Journal of Engineering Science and Technology
Vol. 19, No. 3 (2024) 869 - 879

COMPARATIVE STUDY OF k-MEANS AND MEAN SHIFT CLUSTERING ALGORITHMS FOR WASTE DATA IN WEST JAVA PROVINCE

RONY SETYAWAN*, GERALDI CATUR PAMUJI

Master of Information Systems, Universitas Komputer Indonesia, Indonesia
*Corresponding Author: rony.75122004@mahasiswa.unikom.ac.id

Abstract

This study presents a comprehensive comparative analysis of the k-Means and Mean Shift clustering algorithms, utilizing waste data collected from West Java Province's final disposal site spanning 2016 to 2021, with the primary objective of evaluating their performance and applicability for waste management practices; the analysis encompasses several critical parameters, including the number of clusters generated, variable uniformity, evaluation metrics employed, divergence measures, and processing time efficiency, revealing that k-Means, which formed three clusters, excels in rapid processing and provides finer cluster division, while Mean Shift, yielding two clusters, offers nuanced insights into data patterns, leading to the recommendation that the choice between the two algorithms should be driven by specific project requirements and considerations such as urgency of waste management needs and the depth of understanding desired for effective decision-making, thereby offering a tailored approach to waste data organization that optimizes categorization and contributes to more efficient and sustainable waste disposal practices in the future.

Keywords: k-means clustering, Mean shift clustering, Waste data.

1. Introduction

Clustering, an unsupervised machine learning method, groups unlabelled data into clusters based on similarities and differences [1]. Many algorithms handle numeric or categorical feature values [2]. Numerical features, like height and weight, contrast with categorical features, grouping data into predefined categories (e.g., colour, race). While distance-based measures suit numeric data, calculating categorical similarity is challenging. Diverse techniques address categorical data similarity [3]. In West Java Province, Indonesia, waste data reveals a yearly average of 40,251.93 units over the last 5 years, sourced from opendata.jabarprov.go.id. This dataset, spanning 2016 to 2021, details waste quantities in regencies and cities, contributing to an understanding of regional waste trends. Based on the theory and data, this study will compare two different algorithms, namely k-Means and Mean Shift.

2. Literature

A. k-Means

The k-means, a clustering technique, tackles the challenge of clustering by dividing n data points in a d-dimensional space into k clusters. Its objective is to minimize a cost function, usually representing the total distance between each point and its nearest centroid. Solving this clustering problem exactly is NP-hard, making k-means invaluable for delivering an approximate solution with a runtime of O(nkd) [4].

The k-means algorithm follows a simple procedure: it begins by randomly choosing k points from the dataset to serve as the initial centroids (seeds). Subsequently, every remaining point in the dataset is assigned to the cluster whose centroid is the nearest. The coordinates of these centroids are then recalculated as the mean of all points assigned to the cluster. This iterative process continues until the cost function converges to an optimal solution, although it might not guarantee a global optimum. Hence, the selection of initial centroids during initialization plays a pivotal role in obtaining the most optimal set of centroids [5].

The k-means holds a prominent position as a widely employed and extensively researched technique within the realm of data mining. Its fundamental goal involves the allocation of a dataset $D = \{pi \mid i = 1 \ldots n\}$, where each pi resides in a d-dimensional space, into k clusters, all initiated with arbitrarily selected initial centres. The overarching objective is to minimize the sum of squared error (SSE), as depicted in the provided formula. Within this equation, the term $pi - mj$ signifies the distance from data point pi to cluster center mj, the variable δij serves as a cluster indicator, taking the value $\delta ij = 1$ if pi belongs to cluster Cj and 0 otherwise, and the value of mj is determined as the mean of cluster Cj, computed through the formula.

$$SSE = \Sigma_i \Sigma_j (pi - mj)^2 \quad (1)$$

where: SSE: Sum of squared error (objective function), pi: Data point I, mj: Centroid of cluster j, Σ_i: Summation over all data points, and Σ_j: Summation over all clusters.

In the last five decades, k-means has garnered extensive recognition and favour for its efficacy in clustering endeavors spanning diverse fields [6-9], Despite its simplicity and efficiency, k-means encounters a notable drawback: the random selection of initial centroids can result in suboptimal outcomes, particularly when initialization is subpar. One prevalent challenge associated with k-means is its

propensity to split a substantial cluster into multiple smaller ones or combine small adjacent clusters into more extensive ones, ultimately yielding a minimal sum of squared errors (SSE). As illustrated in Fig. 1, inadequate initialization leads to the erroneous merger of two proximate clusters in the upper-right quadrant and the division of a single cluster into two subclusters in the lower-right corner as in Fig. 1.

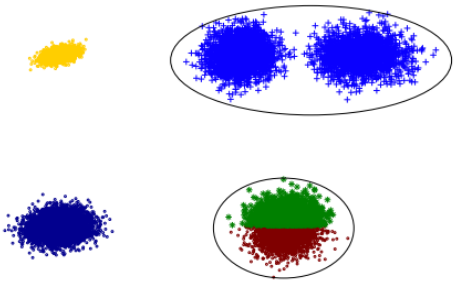

Fig. 1. illustration showcasing clusters starting with a poor initial setup [10].

B. Mean Shift

The Mean Shift algorithm is a data clustering and analysis method used to discover cluster centres in data. The primary objective of this algorithm is to identify cluster centres without requiring a predefined number of clusters. The Mean Shift algorithm operates by moving data points towards areas of higher density within a certain radius [11].

Suppose there are n instances x1, x2, ..., xn within the provided space. The MeanShift vector at position x is established as follows:

$$Mh(x) = 1/k \sum_{x_i \in S_h} (x_i - x) \quad (2)$$

S_h represents a sphere having a radius of h. It encompasses all points that fulfil the subsequent condition:

$$S(x) = \{y : (y - x)^T (y - x) \leq h^2\} \quad (3)$$

Figure 2 illustrates the scope of the larger circle denoting the Sh region. Inside, the smaller red dots symbolize the sample points within this area. The value k corresponds to the quantity of sample points situated in this region. The circle's center signifies the reference point x's location. Depicted as an arrow, the offset vector of the sample point xi concerning the reference point x is also presented. Notably, the yellow arrow vector in Fig. 2 represents the MeanShift vector [11].

In Eq.(2), (xi-x) represents the displacement vector of each data point in the vicinity compared to the reference point. Thus, Mh(x) indicates the average displacement vector of all data points within that region. This is premised on the assumption that all data points adhere to the same probability density distribution. Drawing from gradient knowledge, a non-zero probability density's gradient invariably aligns with the direction of maximal density increment. Consequently, data points in the pertinent vicinity tend to align along the gradient of probability density. It is reasonable to posit that Mh(x) aligns with the probability density function's steepest trajectory. As depicted in Fig. 2, Mh(x) orients towards areas with a denser distribution of data points.

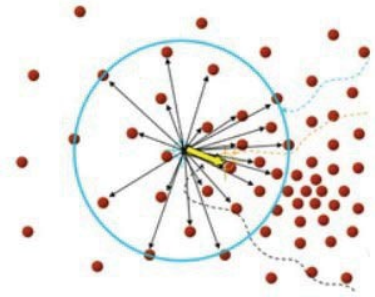

Fig. 2. Illustration depicting the Meanshift concept [11].

3. Method

This section of the research paper explains the method and steps in the algorithm.

A. Clustering process - k-means

1. **Data preprocessing:** Before applying the k-Means clustering algorithm, it is crucial to preprocess the waste data. This involves normalizing waste quantities to ensure consistent scaling across different features or variables. Normalization eliminates the potential bias caused by differences in the measurement scales of various waste parameters, ensuring that each feature contributes equally to the clustering process.
2. **Select number of clusters:** Determining the appropriate number of clusters (k) is a critical step in k-Means clustering. In this research, we employ methods such as the Elbow Method or the Silhouette Score to identify the optimal number of clusters. These techniques analyse the distribution patterns of waste data and help us make an informed choice for the value of k.
3. **Initialize centroids:** The k-Means algorithm begins by randomly selecting k data points from the dataset to serve as the initial centroids of the clusters. These centroids act as the representatives of each cluster and are pivotal in the assignment of data points to clusters.
4. **Assign data points:** In this step, each data point in the dataset is assigned to the nearest centroid based on a distance metric, typically Euclidean distance. This assignment determines which cluster each data point belongs to, based on its proximity to the centroid.
5. **Update centroids:** After all data points have been assigned to clusters, the algorithm recalculates the centroids. The new centroids are determined as the mean (average) of all data points belonging to their respective clusters. This step refines the centroid locations to better represent the data within each cluster.
6. **Repeat:** Steps 4 and 5 are iterated until a convergence criterion is met. Typically, this criterion involves checking if the centroids have stabilized or if a certain number of iterations have been completed without significant changes in cluster assignments.
7. **Results:** Upon convergence, the k-Means algorithm yields k clusters, each characterized by its centroid. These clusters represent distinct groups within the waste data, with data points grouped together based on their similarities.

The results provide valuable insights into how waste quantities are distributed and organized within the final disposal site, aiding in data-driven decision-making for waste management strategies.

B. Clustering process - Mean shift

1. **Data preprocessing:** Similar to k-Means, data preprocessing in the Mean Shift clustering process is essential. It involves normalizing waste quantities to ensure consistent processing. Normalization standardizes the scales of different waste parameters, preventing any disproportionate influence on the clustering results due to the varying measurement units of different features.
2. **Kernel selection:** One of the key decisions in Mean Shift clustering is choosing an appropriate kernel function. The kernel function defines the shape and scale of the region around each data point within which density is estimated. Commonly used kernel functions include Gaussian (Radial Basis Function), Epanechnikov, and others. The choice of kernel function can significantly impact the clustering results, as it affects the shape and sensitivity of the density estimation.
3. **Initial seeds:** Mean Shift starts with each data point considered as a potential seed for a mode (a high-density region). Each data point serves as the initial candidate for a cluster center.
4. **Kernel density estimation:** For each seed (initial data point), the algorithm estimates the density of data points within a certain distance defined by the chosen kernel function. This step involves calculating a weighted average of data points, where data points closer to the seed have a higher influence on the density estimation. This density estimation characterizes the local density around each seed.
5. **Shift seeds:** The algorithm then shifts each seed towards a higher density region within its local neighbourhood. This shift is performed by computing the mean of the data points within the defined kernel-based neighbourhood. The seed is moved to the point with the highest estimated density within its vicinity. This step gradually moves the seeds towards the modes of the data distribution.
6. **Repeat:** Steps 4 and 5 are iterated until the seeds converge. Convergence occurs when the seeds no longer move significantly, indicating that they have settled in high-density regions. The number of iterations required for convergence may vary depending on the data and kernel choice.
7. **Results:** Once convergence is achieved, the Mean Shift algorithm provides clusters represented by the converged seeds. Each seed represents a mode, and data points that are closest to a particular seed belong to the same cluster. The resulting clusters capture the underlying data patterns and density peaks within the waste data, aiding in the identification of distinct waste disposal patterns within the final disposal site.

C. Results comparison steps:

To rigorously compare the outcomes of the k-Means and Mean Shift clustering algorithms, we will consider several key factors, each of which plays a crucial role in assessing the performance and suitability of these algorithms for waste data analysis:

1. **Number of clusters:** The number of clusters formed by each algorithm will be carefully evaluated. k-Means necessitates a predefined number of clusters (k), which is determined beforehand. In contrast, Mean Shift dynamically identifies clusters based on data density, without requiring a predefined k. We will assess whether the number of clusters obtained by each algorithm aligns with the underlying structure of the waste data, ensuring that the clustering solution is meaningful and informative.
2. **Variables:** For a fair and meaningful comparison, both algorithms will employ the same set of variables for clustering. These variables, chosen based on their relevance to waste data analysis, will serve as the basis for grouping data points into clusters. Ensuring consistency in variable selection is vital to isolate the impact of the clustering algorithms on the results.
3. **Measure types:** The choice of similarity measures or distance metrics can significantly influence clustering results. We will thoroughly analyse the impact of different measure types, such as Euclidean distance, Mahalanobis distance, or the Silhouette Score, on the clustering outcomes. This analysis will provide insights into the sensitivity of each algorithm to the chosen similarity measure and its implications for cluster quality.
4. **Divergence:** Divergence measures the degree of dissimilarity between clusters. We will investigate the divergence between clusters formed by each algorithm to understand their separation and distinctiveness. Low divergence suggests that the clusters are well-separated and distinct, while high divergence indicates overlapping or poorly separated clusters. This assessment will shed light on the clustering algorithms' ability to capture meaningful distinctions within the waste data.
5. **Time taken:** The computational efficiency of clustering algorithms is crucial, particularly when dealing with large datasets. We will measure and compare the processing time required by each algorithm to cluster the waste data. This evaluation will help determine the suitability of k-Means and Mean Shift for practical applications, considering both the quality of results and the computational resources expended.

4. Results and Discussion

A. Data Collection

The dataset used in this research is taken from the open data west java website. The data used is waste data in the final disposal site in West Java Province from 2016 to 2021. The total data to be processed is 135 records.

B. k-Means

The k-means algorithm starts with the initialization step, where the number of clusters to form (denoted as 'k') needs to be determined. From the dataset, 'k' data points are randomly selected as initial centroids for each cluster. The selection of these initial centroids can impact the final clustering outcome. Techniques like k-means++ exist for more effective initialization.

After initialization, the subsequent phase involves cluster assignment. Every data point in the dataset is assigned to the cluster with the nearest centroid, often

determined by the Euclidean distance or other metrics. The Euclidean distance measures the separation between a data point and a cluster's centroid. This assignment groups similar data points together. The centroid results formed for each cluster are shown in Table 1.

Table 1. Centroids for each cluster.

Cluster	Garbage amount	Year
Cluster 0	25484.79796992	3229
Cluster 1	1655555.83	3257
Cluster 2	421593.90818182	3230

Following the assignment, the algorithm advances to updating cluster centres. Cluster centroids are recalculated by computing the average of all data points within that cluster. This shift pulls the centroids closer to the center of data points, enhancing cluster representation. Consequently, clusters better embody the data groups they encapsulate.

The assignment and centroid update processes repeat iteratively until convergence. Convergence implies either no further changes in cluster assignments or minimal movement of centroids. The k-means algorithm culminates in established clusters and their respective centroids. Data within a cluster shares similarities, while variations become more pronounced across clusters.

To determine the optimal 'k' value, the elbow method can be employed. The sum of squared distances (SSE) is calculated for various 'k' values. Plotting this SSE against the number of clusters often produces an "elbow" point, indicating an optimal number of clusters where the SSE begins to level off. This method helps choose the most suitable 'k' for clustering. Figure 3 is the scatter plot graph in each cluster.

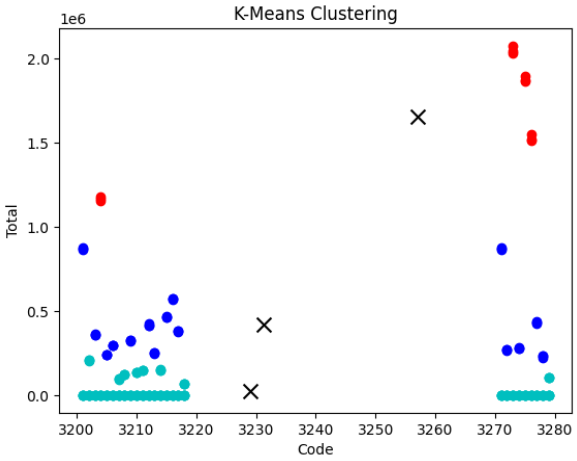

Fig. 3. Visualization of outcomes using a scatter plot obtained through the k-means clustering procedure.

C. Mean shift

The first step in the Mean Shift algorithm is initialization. Each data point in the dataset is given an initial point or "window," which serves as the search for the

mode or mass center. Then, for each initial point, the algorithm will shift towards a higher mass center until it reaches the convergence condition. This process is repeated for each data point until all points converge.

Next, the second step is mass center estimation. For each data point, the algorithm will update the location of the mass center based on the average of all data points within the specified "window" range. If there are other data points within the "window" range, the mass center point will be updated to a higher center. This process will continue until it reaches the convergence condition, where the mass center no longer moves.

After mass center estimation, the next step is data grouping. Each data point will be assigned to the corresponding group based on the identified mass center. Data within the same mass center range will belong to the same group. The final result of this algorithm is the formation of data groups based on the density of data points in the feature space Fig. 4.

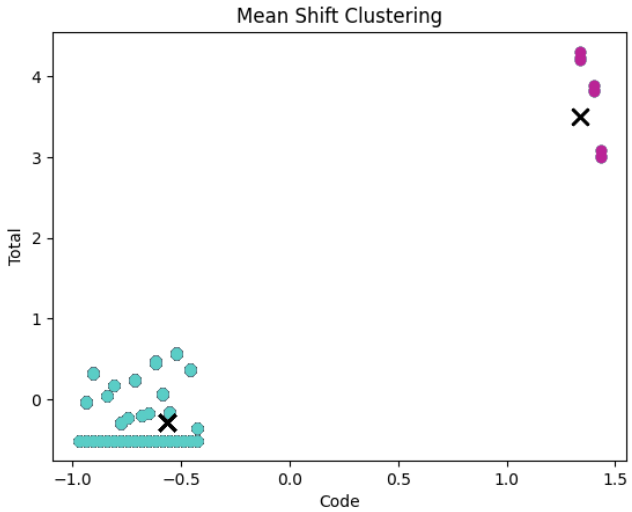

Fig. 4. Scatter plot result from the mean shift algorithm.

D. Result Comparison

In this research, a comparison was conducted among the number of clusters, variables, measure types, divergence, cluster instances, and processing time of the algorithm [12-15]. The results of the comparison between the two algorithms are presented in Table 2. This study gives information regarding this issue, as reported in elsewhere [16-21].

Table 2. Comparison results of k-means and mean shift.

Parameter	k-Means	Mean Shift
Number of Clusters	3	2
Variable	Length, Weight, Diameter	Length, Weight, Diameter
Measure Types	Euclidean Distance	Silhouette Score
Divergence	None	Kernelized Mean Shift
Time taken (seconds)	0.07700	67.84600

5. Conclusion

In this comprehensive analysis, we conducted an in-depth comparison between two prominent clustering algorithms, namely k-Means and Mean Shift, within the context of waste data organization. By evaluating these algorithms across a spectrum of critical parameters, we are now poised to make informed recommendations regarding their application in future waste management strategies.

Number of Clusters: A pivotal observation arising from our study is the discrepancy in the number of clusters generated by these algorithms. k-Means yielded three clusters, while Mean Shift produced two. This distinction carries significant implications. k-Means, with its ability to create more clusters, offers a finer-grained division of waste data into distinct groups, potentially providing clearer and more detailed insights into waste categorization.

Variable Uniformity: A fundamental strength of our comparative analysis lies in the uniformity of variables. Both algorithms leveraged the same attributes: Length, Weight, and Diameter. This consistent approach ensures the fairness and reliability of our comparisons.

Measure Types and Divergence: Methodological differences between k-Means and Mean Shift included the choice of evaluation metrics. k-Means employed the Euclidean Distance, while Mean Shift utilized the Silhouette Score and Kernelized Mean Shift divergence. These diverse metrics underscore the importance of aligning the evaluation method with the dataset's specific characteristics and research objectives. Mean Shift's utilization of nuanced metrics adds depth to our analysis.

Processing Time: An integral practical consideration, the computational efficiency of the clustering algorithms, revealed a significant contrast. The k-Means emerged as the frontrunner, completing its analysis in a mere 0.07700 seconds, while Mean Shift demanded considerably more time, consuming 67.84600 seconds for processing. This profound distinction underscores k-Means' suitability for large-scale datasets where expeditious results are paramount.

In light of these comprehensive findings, it is evident that the K-Means algorithm stands out as the preferred choice. Its swifter processing capabilities, combined with its capacity to generate a higher number of clusters, render it well-suited for real-time decision-making in waste management. While Mean Shift offers nuanced insights, its significantly longer processing time may limit its practicality in scenarios requiring rapid responses to waste data. The implications of these results for future waste organization efforts are profound. The adoption of the k-Means algorithm can enhance waste management strategies by delivering timely and detailed insights, facilitating optimized waste categorization, and contributing to more efficient and sustainable waste disposal practices.

Acknowledgement

We would like to thank Universitas Komputer Indonesia for assisting in writing of this paper.

References

1. Khan, S.S.; and Ahmad, A. (2004). Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, 25(11), 1293-1302.

2. Bulgarevich, D.S.; Tsukamoto, S.; Kasuya, T.; Demura, M.; and Watanabe, M. (2018). Pattern recognition with machine learning on optical microscopy images of typical metallurgical microstructures. *Scientific Reports*, 8(1), 2078.
3. Šulc, Z.; and Řezanková, H. (2019). Comparison of similarity measures for categorical data in hierarchical clustering. *Journal of Classification*, 36, 58-72.
4. Cohen-Addad, V.; Klein, P.N.; and Mathieu, C. (2019). Local search yields approximation schemes for k-means and k-median in Euclidean and minor-free metrics. *SIAM Journal on Computing*, 48(2), 644-667.
5. Dutta, C.; Bishop, L.D.; Zepeda O,J.; Chatterjee, S.; Flatebo, C.; and Landes, C. F. (2020). Imaging switchable protein interactions with an active porous polymer support. *The Journal of Physical Chemistry B*, 124(22), 4412-4420.
6. Alguliyev, R.M.; Aliguliyev, R.M.; and Sukhostat, L.V. (2021). Parallel batch k-means for big data clustering. *Computers and Industrial Engineering*, 152, 107023.
7. Braglia, L.; Lauria, M.; Appenroth, K.J.; Bog, M.; Breviario, D.; Grasso, A.; and Morello, L. (2021). Duckweed species genotyping and interspecific hybrid discovery by tubulin-based polymorphism fingerprinting. *Frontiers in plant science*, 12, 625670.
8. Lee, J.; Choi, I.Y.; and Jun, C. H. (2021). An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data. *Expert Systems with Applications*, 166, 113971.
9. Wang, A.; Liu, H.; Yang, J.; and Chen, G. (2022). Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data. *Computers in Biology and Medicine*, 142, 105208.
10. Ahmed, M.; Seraj, R.; and Islam, S.M.S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8), 1295.
11. Fiyad, H.M.N.; Metwally, H.M.B.; and Abd El, M.A.E.H. (2023). Improved real time target tracking system based on cam-shift and Kalman filtering techniques. *Journal of Applied Research and Technology*, 21(2), 297-308.
12. Pamuji, G.C.; and Rongtao, H. (2020). A comparison study of dB scan and k-means clustering in Jakarta rainfall based on the tropical rainfall measuring mission (TRMM) 1998-2007. *IOP Conference Series: Materials Science and Engineering*, 879(1), 012057.
13. Pangaribuan, I.; Rahman, A.; and Mauluddin, S. (2020). Computer and network equipment management system (CNEMAS) application measurement. *International Journal of Informatics*, *Information System and Computer Engineering (INJIISCOM)*, 1(1), 23-34.
14. Singgih, I.K. (2020). Air quality prediction in smart city's information system. *International Journal of Informatics*, *Information System and Computer Engineering (INJIISCOM)*, 1(1), 35-46.
15. Ginting, S.L.B.; Maulana, H.; Priatna, R.A.; Fauzzan, D.D.; and Setiawan, D. (2021). Crowd detection using YOLOv3-tiny method and viola-jones algorithm at mall. *International Journal of Informatics*, *Information System and Computer Engineering (INJIISCOM)*, 2(2), 13-22.
16. Riza, L.S.; Rosdiyana, R.A.; Pérez, A.R.; and Wahyudin, A. (2021). The k-means algorithm for generating sets of items in educational assessment. *Indonesian Journal of Science and Technology*, 6(1), 93-100.

17. Al Husaeni, D.N.; and Hadianto, D. (2022). The influence of spada learning management system (LMS) on algorithm learning and programming of first grade students at Universitas Pendidikan Indonesia. *Indonesian Journal of Multidiciplinary Research*, 2(1), 203-212.
18. Vijayarani, S.; Sivamathi, C.; and Tamilarasi, P. (2023). A hybrid classification algorithm for abdomen disease prediction. *ASEAN Journal of Science and Engineering*, 3(3), 207-218.
19. Obiwusi, K.Y.; Olatunde, Y.O.; Afolabi, G.K.; Oke, A.; Oyelakin, A.M.; and Salami, A. (2023). Evaluating the performance of supervised machine learning algorithms in breast cancer datasets. *ASEAN Journal of Science and Engineering*, 3(2), 179-184.
20. Vijayarani, S.; Sivamathi, C.; and Prassanalakshmi, R. (2023). Frequent items mining on data streams using matrix and scan reduced indexing algorithms. *ASEAN Journal of Science and Engineering*, 3(2), 123-138.
21. Khaleel, H.Z.; Ahmed, A.K.; Al-Obaidi, A.S.M.; Luckyardi, S.; Al Husaeni, D.F.; Mahmod, R.A.; and Humaidi, A.J. (2024). Measurement enhancement of ultrasonic sensor using pelican optimization algorithm for robotic application. *Indonesian Journal of Science and Technology*, 9(1), 145-162.