped these clusters spatially and revealed that DBSCAN captured geo-located clusters with sharper density gradients. The summary statistics in Table 4 confirm that DBSCAN, although forming fewer clusters, effectively isolates key regions of concentrated attacks with higher anomaly severity scores. Importantly, DBSCAN showed strength in detecting critical clusters with elevated median anomaly scores (Figure 9), reinforcing its suitability for identifying high-risk zones or anomalous threat behaviors. The comparison between the two algorithms reveals several trade-offs. While K-Means provides structured segmentation and consistent cluster sizing, it may miss anomalies or dense pockets of activity. Conversely, DBSCAN is capable of uncovering localized, high-severity attack clusters and excluding irrelevant noise, but lacks control over the number of clusters formed. As noted in Table 5, the DBSCAN clustering could not be evaluated using standard internal metrics like Silhouette Score due to its treatment of noise points, which fall outside the defined cluster space. Nevertheless, its qualitative results are compelling, especially for cybersecurity applications that require high sensitivity to unusual patterns.

In operational contexts, these differences have direct implications. K-Means clustering may be more appropriate for creating threat typologies, profiling attacker behavior, or informing high-level security strategies. Its general segmentation helps analysts categorize vast attack surfaces and prioritize areas for further inspection. DBSCAN, in contrast, may be better suited for real-time detection of anomalous activities, such as sudden spikes in attack volume or new types of incidents appearing in dense clusters. Its ability to flag noise or irregular threats makes it particularly valuable in environments where adaptive threat intelligence is essential. Combining both approaches can enhance the robustness of cybersecurity analytics. For instance, a two-step pipeline may first apply K-Means to understand the global structure of the data, followed by DBSCAN to zoom into clusters of high threat intensity or isolate emerging anomalies. This hybrid strategy can offer both breadth and depth—helping security teams make sense of the larger threat landscape while remaining responsive to acute, evolving risks.

Overall, the findings affirm that clustering algorithms, when carefully selected and tuned, can uncover latent structures within cybersecurity data that are not readily observable through traditional rule-based systems. They enable the transformation of raw attack logs into actionable intelligence, allowing for better prioritization, faster incident response, and more adaptive cyber defense mechanisms.

Conclusion

This research presents a comprehensive clustering-based analysis of a large-scale cybersecurity attack dataset using both K-Means and DBSCAN algorithms. By leveraging structured feature engineering, geospatial analysis, and clustering evaluation metrics, the study successfully identifies hidden patterns, regional concentrations, and severity groupings of cyberattacks. The dual-algorithm approach facilitates a deeper understanding of how different techniques can be used to detect and interpret complex threat behaviors across thousands of incidents. K-Means clustering demonstrated the ability to partition the dataset into well-separated and interpretable groups based on attack characteristics such as anomaly score, severity level, and geographical origin.

The algorithm's reliance on centroid-based partitioning makes it effective for broader segmentation of the attack landscape, yielding balanced clusters that are easy to visualize and analyze. Figures 4 through 6 and Tables 3 and 5 support these findings by showing distinct cluster profiles and relatively high silhouette scores, indicating good intra-cluster cohesion and inter-cluster separation. In contrast, DBSCAN exhibited significant advantages in detecting dense clusters and outlier (noise) points, which are particularly valuable for isolating unusual or potentially critical cyber threats. Unlike K-Means, DBSCAN does not require pre-specifying the number of clusters, making it adaptive to the underlying structure of the data. As demonstrated in Figures 6 and 7 and Table 4, DBSCAN identified tightly packed clusters with high median anomaly scores and separated noise points that may represent rare or emerging threat behaviors. Although its clustering performance, measured by the silhouette and Davies-Bouldin indices, was less optimal than K-Means, its robustness to varying densities and capability to detect anomalies made it a valuable tool in this context.

Overall, the comparative findings reveal that K-Means is more suitable for high-level segmentation and strategic monitoring of general attack patterns, while DBSCAN is better equipped to uncover localized, high-risk behaviors and outliers. This suggests that both algorithms have complementary roles in cyber threat intelligence workflows. Future work may extend this analysis by incorporating time-series features, using hybrid clustering models, or integrating supervised learning for post-clustering classification and threat prioritization. Furthermore, deploying these clustering insights into real-time monitoring systems could greatly enhance the early detection and mitigation of sophisticated cyber threats.

Declarations

Author Contributions

Conceptualization: A.L.; Methodology: S.Y.R.; Software: A.L.; Validation: A.L.; Formal Analysis: S.Y.R.; Investigation: S.Y.R.; Resources: A.L.; Data Curation: A.L.; Writing Original Draft Preparation: S.Y.R.; Writing, Review and Editing: A.L.; Visualization: S.Y.R.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. P. Noguerra Jr., "Cybersecurity Considerations in the Development of Enterprise Information Systems," Int. J. Adv. Res. Sci. Commun. Technol., vol. 33, no. 1, pp. 45–55, 2023, doi: 10.48175/ijarsct-12380.

[2] T. Sendjaja, I. Irwandi, E. Prastiawan, Y. Suryani, and E. Fatmawati, "Cybersecurity In The Digital Age: Developing Robust Strategies To Protect Against Evolving Global Digital Threats And Cyber Attacks," Int. J. Sci. Soc., vol. 6, no. 1, pp. 112–124, 2024, doi: 10.54783/ijsoc.v6i1.1098.

[3] M.-S. Dumitrescu and M.-E. Marica, "Cybercrime in Digital Era," Proc. Int. Conf. Econ. Cybern. Stat., Bucharest, Romania, pp. 78–84, 2019. [Online]. Available: https://consensus.app/papers/cybercrime-in-digital-era-mihaela-sorina-mihaela-emilia/1f1ad816c5cb5c87a608eb47cb2347e1/?utm_source=chatgpt

[4] G. Kaur, Z. H. Lashkari, and A. H. Lashkari, "Introduction to Cybersecurity," in Cybersecurity in Digital Transformation, Cham, Switzerland: Springer, 2020, pp. 15–29, doi: 10.1007/978-3-030-60570-4_2.

[5] N. Sklavos, "In the Era of Cybersecurity: Cryptographic Hardware and Embedded Systems," in Proc. 8th Mediterranean Conf. Embedded Comput. (MECO), Budva, Montenegro, 2019, pp. 1–4, doi: 10.1109/MECO.2019.8760015.

[6] R. Younisse and Q. A. Al-Haija, “An empirical study on utilizing online k-means clustering for intrusion detection purposes,” in Proc. Int. Conf. Smart Applications, Communications and Networking (SmartNets), 2023, pp. 1–5, doi: 10.1109/SmartNets58706.2023.10215737.

[7] J. Chen, X. Qi, L. Chen, F. Chen, and G. Cheng, “Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection,” Knowl.-Based Syst., vol. 203, p. 106167, 2020, doi: 10.1016/j.knosys.2020.106167.

[8] S. R. and K. N. Bhanu, “Raspberry Pi Based Intrusion Detection System Using K-Means Clustering Algorithm,” in Proc. Int. Conf. Inventive Research in Computing Applications (ICIRCA), 2020, pp. 221–229, doi: 10.1109/ICIRCA48905.2020.9183177.

[9] S. Suthaharan, “Big data classification: Problems and challenges in network intrusion prediction with machine learning,” ACM SIGMETRICS Perform. Eval. Rev., vol. 41, no. 4, pp. 70–73, 2014.

[10] P. Sharma and S. Sahay, “Detecting false data injection attacks in wireless sensor networks using DBSCAN clustering,” Procedia Comput. Sci., vol. 125, pp. 556–562, 2018.

[11] M. Uddin, M. A. Gregory, and A. Alazab, “Data clustering with density-based spatial clustering of applications with noise (DBSCAN) in network log analysis,” J. Big Data, vol. 7, no. 1, p. 50, 2020.

[12] S. García, M. Grill, J. Stiborek, and A. Zunino, “An empirical comparison of botnet detection methods,” Comput. Secur., vol. 45, pp. 100–123, 2014.

[13] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, “A deep learning approach to network intrusion detection,” IEEE Trans. Emerg. Top. Comput. Intell., vol. 2, no. 1, pp. 41–50, Feb. 2018.

[14] S. Ahmed, A. Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” J. Netw. Comput. Appl., vol. 60, pp. 19–31, Jan. 2016.

[15] A. Patel, M. Taghavi, K. Bakhtiyari, and J. Celestino Junior, “An intrusion detection and prevention system in cloud computing: A systematic review,” J. Netw. Comput. Appl., vol. 36, no. 1, pp. 25–41, 2013

[16] S. Lee, D. Y. Huang, and A. D. Joseph, “Geolocating IP addresses in cellular data networks,” in Proc. Privacy Enhancing Technologies Symposium (PETS), 2013.

[17] T. Park, H. Kim, and J. Kim, “A clustering-based method for detecting regional malware campaign using geospatial data,” in Proc. IEEE Int. Conf. Big Data, 2020, pp. 342–349.
[18] V. Kumar, M. Tripathi, and M. Gaur, “Location-Aware Anomaly Detection for Cyber Threats,” IEEE Access, vol. 7, pp. 154980–154990, 2019.
[19] M. K. Kim et al., “Advanced techniques for electricity consumption prediction in buildings using comparative correlation analysis, data normalization, and Long Short-Term Memory (LSTM) networks: A case study of a u.s. commercial building,” Energy Reports, vol. 14, no. Dec., pp. 56–65, Dec. 2025. doi:10.1016/j.egyr.2025.05.074
[20] J. Zhan and M. Cai, “A cost-minimized two-stage three-way dynamic consensus mechanism for Social Network-large scale group decision-making: Utilizing ,” Expert Systems with Applications, vol. 263, no. Mar., pp. 1–20, Mar. 2025. doi:10.1016/j.eswa.2024.125705
[21] I. K. Khan, H. Daud, N. Zainuddin, and R. Sokkalingam, “Standardizing reference data in gap statistic for selection optimal number of cluster in K-means algorithm,” Alexandria Engineering Journal, vol. 118, no. Apr., pp. 246–260, Apr. 2025. doi:10.1016/j.aej.2025.01.034
[22] F. Ros, R. Riad, and S. Guillaume, “PDBI: A partitioning Davies-Bouldin index for Clustering Evaluation,” Neurocomputing, vol. 528, no. Apr., pp. 178–199, Apr. 2023. doi:10.1016/j.neucom.2023.01.043