JIRK
Journal of Innovation Research and Knowledge
Vol.6 No.6, Nopember 2026
https://doi.org/10.53625/jirk.v6i6.13489

Integrating Principal Component Analysis and Clustering for High-Resolution Meteorological Forecasting: A Case Study Using BMKG

Helga Charolina Antonia Silubun[1], Arya Dewa Haryanto[2], Rusly[3], Andi Azizah[4]
[1]Physics Education Lecturer, Faculty of Teacher Training and Education, Universitas Musamus
[2]Lecturer in Electrical Dept. Engineering Faculty, Universitas Negeri Yogyakarta
[3,4] Lecturer in Computer Education, Faculty of Teacher Training and Education, Universitas Musamus
Penulis Korespondensi: helga@unmus.ac.id

Received: 01/10/2026 | Accepted: 01/11/2026 | Publication: 05/11/2026

Abstract: *This research aims to integrate Principal Component Analysis (PCA) and clustering techniques in high-dimensional data analysis. The PCA method is used to reduce the dimensionality of the data without losing significant information, while clustering helps in grouping similar patterns in the data. This study applies this approach to the BMKG daily data dataset for the period 1 January 2024 to 19 November 2024, totaling 325 observations from BMKG Online Data at the Sleman DI Yogyakarta Geophysical Station, showing that the combination of PCA and clustering increases the efficiency and accuracy of the analysis. The research results show that the hybrid technique or integration between PCA, clustering (K-Means), and neural networks, is able to provide an effective solution for high-dimensional data analysis, producing accurate and reliable prediction models. This research can be the basis for developing a better weather early warning system and supporting more effective decision making in various sectors.*

Kata Kunci *: PCA, Cluster-K Means, Neural Networks, Weather Forecasting, Hybrid Methods, Data Analysis.*

Copyright © 2026, The Author(s).
This is an open-access article under the CC–BY-SA license.

How to Cite: Silubun, H. C. A., Haryanto, A. D., Rusly, & Azizah, A. (2026). Integrating principal component analysis and clustering for high-resolution meteorological forecasting: A case study using BMKG. *Journal of Innovation Research and Knowledge*, *6*(6), 3307–3324. https://doi.org/10.53625/jirk.v6i6.13489

INTRODUCTION

The integration of Principal Component Analysis (PCA) with clustering techniques for high-resolution meteorological forecasting represents significant progress in improving the accuracy of predictions and data reconstruction. PCA serves as an effective dimensionality reduction method, enabling more efficient processing of complex meteorological data. In this context, PCA helps identify the main features that influence meteorological variables, such as wind speed and temperature, making it easier to analyze and forecast (Geng, Zhang and Wu, 2020) (Balsa et al., 2023). For example, research by Geng et al. showed how PCA can be used to reduce the dimensionality of meteorological data that affects wind speed, which is then combined with an LSTM model to improve prediction accuracy (Geng, Zhang and Wu, 2020).

Additionally, the integration of PCA with clustering techniques allows grouping data based on similar characteristics, which can identify patterns that are not visible in the raw data. This is especially useful in the context of meteorological forecasting, where weather patterns are often complex and irregular. Research by Zhou et al. showed that considering meteorological features and seasonal information in prediction models can significantly improve prediction accuracy (Zhou et al., 2020). By using PCA to group meteorological data, researchers can reduce the noise and bias that often occur in traditional methods, resulting in more accurate and reliable models (Balsa et al., 2023).

Research gaps in the integration of clustering and principal component analysis (PCA) techniques in the context of weather prediction and meteorological data analysis indicate the need for further exploration to optimize these methods. Although PCA has proven effective in reducing dimensionality and increasing prediction accuracy, challenges remain in combining it with more complex machine learning algorithms. Research by Hur et al. showed that the combination of PCA with other methods can improve reconstruction performance, but the specific application in meteorology still requires further research (Lee et al., 2023). This is in line with the findings of Balsa et al. which emphasizes the need to develop hybrid methods that combine PCA and clustering techniques with deep learning algorithms to improve weather predictions and overcome the limitations of traditional forecasting methods (Gunawan, 2024). Integration of PCA with clustering techniques can also provide new approaches in meteorological data analysis. For example, research by Gunawan shows that a combination of PCA and cluster analysis can provide a better alternative compared to single cluster analysis (Duarte et al., 2022). In addition, research by Saputra et al. revealed that the use of hierarchical clustering on principal components (HCPC) techniques can produce better cluster solutions, which shows the potential to improve accuracy in the context of weather prediction (Yu et al., 2022). Thus, further exploration of how these techniques can be effectively integrated is essential to achieve better results in meteorological data analysis. Based on this, the objectives of this research are 1) Exploring a hybrid method that combines PCA and K-Means with a deep learning algorithm (Neural Network) to improve weather prediction performance. 2) Explore how the integration of PCA with clustering techniques can provide a new approach to meteorological data analysis.

LITERATURE REVIEW

1. Principal Component Analysis (PCA)

Concept Principal Component Analysis (PCA) is a widely recognized statistical technique employed for dimensionality reduction, which aims to retain as much variance as possible from the original dataset while transforming the data into a new set of uncorrelated variables known as principal components. This method is particularly beneficial in scenarios where datasets have a high number of features, allowing for a more manageable representation of the data without significant loss of information. PCA achieves this by identifying the directions (principal components) that maximize the variance in the data, effectively summarizing the underlying structure of the dataset (Abdi & Williams, 2010; Karamizadeh et al., 2013).

PCA is not only limited to theoretical applications but has been extensively utilized

across various fields, including bioinformatics, gait analysis, and environmental science. For instance, in gait analysis, PCA has been employed to reduce the complexity of kinematic data, allowing for a clearer understanding of movement patterns in individuals with gait disorders (Soares et al., 2016; Matsushima et al., 2017; Slijepčević et al., 2018). Similarly, in environmental studies, PCA has been used to predict air quality by reducing the dimensionality of the dataset while retaining essential information about air pollutants(Surono *et al.*, 2023).

Moreover, the versatility of PCA is evident in its integration with other machine learning techniques. For example, PCA is often used as a preprocessing step in conjunction with clustering algorithms like K-means, enhancing the performance of these algorithms by reducing noise and redundancy in the data. This combination is particularly useful in high-dimensional datasets, where the curse of dimensionality can significantly hinder the effectiveness of clustering methods (Mathivanan, Ghani and Janor, 2019).

The basic principle of PCA is to find the main components in the data. The main component is a linear combination of original features that maximizes the variance in the data. Following are the main steps of PCA

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Equation 1 is a dataset standardization formula (the feature scale is uniform), to ensure that all features contribute equally when analyzing, where x is the feature value, μ is the average feature, and σ is the standard deviation. Next, to analyze the relationship between features through the covariance matrix (correlation matrix), it can be seen in equation 2 to measure the relationship between variables in the dataset.

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad (2)$$

Next, we will obtain a matrix decomposition to calculate the eigenvector and eigenvalue of the covariance or correlation matrix. Where Eigenvector is a determinant of the direction of the main components, and Eigenvalue is a determinant of the amount of data variation explained by each main component. After that, determine the Main Components where the first main component (PC1) is the direction with the largest variance in the data, the second main component (PC2) is the direction with the next largest variance, and so on. All main components are orthogonal to each other. After that the data is transformed into principal component space using a linear combination of the initial features which can be seen through the following mathematical equation.

Suppose the initial data has dimensions n x m where n is the number of data, m is the number of features, then the covariance matrix is calculated as follows

$$Cov(X) - \frac{1}{n-1}X^T X \quad (3)$$

$$Cov(X)\,.v - \lambda \cdot v \quad (4)$$

Then the eigenvalue and eigenvector values are transformed using the eigenvector matrix

$$Z - X.V \quad (5)$$

Where λ is the Eigenvalue and v is the Eigenvector, Z is the data in the principal component space, and V is the eigenvector matrix. Then the percentage value of variance explained by each component (PC) can be calculated using the following equation

$$PC = \frac{\sigma^2}{\sum PC} x100\% \quad (6)$$

In conclusion, PCA serves as a fundamental tool in data analysis, providing a robust framework for dimensionality reduction while preserving the integrity of the original data. Its applications span numerous domains, demonstrating its utility in simplifying complex datasets and facilitating further analysis through various statistical and machine learning methods.

2. Clustering Concept (K-Means)

In the context of data analysis, K-Means is often used for market segmentation, user grouping, and behavioral analysis. For example, in research on customer segmentation, K-Means is used to group user characteristics based on data obtained from social media platforms such as Instagram (Habibi & Cahyo, 2019; Muhariya et al., 2022). This research shows that K-Means can identify different behavioral patterns among users, which can be leveraged for more effective marketing strategies. In addition, K-Means has also been applied in the field of biology to cluster gene expression profiles, demonstrating the flexibility of this method in various domains (Lèbre et al., 2010 ; Kashyap et al., 2020). One of the advantages of K-Means is its ability to handle large amounts of data quickly. This method is iterative, where at each iteration, the algorithm calculates the distance from each data point to the cluster centroid and updates the centroid position based on the average of the clustered data points (Venkateswarlu & Raju, 2013; Pokharel et al., 2021). However, choosing the right number of clusters (k) remains a challenge, as it can significantly influence the clustering results. Several studies have proposed methods to determine the optimal k value, such as using the bootstrap technique for segmentation analysis (Ernst & Dolničar, 2017; Yoseph et al., 2020). K-Means has also undergone various developments to improve its accuracy and efficiency. For example, improved KMeans algorithms have been developed for specific applications, such as earthquake data analysis and rice genotype clustering (Yuan, 2021; Kashyap et al., 2020). This research shows that with certain adjustments, K-Means can be adapted to meet more complex and specific analysis needs. Overall, K-Means Clustering is a very useful tool in data analysis, allowing researchers and practitioners to identify patterns and structure in unlabeled data. With its ability to adapt and be applied in various fields, K-Means remains one of the most widely used methods in data clustering. We can see the working principle in the following equation

$$J - \sum_{i=1}^{k} \sum_{j \epsilon C_i} \left\| x_j - \mu_i \right\|^2 \quad (7)$$

Where Ci is the ith member, μi is the ith Centroit, and ||xj − μi||2 is the Euclidean distance between xj and μi..

Applications of PCA and Clustering in Meteorology

Principal Component Analysis (PCA) and clustering techniques are increasingly utilized in meteorology to analyze complex datasets, identify patterns, and enhance understanding of atmospheric phenomena. PCA serves as a dimensionality reduction technique that simplifies the analysis of meteorological data by transforming correlated variables into a set of

uncorrelated variables known as principal components. This method is particularly useful in meteorology where datasets can be high-dimensional, containing numerous variables such as temperature, humidity, and wind speed, which may be interrelated. Arroyo et al. highlight that clustering techniques can effectively group meteorological data based on similarity measures, thereby facilitating the identification of distinct weather patterns and trends (Arroyo et al., 2015).

Clustering methods, such as k-means and hierarchical clustering, are widely applied in meteorological studies to categorize data into meaningful groups. Hidalgo and Jougla emphasize the application of k-means clustering to improve climate understanding in urban settings, suggesting that such techniques can enhance city planning and resource management (Hidalgo & Jougla, 2018). Additionally, Galatioto et al. employed Ward's minimum variance algorithm, a hierarchical clustering technique, to analyze pollution data in urban street canyons, demonstrating how clustering can elucidate the relationship between meteorological conditions and air quality (Galatioto et al., 2014). The integration of clustering with PCA allows researchers to not only reduce the dimensionality of the data but also to uncover underlying structures that may not be immediately apparent.

Moreover, the application of clustering techniques extends to the analysis of air pollution and its relationship with meteorological factors. Wang et al. discuss how clustering algorithms can reveal heterogeneity in meteorological observations and improve the understanding of air pollution dynamics (Wang et al., 2023). This is echoed by Zhang et al., who utilized clustering to identify circulation patterns conducive to precipitation extremes, thereby linking meteorological data to significant weather events (Zhang et al., 2022). The ability of clustering methods to handle large datasets and identify patterns makes them invaluable in meteorological research, as highlighted by Tian et al., who provide a comprehensive overview of clustering applications in meteorology (Tian et al., 2014).

In summary, the combination of PCA and clustering techniques in meteorology facilitates a deeper understanding of complex atmospheric data. These methods not only enhance the analysis of meteorological variables but also contribute to practical applications in urban planning and environmental monitoring. The ongoing development and refinement of these techniques promise to further advance meteorological research and its applications.

3. Neural Network

Neural networks have emerged as a pivotal component in artificial intelligence (AI) and machine learning, primarily due to their capacity to capture intricate patterns within data. The fundamental architecture of a neural network is structured into three main components: the input layer, one or more hidden layers, and the output layer. Each layer is composed of interconnected neurons(Malinowski, Rohrbach and Fritz, 2017);(Ba *et al.*, 2023), which function collectively to transform input data into meaningful outputs through a series of weighted connections and activation functions. This hierarchical processing enables neural networks to learn complex relationships and representations, making them particularly effective for tasks such as image recognition, natural language processing, and predictive analytics(Vasileva, Nevedrov and Subbotin, 2020);(Khairati and Putra, 2022).

Neural Network is a computational model consisting of several layers of connected neurons. Each neuron calculates a value based on the input it receives, weights, biases, and

activation functions.

$$Z = \sum_{i=1}^{n} w_i x_i + b \quad (8)$$

Where z is the value calculated by the neuron, wi is the weight for input i, xi is input i, b is the bias. Then the z value will be applied to the activation function a = f(z). The function is what determines how the signal from the neuron is processed, another form of activation function is the sigmod function in equation 9 and the Rectified Linear Unit (ReLu) equation 10 which converts the input value into a range (0,1) for binary classification tasks:

$$f(x) = \frac{1}{1+e^{-x}} \quad (9)$$

$$f(x) = \max(0, x) \quad (10)$$

The neural network training process involves optimizing weights to minimize the missing function, for example, Mean Squared Error equation 11, then to minimize the missing function, the Backpropagation algorithm equation 12 is used:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (11)$$

$$w_i = w_i - \mu \frac{\partial L}{\partial w_i} \quad (12)$$

Where μ is learning rate, $\frac{\partial L}{\partial w_i}$ is gradient of the loss function with respect to the wiights.

The neural network training process is 1). Forward Propagation: calculates network output based on input; 2). Calculating loss: comparing output with target using loss function; 3) Backpropagation: calculating gradients and updating weights and biases to minimize loss; 4) Iteration: repeating Steps 1-3 until the network reaches convergence.

METHOD

Research Design

This research uses a quantitative descriptive analytical research approach, with a case study research method using meteorological data from BMKG (https://dataonline.bmkg.go.id/data_iklim) as the primary data source, namely wind speed, temperature and rainfall data. rain, humidity, duration of sunlight. Secondary data in this research is literature and study results that are relevant to the use of PCA and clusters. The data collection procedure in this research uses BMKG data at the Sleman DI Yogyakarta Geophysical Station, Indonesia, at Latitude -7.82000, and Buju 110.30000, at Elevation 153, time span 01 January 2024 to 19 November 2024.Prosedur dan Teknik Analisis Data.

Data Analysis Procedures and Techniques Data analysis procedures and techniques are carried out in several stages, namely 1) Principal Component Analysis (PCA) aims to reduce data dimensions and identify the main components that influence meteorological variables. The data consists of 325 observations and 7 variables, when PCA analysis of the data was carried out, only 5 variables were taken, namely Average Temperature, Average Humidity, Rainfall, Duration of Sunlight, Average Wind Speed, Dominant Wind Direction. The remainder are wind direction variables and time information. 2). The Grouping Technique (K-MEANS) aims to group data based on PCA results. 3). Deep Learning (Neural Network), by integrating the results of PCA and Grouping analysis (K-MEANS) to increase the accuracy of

https://doi.org/10.53625/jirk.v6i6.13489

predictions and data reconstruction so that we can improve the performance of weather predictions with high resolution with the following network architecture design: Input to neural comes from the results Cluster (KMeans) and PCA, using two hidden layers with 64 neurons each, using the ReLU activation function, and a dropout rate of 40% to prevent overfitting (table 2). The output layer has 5 neurons that correspond to the number of predicted weather variables. The training parameters in Neural Network consist of: 1) Loss Function, namely Mean Square Error (MSE), used to measure the error between predictions and actual values; 2) Optimizer, namely Adam Optimizer, is used to optimize the model training process; 3) Metrics, namely Mean Absolute Error (MAE), are used as additional information to evaluate model performance; 4) Epochs of 50 is the number of model training iterations; and 5) Batch Size of 32, is the number of samples processed before the model is updated.

A	B	C	D	E	F	G
Average Temperature	Average Humidity	Rainfal	Duration of Sunlight	Average Wind Speed	Dominant Wind Direction	Tanggal
30,5	78		3,4	2	SW	01/01/2024
29,9	79	0,8	7,8	4	SW	02/01/2024
27,8	86	0	7,5	2	C	03/01/2024
27,9	83	4,3	2,4	2	S	04/01/2024
27,2	87	26,1	4,1	2	SW	05/01/2024
28,3	85	11,9	2,6	1	C	06/01/2024
27,3	90	17,6	4,6	2	SE	07/01/2024
28,2	86	20,9	0,6	1	C	08/01/2024
29,1	82	0,8	0,7	2	S	09/01/2024
29	83	0	5	2	SW	10/01/2024
29,5	81	8,7	7,2	1	S	11/01/2024
29,3	79	0	4,5	4	SW	12/01/2024
28,7	79	0	8,6	2	W	13/01/2024

Figure 1. Meteorological Data in Excel File

Table 1. Neural Architecture Model: Model: Sequential

Layer (type)	Output Shape	Param
Dense	(None, 64)	1664
Dropout	(None, 64)	0
dense_1 (Dense)	(None, 64)	4160
Dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 5)	325
Total Params: 6,149		
Trainable params: 6,149		
Non-trainable params: 0		

JIRK
Journal of Innovation Research and Knowledge
Vol.6 No.6, Nopember 2026
https://doi.org/10.53625/jirk.v6i6.13489

RESULTS AND DISCUSSION

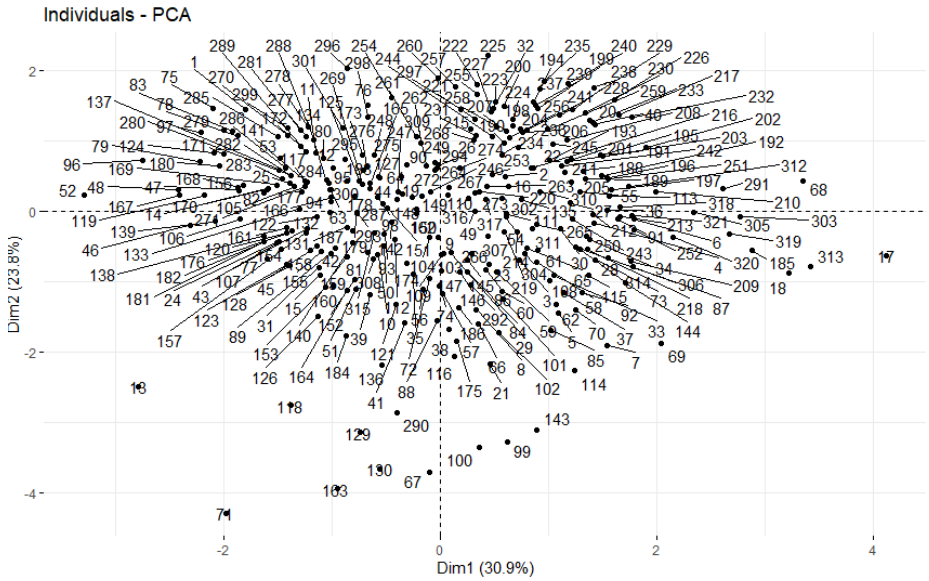
1. Ouput PCA

```
> pca_result
Standard deviations (1, .., p=5):
[1] 1.2421394 1.0915269 0.9647151 0.9612770 0.6410381

Rotation (n x k) = (5 x 5):
                              PC1         PC2         PC3         PC4         PC5
Average Temperature  -0.60391879 -0.41114881 -0.20213261  0.21191126 -0.61682638
Average Humidity      0.71640165 -0.05829611 -0.02766729 -0.09922442 -0.68757492
Rainfall             -0.07607876 -0.52878841  0.23022773 -0.81008096  0.07320424
Duration of Sunlight -0.29914155  0.56029697  0.65881413 -0.18287767 -0.35930675
Average Wind Speed   -0.16363718  0.48374409 -0.68653920 -0.50554201 -0.11093126
> |
```

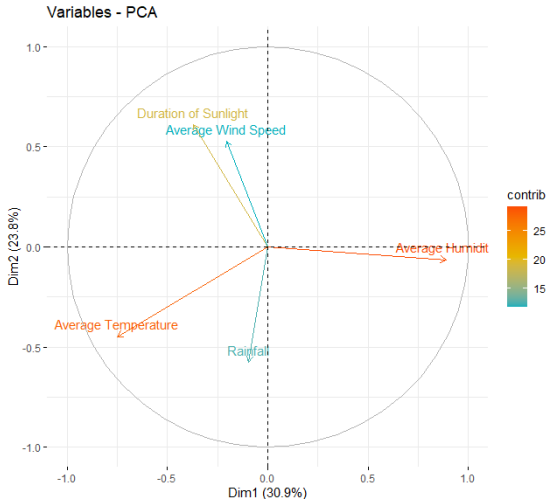
Figure 3. Rstudio Consule display for PCA analysis results output

Figure 4a is an individual PCA graph showing the distribution of individuals (observations) in two dimensions formed by the main components (PC1 and PC2). Dimension 1 (Dim1 30.9%) and dimension 2 (Dim2 23.8%) are the two main components that explain most of the variation in the data. These two dimensions combined account for 54.7% of the total variation in the data. From this picture we can see the distribution of individuals (observations) at each point representing one individual (observation) in the data. Individual positions indicate how they relate to dimensions Dim1 and Dim2. The distribution pattern can be seen from the points spread throughout the graph with some individuals being more separated (such as individual points number 71, 143, 67, 99, and 100), these points may have unique characteristics compared to other groups. Interpretation



(a). Indivudals PCA

https://doi.org/10.53625/jirk.v6i6.13489



(b) Variables PCA

Figure 4. PCA Output Graph: (a) Individuals PCA, (b) Variables PCA

Figure 4b. is a PCA variable graph, which shows the relationship between the original variables and their contribution to the main components. Vector Direction and Length It can be seen that each vector (arrow) represents an original variable (Average Temereture, Rainfall, Average Humidity, Duration of sunlight, Average Wind Speed), with the direction showing its contribution to Dim1 and Dim2. The length of the vector shows the strength of the variable's contribution to the main component, where variables with a long vector are more significant in explaining variance in the data. Interpretation of the Dim1 variable (30.9%), shows that the Average Humidity variable and the Average Temperature variable make a large contribution, as indicated by the vector length that leads to Dim1. Dim2 (23.8%) is influenced by Duration of Sunlight and Average Wind Speed which has a vector direction to Dim2. Meanwhile, the Rainfall variable contributes significantly, but the direction shows an influence on both dimensions, although it is more dominant in the Dim1 direction. The relationship between variables can be seen through the vector direction of each variable. Vectors that approach each other show a positive relationship between variables, such as Average Humidity and Rainfall have a small angle, indicating a positive correlation. Vectors that are perpendicular to each other, such as the variables Avarage Temperature and Duration of Sunlight, show no significant relationship, relatively uncorrelated. For vectors in opposite directions, such as Average Humidity and Duration of Sunlight, which show a negative correlation, they tend to be the opposite.

From the two graphs presented in Figure 4, we can see that individual points that are separate or far from the center, such as individual points 71, 143, and 67, have extreme values for variables that dominate the direction, such as individual points in the positive Dim1 direction have high values for Average Humidity or Average Temperature.

The Meaning of PC Components in the Context of BMKG Data

From these results, the main components can help explain patterns in BMKG, PC1 and PC2 data; Explains the main variability in meteorological data, namely the relationship between wind, humidity, and temperature. PC1 shows that strong winds tend to have an inverse relationship with humidity, PC2 differentiates the effect of temperature on humidity. PC3 and PC4; Highlights the relationship between rainfall and sunlight. PC3 shows regions with different characteristics based on the amount of sunlight versus rainfall. PC4 provides more insight into the dominant rainfall patterns.

2. Ouput Claster (K-Means)

Cluster size, cluster 1 has 138 individual members, cluster 2 has 79 individual members, and cluster 3 has 104 individual members. This data shows that the number of points/individuals in each cluster is not balanced, cluster 2 is smaller than the other two clusters.

Table 1. Cluster coordinate center (Centroid)

Claster	PC1	PC2	PC3	PC4	PC5
1	-1.156.697	0.034590	0.082128	-0.010464	-0.118560
2	0.767432	-0.982774	-0.776297	0.416523	0.119859
3	0.951895	0.700631	0.480709	-0.302512	0.066274

The cluster center (cenetroid) shows average characteristics based on dimensions, in cluster 1: PC1 (-1.1567) has a negative value indicating that PC1 characteristics are brought to the average, PC2 (0.0346) has almost neutral characteristics on this dimension. PC3, PC4, and PC5 have small average values, this shows relatively homogeneous characteristics. Cluster 2: PC1 (0.7674) has a large positive value, indicating that members of this cluster tend to be high, PC2 (-0.9828) has a strong negative value indicating that cluster members have characteristics below the average dimension, PC3, PC4, and PC5 tend to negative value or zero. Cluster 3: PC1 (0.9519) positive value is as high as cluster 2, but higher, PC2 (0.7006) positive value shows a different pattern from cluster 2 which is negative, PC3, PC4, and PC5 have positive values with subtle differences to caster 2.

ISSN 2798-3471 (Cetak)
ISSN 2798-3641 (Online)

https://doi.org/10.53625/jirk.v6i6.13489

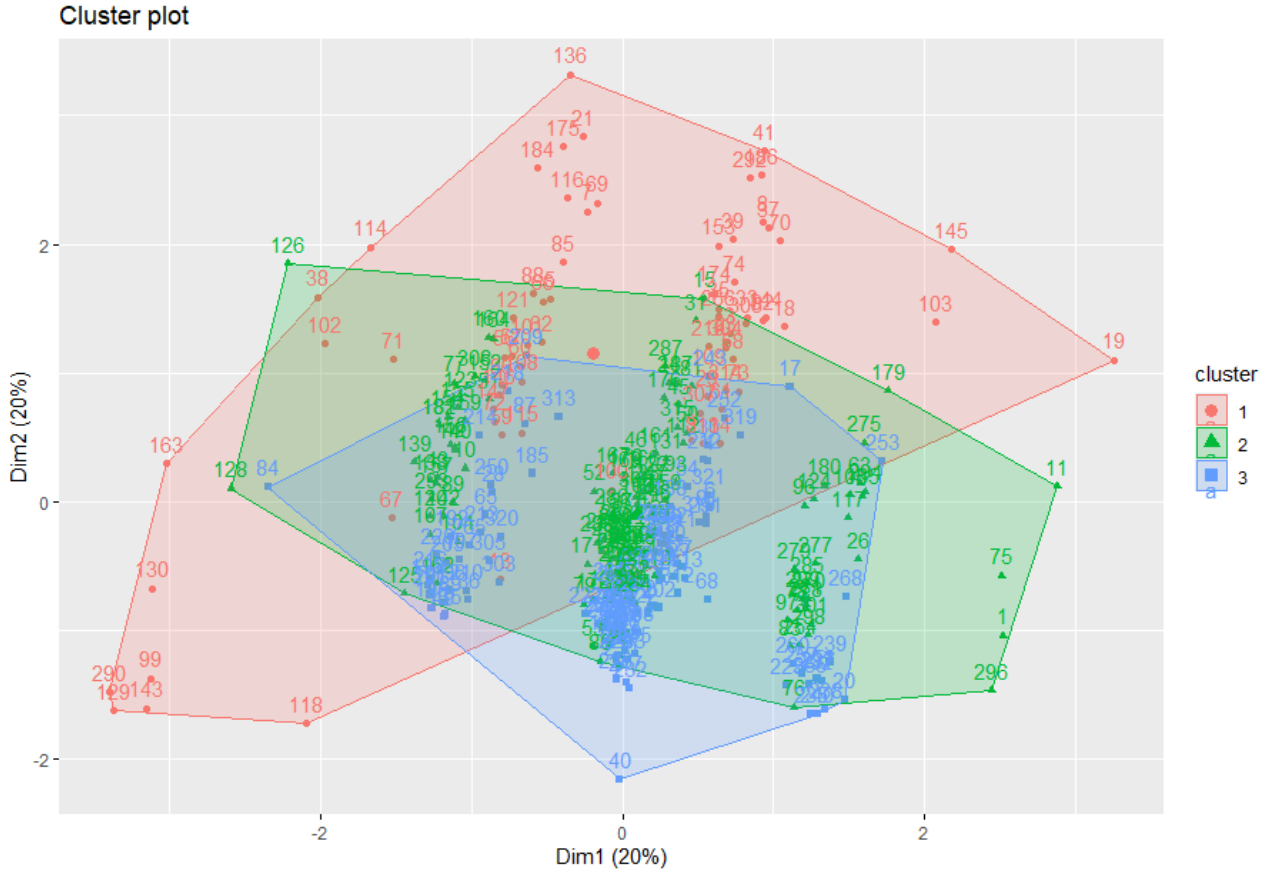


Figure 5. Claster Plot

Figure 5, is a graph of cluster results (K-Means) from the PCA data output, there are two main dimensions, namely Dim1 at 20% and Dim2 at 20% so the total contribution is 40% of the original data which indicates that this projection covers most of the important information , each point in the graph represents one individual (observation data). This data is divided into three clusters marked with different colors and symbols where cluster 1 is red, the circle symbol, represents individuals with certain characters. Cluster 2 is green, the symbol is a triangle, represents a group with a different pattern from cluster 1. Cluster 3 (blue, square), the last group has a different character from the other two clusters. Each cluster is surrounded by colored polygons which describe the coverage area of the cluster members in two-dimensional space. From the picture, it looks like there is an overlap between clusters, especially between cluster 2 (green) and cluster 3 (blue), this shows that several points/individuals in both clusters have similar characteristics, making it difficult to separate them clearly. Cluster 1 (red) tends to be more separated, especially in the top right and bottom left areas. Individual points that are located on the edge or far from the center of the cluster are outliers or in other words members with the most extreme characteristics compared to other members in the same cluster. The data center of each cluster must be projected from five-dimensional space (PC1-PC5) into two-dimensional space (Dim1-Dim2). From the data and graphs we can assume that if Dim1 = PC1, and Dim2 = PC2, then the center of cluster 1 is at coordinates (-1.156697, 0.034590) to the left of the graph, slightly above the

JIRK
Journal of Innovation Research and Knowledge
Vol.6 No.6, Nopember 2026
https://doi.org/10.53625/jirk.v6i6.13489

X axis. Cluster 2's center is at coordinates (0.767432 , -0.982774) at the bottom of the graph on the right side. and cluster 3 at coordinates (0.951895, 0.700631) is located on the top right side of the graph, so that the cluster distribution pattern matches the characteristics of the data center (centroid) described.

3. Ouput Deep Learning (Neural Network)

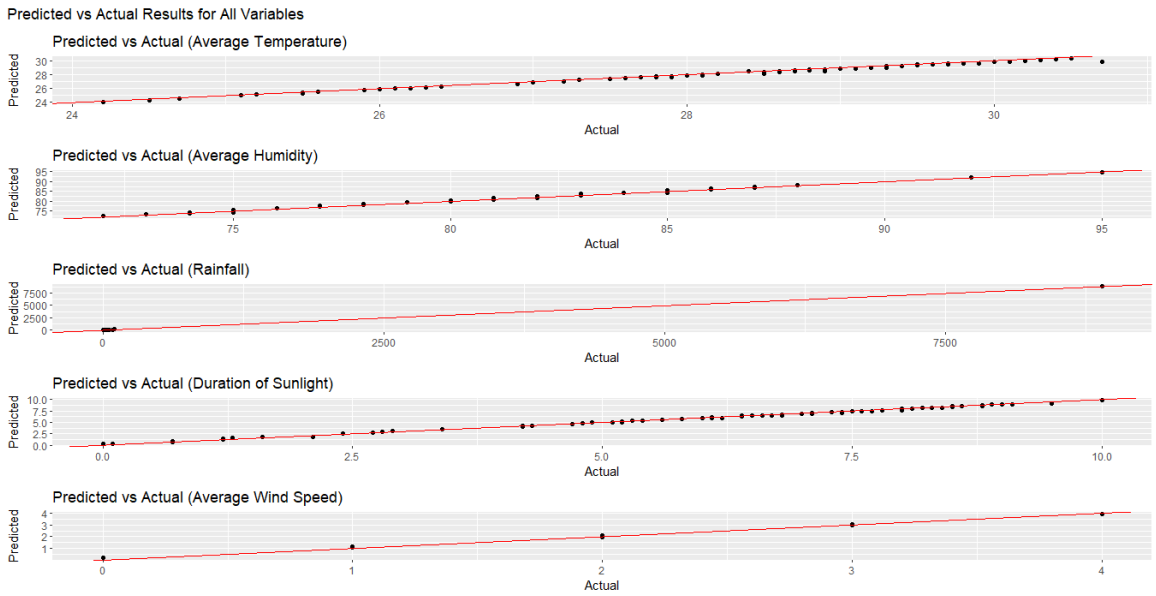


Figure 6. Predicted vs Actual Graph for All Variables

Figure 6 is a Predicted Vs Actual graph for all weather variables based on a Neural Network model that uses data from the results of the Cluster PCA method. From this figure it can be seen that: 1) Average temperature has predictions almost parallel to the red diagnostic line (y = x), p. This shows that the model can predict average temperature quite accurately; 2) Average Humidity for most of the prediction data is near the red diagonal line, but there is a slight spread in certain areas, indicating that humidity accuracy is quite good; 3) Rainfall, there are several outliers in the rain data, but even though in general the pattern is close to a diagonal line, this model has difficulty predicting rain at extreme values; 4) Duartion of Sunlight, very close to the diagonal line, shows good accuracy in predicting the duration of sunlight; 5) Average Wind Speed, the graph shows almost perfect predictions where all points are exactly on the diagonal line, this indicates the model is very reliable for predicting average wind speed.

The performance of this Netural Network Model provides good results for most weather variables, namely temperature, humidity, sunlight and wind speed variables have a high level of accuracy, while Rainfall is a variable that is more difficult to predict well, due to the different rain patterns. more volatile or non-linear. Using Cluster PCA, this method helps simplify data and reduce dimensions, thereby increasing model performance. Using a Neural Network, this algorithm detects patterns more effectively, which can detect non-linear patterns, suitable for complex weather data, thus providing good overall prediction resolution, the best performance is seen in the sunlight and wind speed variables, while the

JIRK
Journal of Innovation Research and Knowledge
Vol.6 No.6, Nopember 2026
https://doi.org/10.53625/jirk.v6i6.13489

accuracy in the rainfall variable still needs to be improved.

Weather Prediction for the next 30 days

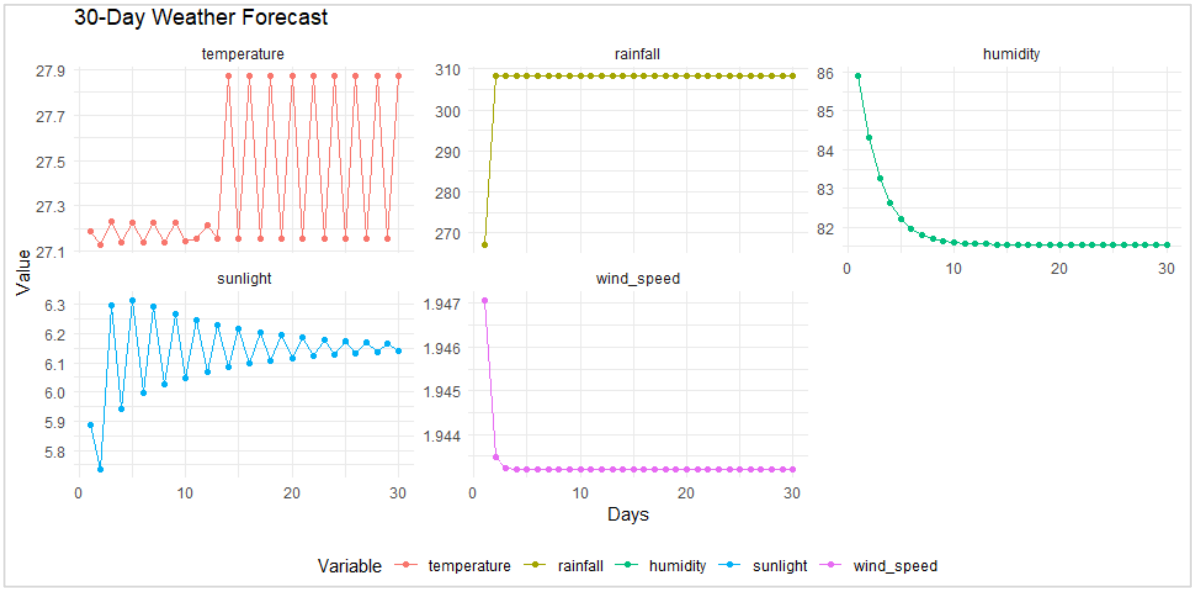


Figure 7. Forecast Weather for 30 Days

Figure 7 is a weather forecast for 30 days starting from November 20 to December 19 2024, this graphic is a combination of all weather variables, in full this forecast graphic can be interpreted as follows: 1) Temperature, has a consistent fluctuating pattern within the range ranges from 27.10C to 27.90C, representing a regular daily pattern of rising and falling temperatures, this states that the temperature looks stable with a consistent daily pattern repetitive,; 2) Rainfall, has a rain pattern that remains almost constant at around 310 mm throughout 30 days. From the graph it can be seen that there is no significant change in rainfall predictions, indicating the limitations of the data in capturing fluctuations in rainfall; 3) Humidity, the humidity pattern shows a drastic decrease at the beginning from around 86% to 82% and then stabilizes at the end of the month. The decrease in humidity at the beginning reflects the weather changes that occur, then after that the humidity appears constant; 4) Sunlight, the pattern of sunlight duration increases gradually during the first few days from 5.8 hours to around 6.3 hours with small daily fluctuations. This prediction indicates that sunlight increases slowly, with brighter days occurring in the middle of the month; 5) Wind Speed, the wind speed pattern experienced a sharp decline at the beginning from 1,947 m/s to 1,944 m/s. The stability of wind speed after the initial decrease shows relatively calm weather in the next 30 days, namely 20 November to 19 December 2024. The graph shows a relatively stable pattern for all variables except humidity which experienced a significant decrease at the beginning.

Iterate a linear regression model from the neural results with 5 predictors and use 10fold-cross-validation to train the model on 9 sections, and test it on the remaining sections then repeat this process 10 times, with the sample size in each fold varying from 26-28

samples. because the data is taken randomly, this is done to reduce the risk of overfitting. Obtained Root Mean Square Error (RMSE) of 0.5235 (52.35%). RMSE measures the average prediction error in the same units as the target variable. This value obtained shows that the model predictions are quite accurate, but there are still moderate deviations. R-squared measures the proportion of variables in the target data that are explained by the model of 0.5578 (55.78%) indicating that around 56% of the variability in the target variable is explained by the predictors used in the model. Mean Absolute Error (MAE) gives an average absolute error in prediction of 0.3963, indicating that the average prediction error is around 0.396 units which is smaller than RMSE.

PCA analysis succeeded in identifying the main components that explain most of the variables in the weather data. These main components provide a clear interpretation of the relationship between weather variables. K-Means succeeded in grouping data into three clusters with different weather characteristics, each cluster has a unit which is reflected in its centroid value. Neural Network modeling trained on grouped data shows good performance in predicting weather variables, especially for temperature, humidity, sunshine duration and wind speed. Rainfall predictions have challenges because they are more volatile.

Implications and Potential Benefits

PCA has succeeded in reducing data dimensions and simplifying complex data structures making it easier to visualize and interpret data. Clustering effectiveness, based on PCA results, allows neural network models to capture more specific patterns in each group of data, thereby increasing prediction accuracy. The performance of the neural network model has demonstrated its ability to capture complex non-linear patterns in weather data, especially for variables that have non-linear relationships. Challenges in research to predict Rainfall is a variable that is difficult to predict because it is influenced by many factors and has a high level of variability. Suggestions for further research: 1) sensitivity analysis of the number of clusters and parameters is needed; 2) Incorporation of Spatial Data, intended for Location data, spatial analysis can be carried out to identify spatial patterns in weather data and increase prediction accuracy; 3) It is necessary to compare the performance of the neural network model with other models such as support vector machines (SVM) or decision-based models. 4) This research has not provided residual analysis of the model to identify prediction error patterns and improve the model, so it is hoped that further research can complete this analysis; 5) This research only predicts 30 days, it would be better if predictions were made for the long term.

The potential of using this hybrid analysis (PCA, Cluster, Machine Learning) can be used in wider applications such as for: 1) Energy, it can help in planning renewable energy production; 2) Agriculture, this analysis can help farmers in planning planting and harvesting; 3) Transportation, this analysis can help in planning transportation routes and schedules; 4) Disaster Management. This is to help with early warning of natural disasters such as floods and droughts.

CONCLUSION

This research successfully shows that integration between PCA and clustering techniques, and neural networks, is able to provide an effective solution for high-dimensional

data analysis, producing accurate and reliable prediction models. This research can be the basis for developing a better weather early warning system and supporting more effective decision making in various sectors.

ACKNOWLEDGMENTS

I am grateful to the Education Fund Management Institute (LPDP), for the scholarship I received. I would also like to express my thanks to the UNMUS campus for giving me the opportunity to continue my doctoral studies, and I would also like to express my thanks to the Professors and Lecturers at Yogyakarta State University who have guided me in my studies.

REFERENCE

Abdi, H. and Williams, L.J. (2010) 'Principal component analysis', *WIREs Computational Statistics*, 2(4). Available at: https://doi.org/10.1002/wics.101.

Ba, A. *et al.* (2023) 'Power optimization of a photovoltaic system with artificial intelligence algorithms over two seasons in tropical area', *MethodsX*, 10. Available at: https://doi.org/10.1016/j.mex.2022.101959.

Balsa, C. *et al.* (2023) 'Reconstruction of Meteorological Records by Methods Based on Dimension Reduction of the Predictor Dataset', *Computation*, 11(5), p. 98. Available at: https://doi.org/10.3390/computation11050098.

Duarte, B. *et al.* (2022) 'Elemental Chemometrics as Tools to Depict Stalked Barnacle (Pollicipes Pollicipes) Harvest Locations and Food Safety', *Molecules*, 27(4), p. 1298. Available at: https://doi.org/10.3390/molecules27041298.

Ernst, D. and Dolničar, S. (2017) 'How to Avoid Random Market Segmentation Solutions', *Journal of Travel Research*, 57(1). Available at: https://doi.org/10.1177/0047287516684978.

Geng, D.Y., Zhang, H. and Wu, H. (2020) 'Short-Term Wind Speed Prediction Based on Principal Component Analysis and LSTM', *Applied Sciences*, 10(13), p. 4416. Available at: https://doi.org/10.3390/app10134416.

Gunawan, G. (2024) 'Disaster Event, Preparedness, and Response in Indonesian Coastal Areas: Data Mining of Official Statistics', *International Journal of Computing and Digital Systems*, 15(1), pp. 249–264. Available at: https://doi.org/10.12785/ijcds/160120.

Habibi, M. and Cahyo, P.W. (2019) 'Clustering User Characteristics Based on the influence of Hashtags on the Instagram Platform', *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 13(4). Available at: https://doi.org/10.22146/ijccs.50574.

Karamizadeh, S. *et al.* (2013) 'An Overview of Principal Component Analysis', *Journal of Signal and Information Processing*, 4(3), pp. 173–175. Available at: https://doi.org/10.4236/jsip.2013.43B031.

Kashyap, A. *et al.* (2020) 'K- Mean and Euclidian Cluster Analysis for Salt Tolerance Rice Genotypes under Alkaline Soil Condition', *International Journal of Current Microbiology and Applied Sciences*, 9(11). Available at: https://doi.org/10.20546/ijcmas.2020.911.043.

Khairati, F. and Putra, H. (2022) 'Prediksi Kuantitas Penggunaan Obat Pada Layanan

Kesehatan Menggunakan Algoritma Backpropagation Neural Network', *Jurnal Sistim Informasi Dan Teknologi*, pp. 128–135. Available at: https://doi.org/10.37034/jsisfotek.v4i3.158.

Lèbre, S. *et al.* (2010) 'Statistical inference of the time-varying structure of gene-regulation networks', *BMC Systems Biology*, 4(1). Available at: https://doi.org/10.1186/1752-0509-4-130.

Lee, S.-H. *et al.* (2023) 'Label-free Atherosclerosis Diagnosis Through a Blood Drop of Apolipoprotein E Knockout Mouse Model Using Surface-enhanced Raman Spectroscopy Validated by Machine Learning Algorithm', *Bioengineering & Translational Medicine*, 8(4). Available at: https://doi.org/10.1002/btm2.10529.

Malinowski, M., Rohrbach, M. and Fritz, M. (2017) 'Ask Your Neurons: A Deep Learning Approach to Visual Question Answering', *International Journal of Computer Vision*, 125(1–3). Available at: https://doi.org/10.1007/s11263-017-1038-2.

Mathivanan, N.M.N., Ghani, N.A.M. and Janor, R.M. (2019) 'A comparative study on dimensionality reduction between principal component analysis and k-means clustering', *Indonesian Journal of Electrical Engineering and Computer Science*, 16(2). Available at: https://doi.org/10.11591/ijeecs.v16.i2.pp752-758.

Matsushima, A. *et al.* (2017) 'Principal component analysis for ataxic gait using a triaxial accelerometer', *Journal of NeuroEngineering and Rehabilitation*, 14(1). Available at: https://doi.org/10.1186/s12984-017-0249-7.

Muhariya, A., Riadi, I. and Prayudi, Y. (2022) 'Cyberbullying Analysis on Instagram Using K-Means Clustering', *JUITA : Jurnal Informatika*, 10(2). Available at: https://doi.org/10.30595/juita.v10i2.14490.

Pokharel, M., Bhatta, J. and Paudel, N. (2021) 'Comparative Analysis of K-Means and Enhanced K-Means Algorithms for Clustering', *NUTA Journal*, 8(1–2). Available at: https://doi.org/10.3126/nutaj.v8i1-2.44044.

Slijepčević, D. *et al.* (2018) 'Automatic Classification of Functional Gait Disorders', *IEEE Journal of Biomedical and Health Informatics*, 22(5). Available at: https://doi.org/10.1109/jbhi.2017.2785682.

Soares, D.P. *et al.* (2016) 'Principal component analysis in ground reaction forces and center of pressure gait waveforms of people with transfemoral amputation', *Prosthetics &Amp; Orthotics International*, 40(6). Available at: https://doi.org/10.1177/0309364615612634.

Surono, S. *et al.* (2023) 'Developing an optimized recurrent neural network model for air quality prediction using K-means clustering and PCA dimension reduction', *International Journal of Innovative Research and Scientific Studies*, 6(2). Available at: https://doi.org/10.53894/ijirss.v6i2.1427.

Vasileva, Y., Nevedrov, A.V. and Subbotin, S.P. (2020) 'Development and Performance Evaluation of a Computer Program Based on Neural Network Mathematical Models for Forecasting By- Product Yield', *E3S Web of Conferences*, 174. Available at: https://doi.org/10.1051/e3sconf/202017403023.

Venkateswarlu, B. and Raju, G. (2013) 'Mine Blood Donors Information through Improved K-Means Clustering', *International Journal of Computational Science and Information*

Technology, 1(3). Available at: https://doi.org/10.5121/ijcsity.2013.1302.

Yoseph, F. *et al.* (2020) 'The impact of big data market segmentation using data mining and clustering techniques', *Journal of Intelligent &Amp; Fuzzy Systems*, 38(5). Available at: https://doi.org/10.3233/jifs-179698.

Yu, T. *et al.* (2022) 'A New Strategy for Microbial Taxonomic Identification Through Micro-Biosynthetic Gold Nanoparticles and Machine Learning', *Advanced Materials*, 34(11). Available at: https://doi.org/10.1002/adma.202109365.

Yuan, R. (2021) 'An improved K-means clustering algorithm for global earthquake catalogs and earthquake magnitude prediction', *Journal of Seismology*, 25(3). Available at: https://doi.org/10.1007/s10950-021-09999-8.

Zhou, B. *et al.* (2020) 'A Hybrid Method for Ultrashort-Term Wind Power Prediction Considering Meteorological Features and Seasonal Information', *Mathematical Problems in Engineering*, 2020, pp. 1–12. Available at: https://doi.org/10.1155/2020/1795486.

https://doi.org/10.53625/jirk.v6i6.13489

HALAMAN INI SENGAJA DIKOSONGKAN

ISSN 2798-3471 (Cetak)
ISSN 2798-3641 (Online)