Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

Clustering and Modeling of Daily Weather Pattern Distribution in Makassar City Using Hybrid DBSCAN-Gaussian Mixture Model

Muhammad Risaldi[1], Ayu Safitri[2], Andi Akram Nur Risal[*3], Dewi Fatmarani Surianto[4], Dyah Darma Andayani[5], Marwan Ramdhany Edy[6], Firdaus[7], Jumadi M Parenreng[8]

[1,2,3,4,5,6,8]Computer Engineering, Faculty of Engineering, Universitas Negeri Makassar, Indonesia
[7]Electrical Engineering, Faculty of Engineering, Universitas Negeri Makassar, Indonesia

Email: [3]akramandi@unm.ac.id

Received : Aug 8, 2025; Revised : Oct 8, 2025; Accepted : Oct 21, 2025; Published : Oct 26, 2025

Abstract

Dynamic and irregular daily weather changes present major challenges in understanding seasonal patterns. Data uncertainty, outliers, and inter-season variability further complicate weather analysis using conventional methods. To address this issue, this study integrates Density-Based Spatial Clustering of Application with Noise (DBSCAN) and Gaussian Mixture Model (GMM) to analyze daily weather patterns in Makassar City. A total of 2,192 daily records from 2019 to 2024, including rainfall, specific humidity, atmospheric pressure, and wind speed, were examined. DBSCAN detected one dominant cluster (2019 data) and 173 outliers. The main cluster was further partitioned by GMM into three sub-clusters representing the wet (511 records, 13.39 mm rainfall), dry (633 records, 0.15 mm), and transition (875 records, 2.53 mm) seasons. GMM identified 1,764 fixed clusters and 255 ambiguous data points, with a log-likelihood of 5091.22 and the highest Silhouette Score of 0.188. Comparative evaluation demonstrated that the hybrid DBSCAN-GMM achieved superior performance (Silhouette Score = 0.1434) compared to DBSCAN or GMM individually. The novelty of this research lies in applying the DBSCAN-GMM integration, which is rarely used in tropical weather analysis, to capture seasonal structure and anomalies adaptively. This study contributes methodologically to clustering-based weather modeling and practically supports applications such as agricultural planning, disaster mitigation, and adaptive climate strategies in tropical regions.

Keywords : *Clustering, Daily Weather, DBSCAN, GMM, Seasonal Pattern.*

1. INTRODUCTION

Weather is the state of the atmosphere in a certain period whose nature continues to change over time. Weather is one of the important factors that affect the implementation of various activities, and has an impact on various other aspects of life [1][2]. Weather observations are generally made in a short period of time because the weather often changes unexpectedly. In addition, humans have limitations in covering large areas of observation [3]. Indonesia, as a tropical country, has highly variable rainfall. There are two types of climate patterns, namely 342 patterns included in the seasonal zone (ZOM) and 65 other patterns included in the non-seasonal zone (Non ZOM) [4].

In 2023, rainfall in Makassar City decreased from 310.16 mm in 2022 to 259.02 mm in 2023 [5]. Some factors that affect the weather include wind speed, wind direction, temperature, air humidity, and rainfall [6]. The diversity of rainfall is influenced by topographical, orographical, and geographical factors, which cause uneven distribution of rainfall between one region and another [7].

Weather forecasting has been done for a long time, usually only based on observations of event patterns, but such forecasting has proven to be unreliable [8][9]. Weather forecasting is the process of predicting the status of the atmosphere for a particular region or area of location by utilizing the latest technology [6][10]. The accuracy in predicting rainfall will affect many sectors such as marine,

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

agriculture, transportation, disaster management and others, so a method is needed in predicting rainfall, one of which is the DBSCAN algorithm and Gaussian Mixture Models (GMM). The DBSCAN and Gaussian Mixture Models (GMM) methods are used simultaneously to analyze daily weather patterns through clustering and data distribution modeling approaches [8][11].

Previous studies have explored the application of machine learning and deep learning in weather forecasting. Convolutional Autoencoders combined with k-means or Sinkhorn-Knopp variants have been shown to extract more representative weather features before clustering [12]. Ensemble learning methods like Random Forest, SVM, and XGBoost have been shown to enhance short-term forecasting acacuracy by minimizing errors in comparison to single models [13]. On the other hand, deep neural networks outperform conventional regression methods, although they still face challenges such as overfitting and the requirement for large datasets [14]. Furthermore, a systematic review of 500 publications highlighted the dominance of popular algorithms such as ANN, DL, RF, XGBoost, SVM, k-means, and PCA, while also emphasizing the research opportunity to develop more efficient and accurate methods for clustering daily weather distribution data [15].

Other studies have compared classical algorithms for weather forecasting, such as Decision Tree, k-NN, and Logistic Regression, which demonstrated relatively high accuracy but remained prone to overfitting due to limited testing sites [16]. The integration of the GEM physics-based model with AI-based GraphCast through spectral nudging improved forecast accuracy by up to 10% and enhanced tropical cyclone prediction, underscoring the importance of hybrid approaches [17]. Additionally, some studies explored the use of observations from Connected Vehicles as an alternative data source, demonstrating the potential of non-traditional data to enrich weather analysis [18]. Lastly, the EBWF ensemble model, which combines RF, GDBT, Naïve Bayes, and k-NN, achieved 94.9% accuracy, 95.7% precision, and 94.9% recall in fast processing for big weather data [19].

Beyond meteorology, several studies have employed Gaussian Mixture Models (GMM). One study, for example, introduced a combination of autoencoder, one-class SVM, and GMM for network intrusion detection based on normal data. The results showed that integrating GMM with representation learning techniques can improve anomaly detection with high accuracy, highlighting the potential of GMM to capture complex data distributions even beyond weather-related contexts [20]. Another study applied GMM in astrophysics to describe galaxy property distributions as a function of halo mass and redshift. Although outside the weather domain, the research demonstrated GMM's strength in modeling multidimensional data distributions [21]. Similarly, LSTM Autoencoder combined with GMM has been proposed for anomaly detection in autonomous vehicle trajectories. The approach could be adapted for daily weather data, which often exhibit time-series characteristics with seasonal variability and outliers [22]. Moreover, GMM has been used for analyzing the evaluation of concrete damage based on acoustic emission signals, showing its ability to classify experimental data in stages. Such applications suggest that GMM could differentiate varying daily weather conditions, such as transitions between extreme and normal states [23].

In addition, some studies have examined DBSCAN, such as the integration of RANSAC and DBSCAN for point cloud segmentation from laser scanning. DBSCAN proved effective in filtering outliers and clustering internal elements automatically, indicating its potential for handling unstructured weather data with uneven distributions [24]. Another study proposed STRP-DBSCAN, a parallel variant with random spatio-temporal partitioning and automatic parameter tuning. The results demonstrated improved clustering accuracy and time efficiency, which are crucial for large-scale and dynamic daily weather data analysis [25]. Furthermore, a comparative study of DBSCAN with other clustering algorithms for college admission data revealed that DBSCAN only formed a single cluster. Highlighting its limitations in datasets with low density. This is an important consideration for its application to weather data, where distributions may vary significantly [26].

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

Various studies have applied clustering approaches such as DBSCAN and Gaussian Mixture Models (GMM) to analyze complex datasets. However, these studies generally applied each method separately, which limits their ability to fully address the diversity and irregularity of dynamic weather data. DBSCAN is effective in detecting outliers and forming density-based clusters, while GMM provides flexibility in modeling probabilistic distributions. Yet, the integration of DBSCAN and GMM for daily weather pattern analysis in tropical cities has not been adequately explored, this research addresses that gap by proposing a hybrid DBSCAN-GMM approach to better capture seasonal variations and anomalies in weather data.

The objective of this study is to cluster daily weather patterns in Makassar City and to model rainfall distribution using DBSCAN to detect clusters and outliers, while GMM is employed for deeper probabilistic modeling of the main cluster. By applying this hybrid method, the study aims to produce a clustering analysis that can reveal seasonal patterns occurring in Makassar over the past six years and demonstrate the advantages of combining density-based and probabilistic approaches.

2. METHOD

In this research, a method with a series of sequential processes is used, including the stages of data collection, data normalization, DBSCAN, Gaussian Mixture Model (GMM), Visualization, and finally Analysis and Evaluation. These stages can be seen in Figure 1.

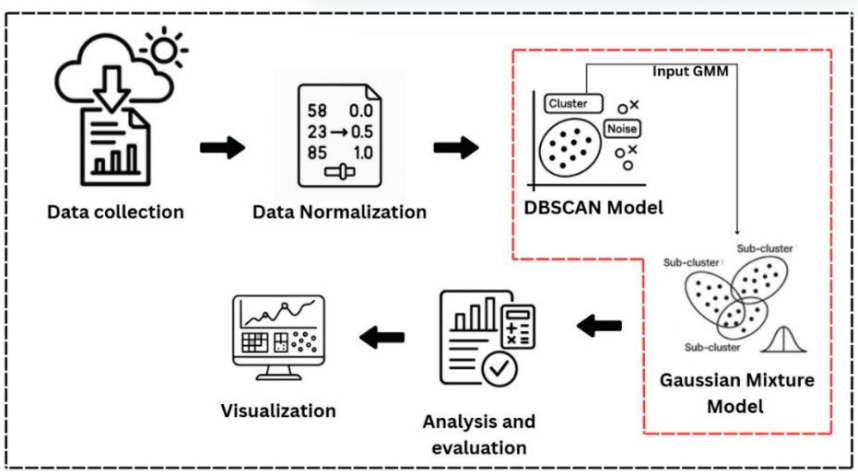

Figure 1. Research Stages

2.1. Data Collection

At this stage, the data used for clustering is obtained from NASA POWER on the website https://power.larc.nasa.gov/data-access-viewer/ with a total of 2,192 data from 2019-2024 which provides structured information related to daily weather data in Makassar. The data collected includes information on rainfall (mm), specific humidity (g/kg), atmospheric pressure (kPa), and wind speed (m/s). Data samples can be seen in Table 1.

Table 1. Data Sample

Year	Day	Rainfall (mm)	Specific Humidity (g/kg)	Atmospheric Pressure (kPa)	Wind Speed (m/s)
2019	1	9.1	18.88	99.89	3.21
2019	2	17.63	19.17	99.9	3.7
2019	4	7.17	18.87	100.07	2.99
2019	5	5.04	18.61	100.07	1.55

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

2.2. Data Normalization

At this stage, the data used is first normalized using the MinMaxScaler function which converts each value in the feature into a range between 0 and 1 [27]. This normalization is done because of the difference in units and value ranges between features. The equation used in the MinMaxScaler function is equation 1 [28]:

$$x_{scaller} = \frac{x - min\,(x)}{max\,(x) - min\,(x)} \quad (1)$$

With this method, all features have a balanced contribution in the clustering process, without the dominance of features with a larger value scale. The results of normalization using MinMaxScaller can be seen in Table 2.

Table 2. Normalization Results

Rainfall (mm)	Specific Humidity (g/kg)	Atmospheric Pressure (kPa)	Wind Speed (m/s)	Date
0.03885	0.70439	0.60439	0.49094	2019-01-01
0.07527	0.74300	0.61538	0.58953	2019-01-02
0.03437	0.73901	0.73626	0.56539	2019-01-03
0.03061	0.70306	0.80219	0.4466	2019-01-04
...	...	...	...	...
0.03676	0.77496	0.48351	0.40643	2024-12-31

2.3. Density-Based Spatial Clustering of Applications with Outliers (DBSCAN)

After the data is normalized, the initial clustering process is performed using the DBSCAN (Density-Based Spatial Clustering of Applications with Outliers) algorithm. In this study, the eps parameter value of 0.08 and min_samples of 6 were selected based on visual experiments using the k-distance graph. The eps value of 0.08 corresponds to the optimal neighborhood radius for separating dense regions from sparse ones. Meanwhile, min_samples = 6 was chosen by considering the data dimensionality (four main features: rainfall, humidity, pressure, and wind speed), following the common guideline that min_sample = 2 x dimension, while also adjusting to avoid excessive noise. DBSCAN was applied to the normalized data using MinMaxScaler, resulting in several clusters and a number of data classified as outliers (marked with a -1 label). The main cluster with the largest amount of data (label 0) was selected for further analysis using the Gaussian Mixture Model (GMM) method to explore the more complex distribution structure within it.

2.4. Gaussian Mixture Model (GMM)

After initial clustering using DBSCAN, further analysis was conducted using the Gaussian Mixture Model (GMM) method. This approach aims to understand the internal distribution structure in the main cluster obtained by DBSCAN. In this study, the GMM parameters used are n-components of 3, which are assumed to represent different weather distribution sub-patterns, a and random_state of 42. The choice of three components was not arbitrary, it was based on domain knowledge of Indonesia's climate, which is generally divided into three main seasonal conditions: rain season, dry season, and transitional period. Thus, n-components = 3 was selected to align with these natural seasonal categories while allowing probabilistic modeling of their internal variability. The GMM model is trained (fit) on the selected x_cluster data. The Gaussian Mixture Model (GMM) equation used is equation 2 [29][30]:

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot N(x \mid \mu_k, \Sigma_k) \quad (2)$$

The log-likelihood for models such as the Gaussian Mixture Model (GMM) is calculated by measuring how likely the model is to produce the data at hand [31]. The log-likelihood equation used is equation 3 [32]:

$$ln\, L_c(\psi) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} ln\, [\pi_k f_k(x_i; \theta_k)] \quad (3)$$

2.5. Analysis and Evaluation

The amount of data in each cluster of both DBSCAN and GMM was calculated and compared. In addition, the average value of each weather feature is calculated for each GMM sub-cluster, so that the main characteristics of each weather group can be interpreted. Silhouette score is also used to evaluate the quality of separation between clusters, with values ranging between -1 and 1 [33]. The equation used for Silhouette Score calculation is equation 4 [34][35]:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (4)$$

2.6. Visualization of Results

The GMM clustering results are visualized using PCA to reduce the four weather features into two dimensions, so that the sub-cluster distribution pattern can be seen clearly through the scatter plot. In addition, to see the seasonal trend of each sub-cluster formed through GMM, a time-based visualization using a line plot is performed. This visualization aims to observe changes in daily weather patterns occurring throughout the year in each sub-cluster.

3. RESULT

At this stage, the clustering results of the four main variables namely rainfall, specific humidity, atmospheric pressure, and wind speed are presented. In this section, the results of data clustering and average cluster formation using the DBSCAN (Density-Based Spatial Clustering of Applications with Outliers) algorithm are discussed, which aims to detect seasonal patterns and dominant weather groups without determining the number of clusters from the start. The results of DBSCAN clustering can be seen in Table 3.

Table 3. DBSCAN Clustering Results

DBSCAN Cluster	Number of Data
-1	173
0	2019

In Table 3, the results of applying the DBSCAN algorithm to normalized weather data resulted in the main cluster, cluster 0, which includes 2019 data. Based on these values, this cluster can be attributed to weather conditions that tend to be normal or common, possibly representing a mild dry season or stable weather in the region under study. Meanwhile, cluster -1 is noise data that was successfully detected by the DBSCAN model. This indicates that these outliers could represent weather anomalies

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

such as sudden rain, low pressure conditions, or other extreme weather events. Visualization of DBSCAN clustering results can be seen in Figure 2.

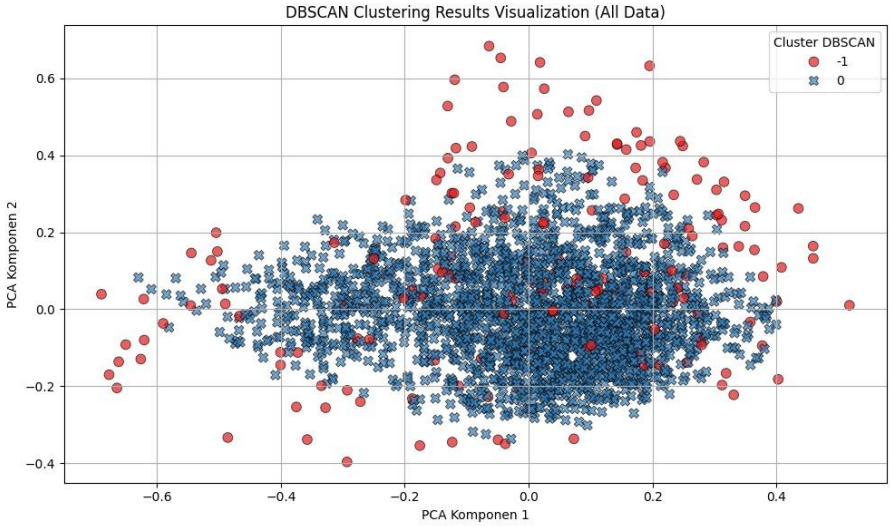

Figure 2. Visualization of DBSCAN Clustering Using Scatter Plot

Figure 2 shown the DBSCAN clustering results, where cluster 0 was identified as the dominant cluster, while the red points represent outlier. Based on this finding, a further clustering stage was conducted using the Gaussian Mixture Model (GMM) to identify seasonal patterns or finer sub groups within the main cluster. In this experiment, data from Cluster 0 was re-analyzed using GMM with a predefined number of three clusters, representing different weather characteristics, as shown in Table 4.

Table 4. Gaussian Mixture Models (GMM) Clustering Results

DBSCAN Cluster	Number of Data	Rainfall (mm)	Specific Humidity (g/kg)	Atmospheric Pressure (kPa)	Wind Speed (m/s)
0	511	13.3906	19.0244	99.7659	2.5002
1	633	0.1524	17.2043	99.8488	2.7761
2	875	2.5384	18.6788	99.7751	2.1172

Table 4 shows that GMM Cluster 0 consists of 511 data and represents wet weather conditions. This is indicated by the highest rainfall value of 13.3906 mm and the specific humidity which is also high at 19.0244 g/kg. Atmospheric pressure was 99.7659 kPa and wind speed was 2.5002 m/s. These characteristics indicate that this cluster is very likely to represent the monsoon period, where rainfall and humidity tend to increase significantly.

GMM cluster 1 consists of 633 data and represents dry weather conditions. This is indicated by the lowest rainfall value at 0.1524 mm and specific humidity at 17.2043 g/kg, as well as the highest atmospheric pressure of 99.8488 kPa and wind speed of 2.7761 m/s among the other clusters. This combination indicates that this cluster most likely represents the peak of the dry season, where air pressure tends to be high and winds are stronger with relatively low humidity. GMM cluster 2 consists of 875 data and reflects weather conditions that are in the transition phase between seasons, such as the transition from the dry season to the rainy season. This can be seen from the moderate rainfall value at 2.5384 mm and a fairly high specific humidity of 18.6788 g/kg, while atmospheric pressure is at 99.7751 kPa and wind speed is 2.1172 m/s. These characteristics indicate that the weather in this cluster is not too extreme, but has early signs of change towards the rainy season. Visualization of GMM clustering results can be seen in Figure 3.

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

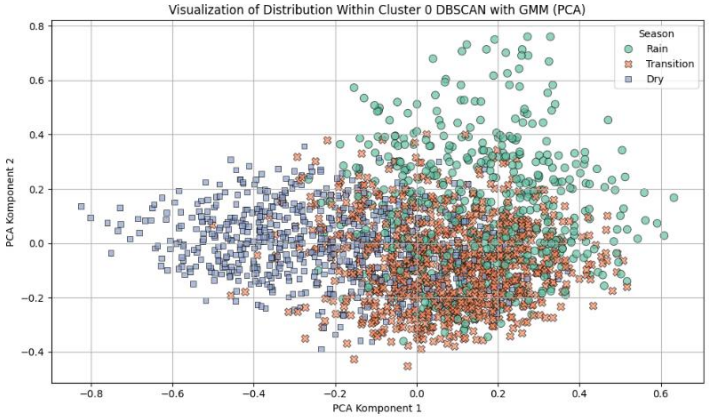

Figure 3. Visualization of GMM Using Scatter Plot

In Figure 3, the clusters formed can be measured by how close the data points are to their cluster compared to other nearby clusters using Silhouette Score. The average Silhouette Score results for each sub-cluster are presented in Table 5.

Table 5. Average Silhouette Score Results for Each Sub-Cluster

GMM Cluster	Silhouette Score	Log-Likelihood
0	0.046	
1	0.159	5091.22
2	0.188	

In Table 5, it can be seen that GMM cluster 0 has the lowest Silhouette Score value which indicates that the data in this cluster is similar to other clusters. GMM cluster 1 also has a fairly good separation. Meanwhile, GMM cluster 2 has the best separation quality compared to other clusters. The quality of GMM in clustering can be measured using log-likelihood which is designed for probabilistic models. Although the resulting Silhouette Score is low on average, it reflects the natural overlap in the daily weather data, which is not fully separated between seasons. The GMM is able to realistically map this transition by obtaining a log-likelihood value of 5091.22. This is quite a high value in matching the data to a Gaussian distribution.

A comparison of model performance was conducted to test the effectiveness of each method in classifying weather data accurately and meaningfully. The goal is to determine the extent to which the DBSCAN, GMM, and their combination (DBSCAN + GMM) models are able to identify weather seasonal patterns as well. By evaluating the number of clusters formed, the presence of noise and silhouette score, the results are shown in Table 6.

Table 6. Model Performance Comparison

Method	Number of Clusters	Noise Data	Silhouette score	Computation time
DBSCAN	1	173	-	0.0360s
GMM	3	-	0.0880	0.0483s
DBSCAN + GMM (Hybrid)	3	-	0.1434	0.1032s

Based on Table 6, the single DBSCAN approach only forms one main cluster and detects 173 data as noise or outliers. Since only one cluster is formed, the Silhouette Score calculation cannot be

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

done, considering that this metric requires a minimum of two clusters to measure the separation between groups. This method also has the fastest computation time of 0.0360 seconds. Meanwhile, the GMM approach run independently is able to form three clusters, but does not have a mechanism to detect outliers, so all data is considered valid and falls into one of the clusters. GMM produces a Silhouette Score value of 0.0880 with a computation time of 0.0483 seconds. The best results were obtained from the combined DBSCAN + GMM approach, where DBSCAN was used to identify and filter outliers first, then GMM was applied to the main cluster to form three sub-clusters. This approach resulted in the highest Silhouette Score value of 0.1434 and computation time of 0.1032 seconds, indicating that the integration of the two methods was able to improve the clustering quality despite requiring higher processing time. The difference in computation time across methods is influenced by algorithmic complexity: DBSCAN performs local density calculations once, leading to faster execution, while GMM requires iterative Expectation–Maximization steps that increase processing time. The hybrid DBSCAN–GMM approach combines both procedures, thereby introducing additional overhead but producing higher clustering quality.

The relationship between GMM clustering results and seasonality is proven based on the number of days of each calendar season. The heatmap results show that GMM Cluster 0 dominates the Rainy Season with 246 days, GMM Cluster 1 dominates the Dry Season with 408 days, and GMM Cluster 2 appears most in the Transition Season with 445 days. This pattern is in line with the characteristics of each cluster, where Cluster 0 has an average rainfall of 13.39 mm, Cluster 1 a low rainfall of 0.15 mm, and Cluster 2 of 2.53 mm, reflecting rainy, dry, and transitional conditions. Heatmap visualization can be seen in Figure 4.

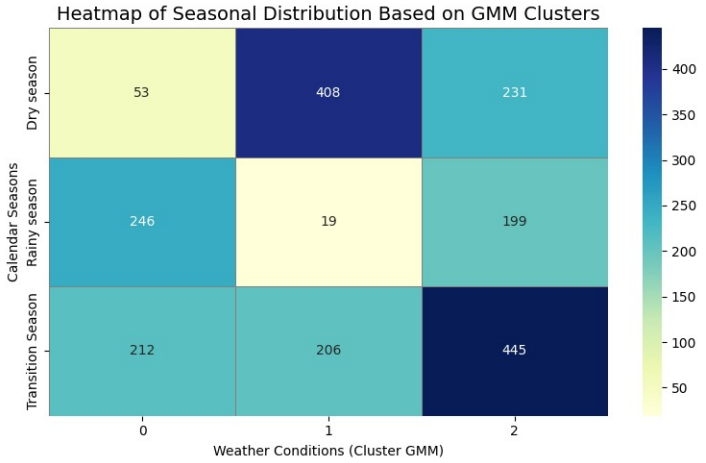

Figure 4. GMM Visualization Using Heatmap

Figure 4 illustrates the Gaussian Mixture Model (GMM) not only divides the data into several sub-clusters, but also provides probabilistic information on each data point, namely how likely a data point belongs to each cluster component formed. Through a probabilistic approach, GMM calculates the probability that an observation comes from each sub-cluster such as rain, dry, transition. This probability value is utilized to detect data that has a high probability of belonging to a cluster other than the cluster in which the data is classified.

In this study, the probability threshold for each season is determined based on the probability distribution of each cluster. Each season (Rain, Dry, or Transition) has a minimum and maximum probability range. The probability threshold is calculated as the average value between the minimum and maximum probabilities for each cluster. In other words, data that has a probability higher than this threshold is considered as data that definitely belongs to that cluster, which is called a fixed cluster. Conversely, data with a probability lower than the threshold is considered ambiguous to the cluster,

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

referred to as anomalous, as it has the potential to belong to other clusters. The threshold determination for each cluster is shown in Table 7.

Table 7. Determination of Threshold for Each Cluster

Season	Min Probability	Max Probability	Threshold
Rain	0.4925	1.0000	0.7462
Dry	0.4621	1.0000	0.7311
Transition	0.4468	1.0000	0.7234

In Table 7 the Rainy season has a threshold of 0.7462, indicating that data with a probability lower than this value is considered to not match the expected rainfall pattern. Likewise, the Dry and Transitional seasons have thresholds of 0.7311 and 0.7234 respectively, indicating the threshold probability for detecting anomalies within each season. This threshold value is used to distinguish data that definitely belongs to the season's cluster (Fixed Cluster) and data that is considered ambiguous or anomalous. Fixed cluster and ambiguous cluster can be seen in Table 8.

Table 8. Fixed vs Anomalous Data Per Season

Season	Fixed cluster	Ambiguous Cluster	Total
Rain	458	53	511
Dry	556	77	633
Transition	750	125	875

Based on Table 8, Rainy season has 458 data in Fixed Cluster and 53 data in Ambiguous Cluster out of 511 total data. The dry season has 556 data in fixed cluster and 77 data in ambiguous cluster out of 633 total data. Finally, the Transition season has 750 data in fixed clusters and 125 data in ambiguous clusters out of a total of 875 data. Visualization of the distribution of normal and anomalous data for each season can be seen in Figure 5.

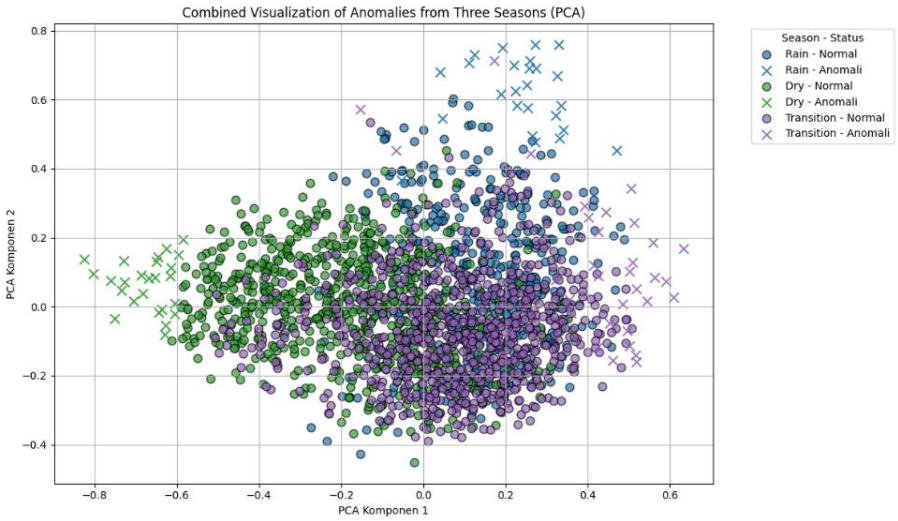

Figure 5. Seasonal Weather Distribution Map: Normal and Anomaly

In Figure 5, the status of anomalous data (ambiguous cluster) is marked with a cross symbol, while normal data (fixed cluster) is marked with a round symbol. Each color represents a different season. Blue for rainy season, green for dry season, and purple for transition season. Points that are distributed far from the center of the cluster usually represent anomalies, indicating weather conditions

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

that do not conform to typical seasonal patterns. This representation makes it clear how to visually recognize anomalies from the distribution of data in the reduced feature space.

The cluster distribution generated by the model provides an overview of seasonal patterns that can be seen from the distribution of data generated by the four variables used. Each variable has a characteristic seasonal pattern that provides information about the range of seasonal dominance in each month which can be seen in Figure 6.

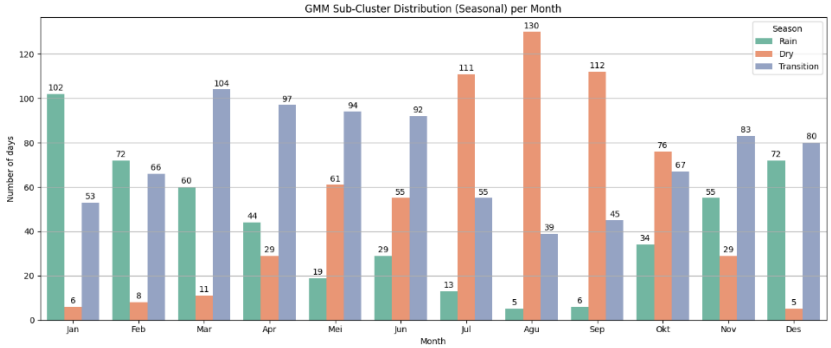

Figure 6. Visualization of Weather Pattern Seasonal Time

Figure 6 illustrates that each variable displays characteristics in each year and illustrates the dominance pattern in a particular month. The visualization in Figure 6 shows that the sub-clusters formed have different dominance tendencies throughout the year. The Rain sub-cluster appears most in January (102 days) and February (72 days), and again increases in November (55 days) and December (72 days), which shows the peak pattern of the rainy season in the tropics. Meanwhile, the Dry sub-cluster began to dominate from May (61 days), with a peak in August (130 days), before declining in October (76 days). The Transition sub-cluster dominates in transitional months such as March (104 days), April (97 days), and May (94 days), and increases again at the end of the year in November (83 days) and December (80 days).

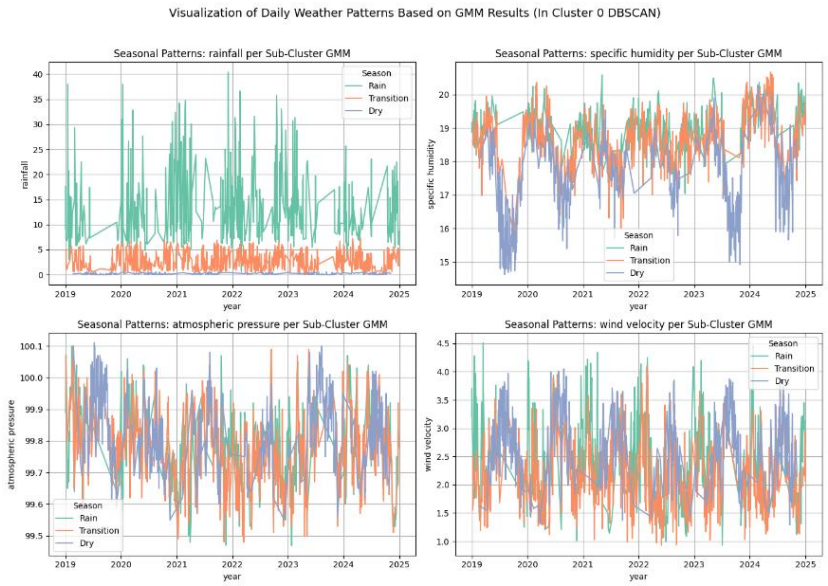

Figure 7. Visualization of Seasonal Pattern for Each Variable

Based on the visualization in Figure 7, the seasonal weather pattern that occurs in Makassar for the last 6 years shows that the rainy season occurs in the range of December to March. Conversely, very low rainfall in the range of June to September indicates the dry season. There are two transitional season

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

patterns, namely the transition to Drought occurs in April to May and the transition to Rain occurs in October to November. The conclusion of the seasonal timing of weather patterns is presented in Table 9.

Table 9. Conclusion of Seasonal Timing of Weather Patterns

Time	Dominant Cluster	Season Character
Dec- Feb	GMM 0	Rainy Season
Apr-May	GMM 2	Transition to Dry Season
Jun-Sep	GMM 1	Dry Season
Oct-Nov	GMM 2	Transition to Rain

Based on Table 9, the seasonal division is quite clear, depicting four distinct time periods grouped into three main clusters that characterize the seasonal pattern. The results prove that the annual seasonal pattern at the Makassar location can be well recognized using the DBSCAN-GMM combination.

4. DISCUSSION

The results of this study confirm that the hybrid DBSCAN–GMM approach is effective for analyzing daily weather patterns in Makassar, especially in distinguishing rainy, dry, and transitional seasons while detecting anomalous events. Compared to previous studies that applied machine learning and deep learning for weather forecasting, the hybrid clustering approach demonstrates a different strength. While deep learning models such as Convolutional Autoencoders or ensemble methods like Random Forest, SVM, and XGBoost often achieve high short-term prediction accuracy, they require large datasets and face challenges such as overfitting and low interpretability. In contrast, the DBSCAN–GMM hybrid offers a more computationally efficient and interpretable method, which is important for tropical regions with limited data resources.

Classical algorithms such as Decision Tree, k-NN, and Logistic Regression were previously reported to deliver good accuracy but remained vulnerable to overfitting due to limited testing sites. The hybrid method used in this study addresses this limitation by not relying on supervised learning or site-specific calibration, but instead by adaptively clustering daily weather data. Furthermore, while physics-based hybrid models such as GEM combined with AI-based approaches showed improved forecasting, they mainly focus on predictive accuracy, whereas the DBSCAN–GMM hybrid provides added value by explicitly identifying anomalies that represent extreme events in the tropical climate. This ability to capture both seasonal structures and anomalies highlights the novelty of the proposed framework within computer science-based weather analysis.

The findings are also consistent with regional meteorological observations in Indonesia and Southeast Asia, where rainfall patterns follow strong seasonal variability with transitional periods that are critical for agricultural and disaster management planning. The hybrid clustering successfully highlighted December–February as the rainy season, June–September as the dry season, and April–May and October–November as transitional months, which aligns with regional climatology reports. Importantly, about 13% of the data was identified as anomalous, representing unusual rainfall or atmospheric pressure shifts that conventional methods may overlook. This strengthens the relevance of the hybrid approach for tropical meteorology, where irregularities often carry significant societal impacts.

From a practical perspective, the results can directly support decision-making systems in agriculture, disaster mitigation, and early warning applications. Identifying transition periods is valuable for crop scheduling, while anomaly detection provides actionable insights for anticipating floods or droughts. For informatics and computer science, this research demonstrates how clustering and

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

probabilistic modeling can be combined to enhance the interpretability of climate data, bridging methodological advances with domain-specific applications.

Nevertheless, limitations exist. The dataset is restricted to one city over a six-year period, which constrains the generalizability of the findings to broader regions. In addition, the relatively low Silhouette Score indicates natural overlap in tropical weather patterns, suggesting that hybrid clustering alone cannot fully separate seasonal boundaries. Future studies should consider extending this framework to multi-city or multi-regional datasets, incorporating temporal dynamics, and exploring integration with deep learning models to improve predictive capabilities.

Overall, this study contributes a methodological advancement by integrating DBSCAN and GMM for clustering tropical weather data. The approach not only validates seasonal structures against regional meteorological patterns but also enhances anomaly detection, offering practical and scientific value for both computer science and climate-related applications in tropical regions.

5. CONCLUSION

The application of Gaussian Mixture Model (GMM) in the analysis of daily weather patterns in Makassar proved to be effective in identifying complex weather distributions through a probabilistic approach. The hybrid method (DBSCAN + GMM) provides the best performance compared to when used separately. This hybrid approach produces the highest Silhouette Score value of 0.1434. DBSCAN works to filter out outliers and form one main cluster while GMM divides the main cluster into three sub-clusters representing rainy, dry and transition seasons, and classifies the data into fixed cluster (87%) and ambiguous cluster (13%) categories. This ambiguous data reflects the presence of weather anomalies, such as sudden rains or unusual pressure changes, that do not fully fit the general seasonal pattern. The GMM's ability to reveal and map these anomalies makes it a powerful tool to support early warning systems and adaptive planning for extreme weather conditions. Practically, the findings can support stakeholders such as BMKG and local governments in developing adaptive strategies for agriculture, disaster mitigation, and early warning systems. Nevertheless, this study is limited to a single-city dataset (Makassar) covering six years, which may restrict generalizability to broader regions. future research should extend the approach by incorporating multi-city datasets, integrating with deep learning models, and testing long-term climate projections to enhance the robustness of weather pattern clustering.

REFERENCES

[1] M. Yusuf, A. Setyanto, and K. Aryasa, "Analisis Prediksi Curah Hujan Bulanan Wilayah Kota Sorong Menggunakan Metode Multiple Regression," *Jurnal Sains Komputer & Informatika (J-SAKTI)*, vol. 6, no. 1, pp. 405–417, 2022.

[2] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate Medium-ange global weather forecasting with 3D neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023, doi: 10.1038/s41586-023-06185-3.

[3] E. Fiola, F. Yulius, and D. M. Risani, "Metode Seleksi Variabel dalam Pemodelan Regresi Linear Data Curah Hujan Provinsi Lampung," in *Seminar Nasional Sains Data 2024 (SENADA)*, 2024, pp. 351–366. doi: 10.33005/senada.v4i1.213.

[4] D. E. Youlistia, R. Widiastuti, D. P. Meirylia, F. D. T. Amijaya, and A. A. Fauzi, "Analisis dan Prediksi Pengaruh Kelembaban Udara Terhadap Curah Hujan Bulanan 3 Wilayah Provinsi Kalimantan Utara Menggunakan Metode Regresi Berganda," *INTERVAL : Jurnal Ilmiah Matematika*, vol. 4, no. 2, pp. 72–78, 2024, doi: 10.33751/interval.v4i2.11653.

[5] A. Hafid and A. P. Islamy, *Statistik Daerah Kota Makassar 2024*, vol. 10, no. 1. Badan Pusat Statistik (BPS) Kota Makassar, 2024.

[6] M. Maulita and N. Nurdin, "Pendekatan Data Mining Untuk Analisa Curah Hujan Menggunakan Metode Regresi Linear Berganda (Studi Kasus: Kabupaten Aceh Utara)," *IDEALIS : Indonesia*

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

Journal Information System, vol. 6, no. 2, pp. 99–106, 2023, doi: 10.36080/idealis.v6i2.3034.

[7] R. Ismayanti and W. M. Baihaqi, "Prediksi Potensi Suatu Wilayah Untuk Menjadi PLTS Dengan Machine Learning," *Jurnal Informatika dan Teknologi Interaktif*, vol. 1, no. 2, pp. 66–72, 2024, doi: 10.63547/jiite.v1i2.6.

[8] A. Hot Iman, F. Ready Permana, G. Putro Wardana, R. Kemmy Rachmansyah, and M. Mega Santoni, "Perbandingan Algoritma Klasifikasi Random Forest dan Extreme Gradient Boosting pada Dataset Cuaca Provinsi DKI Jakarta Tahun 2018," in *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, 2022, pp. 593–601.

[9] Z. Ben Bouallègue *et al.*, "The Rise of Data-Driven Weather Forecasting A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context," *Bulletin of American Meteorological Society*, vol. 105, no. 6, pp. E864–E883, 2024, doi: 10.1175/BAMS-D-23-0162.1.

[10] S. Ventura, J. R. Miró, J. C. Peña, and G. Villalba, "Analysis of Synoptic Weather Patterns of Heatwave Events," Climate Dynamics, vol. 61, no. 9–10, pp. 4679–4702, 2023, doi: 10.1007/s00382-023-06828-1.

[11] E. Giama *et al.*, "Building Energy Simulations Based on Weather Forecast Meteorological Model: The Case of an Institutional Building in Greece," *Energies*, vol. 16, no. 1, 2023, doi: 10.3390/en16010191.

[12] T. Kurihana *et al.*, "Identifying Climate Patterns Using Clustering Autoencoder Techniques," *Artificial Intelligence for the Earth Systems*, vol. 3, no. 3, pp. 1–18, 2024, doi: 10.1175/aies-d-23-0035.1.

[13] E. A. Mohammed, X. Zhi, and K. A. Abdela, "Extreme Weather Patterns in Ethiopia: Analyzing Extreme Temperature and Precipitation Variability," *Atmosphere*, vol. 16, no. 2, p. 133, 2025, doi: 10.3390/atmos16020133.

[14] B. V. Malozyomov, N. V. Martyushev, S. N. Sorokova, E. A. Efremenkov, D. V. Valuev, and M. Qi, "Analysis of a Predictive Mathematical Model of Weather Changes Based on Neural Networks," *Mathematics*, vol. 12, no. 3, pp. 1–17, 2024, doi: 10.3390/math12030480.

[15] B. Bochenek, "Machine Learning in Weather Prediction and Climate Analyses — Applications and Perspectives," *Atmosphere*, pp. 1–16, 2022, doi: 10.3390/atmos13020180.

[16] M. Abdulraheem, J. B. Awotunde, A. E. Adeniyi, I. D. Oladipo, and S. O. Adekola, "Weather Prediction Performance Evaluation on Selected Machine Learning Algorithms," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 4, pp. 1535–1544, 2022, doi: 10.11591/ijai.v11.i4.pp1535-1544.

[17] S. Z. Husain *et al.*, "Leveraging Data-Driven Weather Models for Improving Numerical Weather Prediction Skill Through Large-Scale Spectral Nudging," *Weather and Forecasting (American Meteorological Society)*, vol. 40, no. 9, pp. 1749–1771, 2025, doi: 10.1175/waf-d-24-0139.1.

[18] A. R. Siems-Anderson, "The use of Vehicle-Based Observations in Weather Prediction and Decision Support," *Meteorological Applications*, vol. 31, no. 4, pp. 1–11, 2024, doi: 10.1002/met.2225.

[19] H. Shaiba *et al.*, "Weather Forecasting Prediction Using Ensemble Machine Learning for Big Data Applications," *Computers, Materials and Continua*, vol. 73, no. 2, pp. 3367–3382, 2022, doi: 10.32604/cmc.2022.030067.

[20] C. Wang, Y. Sun, S. Lv, C. Wang, H. Liu, and B. Wang, "Intrusion Detection System Based on One-Class Support Vector Machine and Gaussian Mixture Model," *Electronics*, vol. 12, no. 4, pp. 1–16, 2023, doi: 10.3390/electronics12040930.

[21] Y. Zhang *et al.*, "Characterizing the Conditional Galaxy Property Distribution Using Gaussian Mixture Models," *Astrophysical Journal*, vol. 16, no. 2, p. 159, 2023, doi: 10.3847/1538-4357/accb90.

[22] B. Wang, W. Li, and Z. H. Khattak, "Anomaly Detection in Connected and Autonomous Vehicle Trajectories Using LSTM Autoencoder and Gaussian Mixture Model," *Electronics*, vol. 13, no. 7, 2024, doi: 10.3390/electronics13071251.

[23] B. Yu, J. Liang, and J. W. W. Ju, "Damage Evolution Analysis of Concrete Based on Multi-Feature Acoustic Emission and Gaussian Mixture Model Clustering," *International Journal of Damage Mechanics*, vol. 33, no. 6, pp. 474–494, 2024, doi: 10.1177/10567895241235581.

Vol. 6, No. 5, October 2025, Page. 5237-5250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5254

[24] Harintaka and C. Wijaya, "Automatic Point Cloud Segmentation using RANSAC and DBSCAN Algorithm for Indoor Model," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 21, no. 6, pp. 1317–1325, 2023, doi: 10.12928/TELKOMNIKA.V21I6.25299.

[25] X. An *et al.*, "STRP-DBSCAN: A Parallel DBSCAN Algorithm Based on Spatial-Temporal Random Partitioning for Clustering Trajectory Data," *Applied Sciences*, vol. 13, no. 20, 2023, doi: 10.3390/app132011122.

[26] E. L. Cahapin, B. A. Malabag, C. S. Santiago, J. L. Reyes, G. S. Legaspi, and K. L. Adrales, "Clustering of Students Admission DataUsing K-Means, Hierarchical, and DBSCAN Algorithms," *Bulletin of Electrical Engineering and Informatics,* vol. 12, no. 6, pp. 3647–3656, 2023, doi: 10.11591/eei.v12i6.4849.

[27] M. A. Harriz and H. Setiyowati, "Komparasi Algoritma Decision Tree Dan Knn Dalam Mengklasifikasi Daerah Berdasarkan Produksi Listrik," *JIKO (Jurnal Informatika dan Komputer)*, vol. 7, no. 2, p. 167, 2023, doi: 10.26798/jiko.v7i2.787.

[28] A. I. Caniago, W. Kaswidjanti, and Juwairiah, "Recurrent Neural Network With Gate Recurrent Unit For Stock Price Prediction," *Telematika: Jurnal Informatika dan Teknologi Informasi*, vol. 18, no. 3, pp. 345–360, 2021, doi: 10.31515/telematika.v18i3.6650.

[29] N. N. Alyarahma, G. Kholijah, and C. Sormin, "Pengelompokan Provinsi di Indonesia Menggunakan Gaussian Mixture Model Berdasarkan Indikator Kemiskinan," *J Journal of Mathematic Theory and Applications (JOMTA)*, vol. 6, no. 2, pp. 158–167, 2024, doi: 10.31605/jomta.v6i2.4032.

[30] P. K. Mayakuntla, A. Ganguli, and D. Smyl, "Gaussian Mixture Model-Based Classification of Corrosion Severity in Concrete Structures Using Ultrasonic Imaging," *Journal of Nondestructive Evauation.*, vol. 42, no. 2, pp. 1–20, 2023, doi: 10.1007/s10921-023-00939-9.

[31] E. D. Suanda, W. Somayasa, and M. K. Djafar, "Estimasi Maksimum Likelihood Untuk Parameter Dalam Model Regresi Beta," *Jurnal Jurusan Matematika FMIPA*, vol. 3, no. 2, pp. 288–297, 2023, doi: 10.33772/jmks.v3i2.44.

[32] Z. Ren, "Optimal distribution-free concentration for the log-likelihood function of Bernoulli variables," *Journal of Inequalities and Application*, vol. 2023, no. 1, pp. 1–11, 2023, doi: 10.1186/s13660-023-02995-1.

[33] D. K. Wardy, I. K. G. D. Putra, and N. K. D. Rusjayanthi, "Clustering Artikel pada Portal Berita Online Menggunakan Metode K-Means," *Journal Ilmiah Teknologi dan Komputer (JITTER)*, vol. 3, no. 1, pp. 3–11, 2022.

[34] B. Rodrigues de Oliveira *et al.*, "Temporal Variability in Soybean Sowing and Harvesting According to K-Means and Silhouette Scores," *Trends in Agricultural and Environmental Sciences*, vol. 2, pp. 1–15, 2024. doi: 10.46420/taes.e240010.

[35] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International of Journal Online Biomedical Engineering*, vol. 19, no. 4, pp. 174–182, 2023, doi: 10.3991/ijoe.v19i04.37059.