JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 3, July 2026, pp. 926-944

Handling Missing Values using Weighted Linear Combination of KNN-SVD: A Case Study of Rainfall Data in West Java

Rizkian Agung Jamaesa[1], Sri Nurdiati[2*], Elis Khatizah[2], Mohamad Khoirun Najib[2], Lilis Sri Wahyuni[1]
[1]Magister of Applied Mathematics, School of Science Data, Mathematics, and Informatics, IPB University, Indonesia
[2]Division of Computational Mathematics, School of Science Data, Mathematics and Informatics, IPB University, Indonesia
nurdiati@apps.ipb.ac.id

Article History:
Received : 25-11-2025
Revised : 01-01-2026
Accepted : 03-01-2026
Online : 01-07-2026

Keywords:
Imputation;
Missing Value;
Integration;
Rainfall Dataset.

ABSTRACT

This study is an experimental and comparative quantitative research that evaluates missing value imputation methods for daily rainfall data in West Java. Rainfall data are crucial for environmental policies, particularly in flood control and water resource management. Daily rainfall records from five BMKG stations in West Java were used in this study. Although these stations provide accurate data through direct measurement, missing values often occur due to human error or equipment problems. To solve this, we introduce an integrated imputation method that combines K-Nearest Neighbors (KNN) and Singular Value Decomposition (SVD) with a Weighted Linear Combination (WLC) approach. This method represents a significant improvement over the single-model imputation methods employed in earlier research. We split the dataset into training and testing sets using five different ratios (95:5%, 90:10%, 80:20%, 70:30%, and 64:40%) to test the model's performance. We measured effectiveness using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The results show that the combined KNN–SVD method outperforms KNN or SVD alone in all cases. The best results were obtained from the 95:5% split, with the lowest MAE and RMSE values of 7.35 and 13.22, respectively. These results suggest that the integrated KNN–SVD imputation model enhances the reliability of rainfall datasets, thereby improving climate information for hydrological studies, disaster risk reduction, and policy-making in West Java.

https://doi.org/10.31764/jtam.v10i3.36708

A. INTRODUCTION

Rainfall is an essential factor that supports life. Precipitation refers to the amount of rainwater that falls to the Earth's surface over a specific period. In conducting weather studies, rainfall data are one of the key components required (Azman et al., 2021). Excessive rainfall events, which have become more frequent due to climate change, are difficult to predict and can harm communities (Rafhida et al., 2024), one primary source of rainfall data is the BMKG observation stations, which are considered more reliable than satellite data because they represent actual ground measurements (Sriwahyuni et al., 2025).

However, rainfall data recorded at stations often face the issue of missing values. Missing values are a common problem in rainfall datasets (Duarte et al., 2022; Kim et al., 2019) and are considered critical to data quality. When the proportion of missing values exceeds 15%, it can significantly distort data interpretation and reduce the reliability of subsequent analyses

(Odhiambo, 2020). The quality of analytical results obtained from complete datasets differs from those derived from datasets containing missing values (Mohammed et al., 2021). Several factors contribute to missing values, including human error and machine malfunction (Agrawal, 2023). Therefore, appropriate handling of missing values is necessary.

Deletion and imputation are standard approaches for handling missing values. Deletion methods, although simple to implement, may introduce bias and lead to the loss of valuable information, particularly when the proportion of missing data is high (Shantal et al., 2023). As a result, imputation techniques are more commonly adopted to preserve data structure and minimize analytical bias. Imputation methods have been widely applied across various types of datasets, including medical datasets (Liu et al., 2020), heart disease data (Moatadid et al., 2023), electric vehicle charging data (Lee et al., 2020), weather data (Nida et al., 2023). Given the critical role of data completeness in climate analysis, selecting an appropriate imputation method is a crucial step in improving rainfall data quality.

Classical techniques such as K-Nearest Neighbors (KNN) (Fadlil et al., 2022; Hariyanto, 2020; Saeipourdizaj et al., 2021) and Singular Value Decomposition (SVD) (Arciniegas-Alarcón et al., 2020; Brand, 2002; Yuan et al., 2019) are among the most commonly used methods for missing value imputation. KNN is recognized for its simplicity and effectiveness in capturing local data patterns, while SVD is capable of exploiting the global structure of the dataset through matrix decomposition. However, each method has inherent limitations when applied independently, particularly in complex climate dataset with spatial and temporal variability.

Research on data imputation methods has been extensively conducted using various approaches. Several techniques, such as Mean Imputation (MI), Predictive Mean Matching (PMM), Sample Imputation (SI), and K-Nearest Neighbors (KNN), have been applied for data imputation; however, results indicate that the KNN method outperforms other techniques (MI, PMM, SI) for climate-related datasets (Nida et al., 2023). A predictive model for assessing stroke risk factors in elderly patients has also been developed using Singular Value Decomposition (SVD) for missing value imputation, which has proven effective in improving prediction accuracy (Zhang et al., 2022). Nevertheless, most existing studies focus on applying individual imputation methods or integrated approaches in non-regional or non-climate-specific contexts, with limited emphasis on regional rainfall datasets characterized by spatial heterogeneity.

While single-method imputations are commonly used, integrated imputation approaches have the potential to yield better results. Furthermore, integrated approaches such as the KNN–MCF method have been shown to produce superior imputation performance compared to standalone KNN (Sanjar et al., 2020). In the context of regional rainfall data in West Java, studies that systematically evaluate and integrate multiple imputation techniques remain limited. Specifically, the potential of combining KNN and SVD to leverage both local similarity and global data structure for rainfall data imputation has not been sufficiently explored.

Therefore, this study aims to evaluate and compare the performance of KNN, SVD, and an integrated KNN-SVD approach for handling missing values in the rainfall dataset from West Java. The proposed integration is implemented using a weighted linear combination scheme to balance local similarity and global structural information contained in the rainfall dataset. This research regularly examines the effect of key parameters, including the number of neighbors in KNN, the rank in SVD, and the weighting proportion in the integration model, on imputation

accuracy. By improving the completeness and reliability of rainfall datasets, this study is expected to support more robust hydrological analysis, climate-related studies, and disaster risk management in West Java.

B. METHODS

1. Study Area and Dataset

This study adopts an experimental and comparative quantitative research design to evaluate the performance of different missing value imputation methods on daily rainfall data. To provide a clear overview of the research methodology, this study was conducted through several sequential stages. First, daily rainfall data were collected from five BMKG observation stations in West Java for the period 2015–2024. Second, missing values were identified and partially simulated to enable controlled evaluation. Third, the dataset was divided into training and testing subsets with various proportions. Fourth, missing value imputation was performed using KNN, SVD, and the integrated KNN–SVD method based on the Weighted Linear Combination approach. Finally, the imputation performance was evaluated using MAE and RMSE across different stations and data proportions.

This study focuses on West Java Province, which has the largest population in Indonesia and the highest exposure to extreme weather disaster risks according to the Indonesian Disaster Risk Index (RBI). West Java features diverse topography, consisting of hilly areas in the central region, mountainous zones in the south and east, and lowlands in the north. The province has a humid tropical climate with relatively stable temperatures and high annual rainfall ranging from 2,000 to 4,000 mm (Purnamasari et al., 2021). The data used in this study were obtained from the official BMKG (Meteorology, Climatology, and Geophysical Agency) website and consist of daily rainfall data (mm/day) recorded by BMKG observation stations. The rainfall data are categorized into four intensity levels as defined by BMKG (2021), as shown in Table 1.

Table 1. Classification of daily rainfall intensity

Category	intensity (mm)
Low rainfall	0-50
Moderate rainfall	50-100
High rainfall	100-300
Very high rainfall	>300

Table 1 categorizes daily rainfall intensity into four levels (low, moderate, high, and very high) based on the corresponding millimeter ranges. Rainfall observation stations managed by BMKG play a crucial role in providing climate data that serve as the foundation for various climatological and hydrological analyses (Radjab et al., 2020). The data produced by these stations generally consist of daily rainfall measurements in millimeters, which are utilized for weather forecasting, climate pattern analysis, and hydrometeorological disaster mitigation. There are five BMKG observation stations in West Java, each exhibiting varying proportions of missing values, as shown in Table 2.

Table 2. Missing value of rainfall data at BMKG stations in West Java (January 2015-June 2024)

Station Name	Filled Percentage	Missing Percentage
Geofisika Bandung	80.05%	19.95%
Meteorologi Kertajati	88.90%	11.10%
Meteorologi Penggung	64.28%	35.72%
Klimatologi Jawa Barat	62.35%	37.65%
Meteorologi Citeko	85.82%	14.18%

Table 2 shows that the highest percentage of missing values occurs at the West Java Climatology Station, amounting to 37.65%, while the lowest percentage is found at the Kertajati Meteorological Station, with 11.10%. The average proportion of missing values across all rainfall stations in West Java is 23.72%. These results indicate that missing values in rainfall data across West Java are significant and non-negligible. To evaluate the imputation performance in a controlled manner, the original complete observations were partially masked to simulate missing values. The masking process was conducted randomly following the original missing value distribution of each station. The dataset was then divided into training and testing subsets using five different proportions, namely 95:5, 90:10, 80:20, 70:30, and 60:40. Each experimental configuration was repeated 100 times to reduce the randomness effect and ensure result stability.

2. K- Nearest Neighbour (KNN)

The K-Nearest Neighbors (KNN) method is a classical technique commonly used for data imputation. KNN is an algorithm that classifies an object based on the closest distance between that object and its neighboring data points (Pertiwi et al., 2020). In rainfall data analysis, each observation station is represented as a vector containing rainfall observations over a specific time period. If there are n time periods and m stations, then the entire dataset can be represented in the following matrix form.

$$\boldsymbol{A} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(m)} \end{bmatrix} \in \mathbb{R}^{n \times m} \quad (1)$$

The matrix $\boldsymbol{A}$ serves as the main object of analysis, containing missing values that will be imputed, where $x_i^{(h)}$ denotes the rainfall data at the h-th station and at the i-th time period. In this study, a norm vector is defined as a function $\|\cdot\|$ that maps $\boldsymbol{x} \in \mathbb{R}^n$ to a non-negative real number and satisfies the following fundamental properties:

a. $\|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = 0$ (*positif definiteness*),
b. $\|\alpha\boldsymbol{x}\| = |\alpha| \cdot \|\boldsymbol{x}\|$ for every *scalar* α (*homogenitas*),
c. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ (triangle inequality).

The norm serves to measure the "magnitude" or "length" of a data vector, allowing it to be used in analysis and information processing. In the context of this study, the Euclidean norm is formulated as follows (Boyd dan Vandenberghe 2017):

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}, \tag{2}$$

In this study, the Euclidean distance was chosen as the proximity measure in the KNN method due to several advantages that align with the characteristics of rainfall data. First, the Euclidean distance is the most common and straightforward metric for measuring the closeness between points in a multidimensional space, making its interpretation more intuitive compared to other distance metrics (Steinbach & Tan, 2009). Second, the Euclidean distance is sensitive to differences in values across dimensions, allowing it to capture both spatial and temporal variations present in rainfall data. This is particularly important since the rainfall distribution across stations tends to have a homogeneous scale after normalization, ensuring that the fundamental assumption of the Euclidean distance remains valid (Ha et al., 2011). Third, the use of the Euclidean distance is supported by previous studies demonstrating its more stable performance compared to other metrics, such as Manhattan or Minkowski distance, in the context of hydrological data imputation (Junninen dkk., 2004). Therefore, the selection of the Euclidean distance is not only practical for implementation but also grounded in theoretical and empirical justification. The Euclidean distance formula is expressed as follows (Syauqi et al., 2023):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

In the context of rainfall data, this distance represents the similarity of temporal rainfall patterns between observation stations, allowing missing values at a given station to be estimated based on a station with similar rainfall behavior. Rainfall data exhibit spatio-temporal variation, meaning that accurate estimation requires considering a larger number of neighboring points for comparison. The literature suggests that, for environmental data, the optimal value of k is often larger than that used in classical datasets. Mathematically, if $X(t, s)$denotes the rainfall value at time $t$and location s, then it can be expressed as:

$$Cov[X(t, s), X(t', s')] = f(\|t - t'\|, \|s - s'\|)\ f\ decreasing\ function, \tag{4}$$

The larger the value of k, the KNN algorithm can explore neighbors within a broader spatio-temporal neighborhood, allowing the estimation $\hat{x}_i$to approach the conditional expectation, expressed as $\hat{x}_i \to \mathbb{E}[x_i|neighborhood]$. In this study, the number of nearest neighbors k was varied from 2 to 70 to examine its influence on imputation accuracy. The optimal value of k was selected based on the minimum Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) obtained from repeated experiments. Prior to distance computation, rainfall data were normalized to ensure that all variables contributed equally to the Euclidean distance calculation. The k value that gives the smallest error value will be used in the KNN-SVD integration.

3. Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) performs numerical computations to represent and manipulate a matrix in a simpler form through matrix decomposition (Erichson et al., 2019). SVD performs numerical computations to represent and manipulate matrices in a simpler form through matrix decomposition (Erichson et al., 2019). It is one of the commonly used methods because SVD can reduce large-scale data without losing the main information contained within it (Septiawan dkk., 2019).Mathematically, SVD decomposes a matrix $\boldsymbol{A}$ of size $m \times n$ into three components: an orthogonal matrix ($\boldsymbol{U}$), a diagonal matrix ($\boldsymbol{\Sigma}$), and the transpose of an orthogonal matrix ($\boldsymbol{V}^T$). SVD is employed to capture hidden patterns within the data and to approximate the data through dimensionality reduction (Afshar & Usefi, 2021). Mathematically, the imputation process using the SVD model begins with the rainfall data $\boldsymbol{A} \in \mathbb{R}^{m \times n}$with entries a_{ij}, where some of the a_{ij}values are missing. During imputation, a rank-l approximation of a matrix $\boldsymbol{A}$ performed, where $l < s = \min(m, n)$, by minimizing the following optimization problem:

$$\min_{\boldsymbol{A}l} \|\boldsymbol{A} - \boldsymbol{A_k}\| \text{ subject to rank } (\mathbf{A}_l) \leq \boldsymbol{l}, \tag{5}$$

the best rank-l an approximation is obtained from the sum of the first $l$outer products of the left ($\boldsymbol{u}$) and right ($\boldsymbol{v}$) singular vectors, each scaled by their corresponding singular values ($\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_s$), as follows**:**

$$\boldsymbol{A}_l = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T + \ldots + \sigma_l \boldsymbol{u_l} \boldsymbol{v}_l^T. \tag{6}$$

The norm of the difference between the best approximation and the original matrix under the induced 2-norm condition is equal to the $(l+1)^{th}$ singular value of the matrix.

$$\widetilde{\boldsymbol{A}} = \boldsymbol{U}_l \boldsymbol{\Sigma}_l \boldsymbol{V}_l^T \approx A, \tag{7}$$

with $\boldsymbol{U}_l \in \mathbb{R}^{m \times l}$, $\boldsymbol{\Sigma}_l \in \mathbb{R}^{l \times l}$, and $\boldsymbol{V}_l \in \mathbb{R}^{n \times l}$, which represent the reduced decomposition of $\boldsymbol{A}$**.** The selection of the value of l, as well as the applied regularization or optimization technique, determines the SVD mode being tested. For rainfall datasets, the low-rank approximation captures dominant seasonal and spatial rainfall patterns while filtering noise caused by extreme events or recording errors. The rank parameter l in the SVD-based imputation was varied from 1 to 5 to analyze its effect on reconstruction accuracy. The optimal rank was determined by comparing MAE and RMSE values across all stations. The value of l that gives the smallest error value will be used in the KNN-SVD integration.

4. KNN-SVD Integration

The integration of the K-Nearest Neighbors (KNN) and Singular Value Decomposition (SVD) methods in this study is designed using the Weighted Linear Combination (WLC) approach, which is based on the need to produce more accurate and stable missing value imputations. Conceptually, the imputation results from each method are assigned specific weights according

to their relative contributions and then combined into a single final estimate. The mathematical formulation of the WLC integration is expressed as follows:

$$I_{int} = (1 - \alpha)I_{KNN} + \alpha \times I_{SVD}, \tag{8}$$

where I_{int} represents the integrated imputation result, I_{KNN} is the imputation result obtained using KNN, I_{SVD} is the imputation result obtained using SVD, and α is the weight that indicates the relative contribution of each method. The weighting parameter α in the Weighted Linear Combination, the value was varied between 0 and 1 wth an increment of 0,1 to examine the relative contribution of KNN and SVD. For each weight configuration, MAE and RMSE were computed to identify the optimal balance between local and global rainfall information. The experimental results indicate that different weight combinations influence MAE and RMSE differently, confirming the complementary nature of KNN and SVD in rainfall imputation.

5. Evaluation Metrics

After obtaining the imputation model to be used, it is necessary to measure the accuracy of the imputation results using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE represents the average absolute deviation, while RMSE is more sensitive to large errors; thus, both complement each other in assessing model performance. In this study, the Root Mean Squared Error (RMSE) is employed as the metric to evaluate the accuracy of the missing value imputation. RMSE measures the difference between the observed value (X_{obs}) and the imputed value (X_{input}) for the i-th vector, where the elements represent the original and imputed values. Each (i) corresponds to the number of missing values used (Cihan & Ozger, 2019).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i^{obs} - X_i^{input})^2}{n}} \tag{9}$$

Mean Absolute Error (MAE) is one of the evaluation metrics used to measure the magnitude of prediction errors. MAE calculates the average absolute difference between the predicted values and the actual values. MAE can be defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|X_i^{obs} - X_i^{input}\right| \tag{10}$$

The formula above shows that MAE is calculated by summing all the absolute differences between the observed data (X_i^{obs}) and the imputed data (X_i^{input}) then dividing by the total number of data points. Thus, MAE provides an average measure of the magnitude of errors regardless of the direction of deviation, making the resulting value easier to interpret.

C. RESULT AND DISCUSSION

1. Pre-Processing of Datasets

Preprocessing is an essential stage in data processing, as the quality of the data used greatly influences the outcomes of the analysis and modeling. In studies involving rainfall data, this process becomes crucial since the data often contain missing values, outliers, or recording errors (Duarte et al., 2022). Through proper preprocessing, the data can be transformed to become more consistent, clean, and ready for imputation model development. Moreover, this stage helps improve the accuracy and reliability of the models developed in the research (García et al., 2016). The rainfall data obtained were categorized as 1 for available data and 0 for missing values. The prepared data, consisting only of available and missing entries, were processed using MATLAB, resulting in an average missing value proportion of 23.72% across all stations.

2. Developing the KNN Model

The preprocessed rainfall data were then processed using MATLAB software. First, the imputation model began by determining the range of k values to be used. Second, the percentage of random samples to be utilized was specified. Third, imputation was performed using the built-in KNN package in MATLAB. Fourth, the station to be analyzed for Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) was selected, and the imputation process was repeated 100 times to obtain the smallest MAE and RMSE values. In evaluating the performance of the k-Nearest Neighbors (KNN)-based imputation model, a crucial aspect lies in the selection of the number of nearest neighbors (k), as this parameter directly affects the level of prediction accuracy.

However, there is no definitive method for determining the optimal value of k, a k value that is too small may lead to overfitting, while a value that is too large risks underfitting. Therefore, in this study, several k values were tested using performance indicators such as MAE and RMSE. These two metrics are widely used in the literature because they provide a comprehensive overview of the average error and are more sensitive to extreme values (Willmott & Matsuura, 2005). According to Nti et al. (2021), in the KNN model, varying the value of k aims to achieve a balance between variance and bias. Furthermore, the effectiveness of MAE and RMSE as evaluation measures has been demonstrated in research involving time series and environmental data imputation, making them highly relevant for rainfall data analysis (Farhangfar et al., 2007). The MAE and RMSE values for different k values are shown in Figure 1.

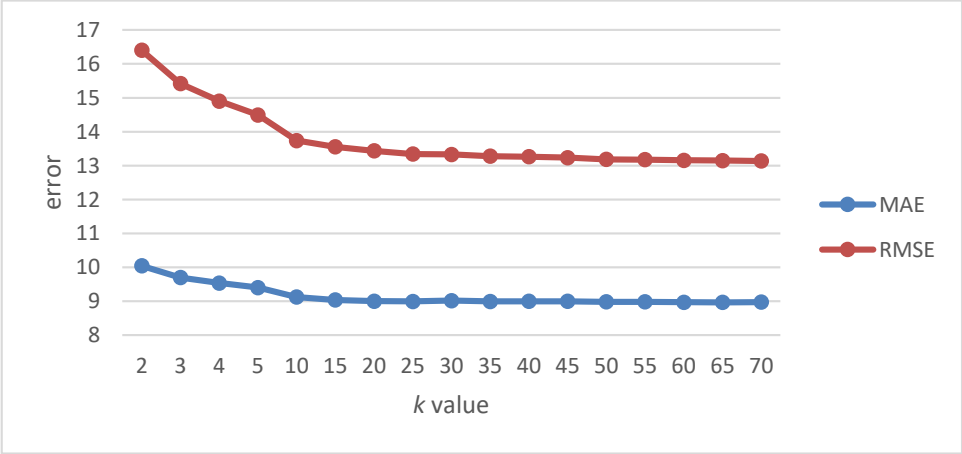


Figure 1. MAE error values of the KNN model

Figure 1 illustrates a decrease in error values (MAE and RMSE) as the k value increases in the KNN model. This pattern indicates that increasing the number of neighbors initially enhances the stability of the estimation by reducing prediction variability. However, once k reaches approximately 10, both MAE and RMSE tend to converge and show no significant improvement. The value $k = 10$ signifies that the model has achieved an optimal balance between complexity and generalization.

The decreasing trend in MAE and RMSE observed for increasing values of k indicates that incorporating a larger number of neighboring observations enhances the robustness of the imputation by reducing the variance associated with localized rainfall anomalies. This behavior aligns with the theoretical bias-variance trade-off inherent in nonparametric learning methods, where increasing neighborhood size mitigates variance at the expense of increased bias (Hastie et al., 2006). The convergence of both MAE and RMSE beyond $k = 10$ suggest that the KNN model has reached a stable operating regime in which further enlargement of the neighborhood yields diminishing marginal gains. From a statistical standpoint, this plateau implies that additional neighbors contribute redundant information rather than improving predictive accuracy. Similar convergence patterns have been consistently reported in hydrological time series and precipitation imputation studies, where moderate k values are sufficient to capture dominant spatial and temporal dependencies without over-smoothing extreme rainfall events (Xie et al., 2024).

In the context of rainfall data in West Java, which are characterized by pronounced spatial heterogeneity and episodic extreme precipitation, selecting an excessively large k may obscure localized raingall dynamics that are climatologically meaningful. Conversely, very small k values risk overemphasizing noise arising from measurement errors or short-term variability. The selection of $k = 10$ therefore, it represents an optimal compromise between preserving local rainfall characteristics and ensuring numerical stability of the imputation process. This finding reinforces the suitability of the chosen k value for subsequent analyses and supports its use as a reliable configuration for KNN-based computation of rainfall data in regions with complex hydrological behavior (Xie et al., 2024).

3. Developing the SVD Model

The prepared rainfall data were processed using MATLAB software. First, the imputation model was initiated by determining the range of l values to be used. Second, the percentage of random samples was specified. Third, imputation was performed using the SVD package available in MATLAB. Fourth, the station to be displayed for the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) results was selected, and the imputation process was repeated 100 times to obtain the smallest MAE and RMSE values. In this dataset, the singular values $l = 1, 2, 3, 4,$ and 5 were derived from a matrix of size $min(3496, 5)$. The performance analysis of the SVD model in data imputation is highly influenced by the number of singular components used. Selecting an appropriate number of components is crucial, as it plays a key role in capturing the dominant structure of the data (Halko et al., 2011). In addition to their sensitivity to prediction errors, MAE and RMSE are employed as evaluation metrics, particularly for datasets prone to variability (Junninen et al., 2004; Willmott & Matsuura, 2005). The MAE and RMSE values based on l are presented in Table 3.

Table 3. SVD model error values

Station	MAE					RMSE				
	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
Bandung	6.51	6.51	6.84	7.11	6.38	11.27	11.16	11.98	12.41	12.52
Citeko	8.06	8.23	9.20	9.69	8.13	13.51	13.96	15.84	17.20	15.41
Jabar	15.06	16.94	17.30	13.89	13.78	22.71	26.05	26.68	22.67	22.85
Kertajati	8.54	8.71	7.70	8.68	6.55	15.12	17.19	15.39	17.55	15.58
Penggung	8.86	8.36	8.73	8.59	6.66	14.99	16.42	16.91	17.09	15.91
Average	9.41	9.75	9.95	9.59	**8.30**	**15.52**	16.96	17.36	17.39	16.45

Table 3 presents the evaluation results of varying l values $(1, 2, 3, 4, 5)$ across the five rainfall stations. The evaluation was conducted by calculating the MAE and RMSE values from the imputation results. Based on the average values across all stations, the smallest MAE was obtained at $l = 5$ with a value of 8.30, while the smallest RMSE occurred at $l = 1$ with a value of 15.52. The divergence between the optimal values of l obtained using MAE and RMSE reflects the fundamentally different error sensitivities of these metrics. MAE measures the average magnitude of errors and is less influenced by extreme deviations, whereas RMMSE favors penalizing larger errors more heavily by squaring the residual (Willmott & Matsuura, 2005). As a result, RMSE favors simpler models that suppress extreme fluctuations, while MAE tends to reward models that preserve broader structural patterns. The selection of $l = 5$ over $l = 1$ is justified by its superior ability to retain more information from the rainfall data, resulting in more representative imputation outcomes. At $l = 1$ only the dominant singular value is retained, capturing the largest variance component of the rainfall data. While this configuration minimizes RMSE, it substantially reduces station variability and suppresses localized rainfall patterns, resulting in a higher average absolute error. Such oversimplification is particularly problematic for rainfall datasets, which are inherently heterogeneous and influenced by localized climatic and topographic factors.

In contrast, at $l = 5$ allows the model to capture a richer representation of spatial and temporal variability without introducing excessive noise. The first few singular values capture the dominant patterns more comprehensively without introducing excessive noise, while

subsequent components contribute meaningful secondary patterns related to station differences (Junninen et al., 2004). The improved MAE at $l = 5$ indicates that this configuration more effectively reconstructs missing rainfall values across stations, resulting in more representative imputation outcomes. From a hydrological standpoint, preserving moderate variability is essential to avoid over-smoothing precipitation extremes that are critical for downstream analyses. Therefore, despite the slightly higher RMSE at $l = 5$, this configuration provides a superior balance between model complexity and imputation accuracy. Similar findings have been reported in previous studies, where retaining several leading singular components yields more stable and interpretable imputations for environmental and hydro-meteorological datasets. Therefore, $l = 5$ provides an optimal balance between model complexity and imputation accuracy, and is selected as the best configuration for the SVD-based missing value imputation. It will also be applied in the integration of KNN–SVD in this study.

4. Development of the Integration Model (KNN-SVD)

The integration of the k-Nearest Neighbors (KNN) and Singular Value Decomposition (SVD) methods using the Weighted Linear Combination (WLC) approach is designed to optimize data imputation performance. WLC functions as an adaptive fusion mechanism, allowing the contribution weights between KNN and SVD to be proportionally adjusted. Consequently, this integrated model not only reduces dependency on a single method but also enhances robustness against complex data variability. Mathematically, WLC enables a weighted linear combination that preserves the fundamental properties of both techniques. This integrated approach has been shown to improve prediction accuracy in studies involving environmental and time-series data (Carter & Rinner, 2014; Malczewski, 2000).

In this study, the best parameters obtained from the initial stage, namely $k = 10$ in KNN and $l = 5$ in SVD, were selected because they produced the smallest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The selection of these optimal parameters is crucial so that the integration with WLC operates under the most representative configuration, thereby consistently minimizing prediction errors (Willmott & Matsuura, 2005). The subsequent discussion focuses on the performance of the integrated model in providing more stable and accurate imputation results. The MAE and RMSE values for $k = 10$ and $l = 5$, based on weights, can be seen in Figure 2.

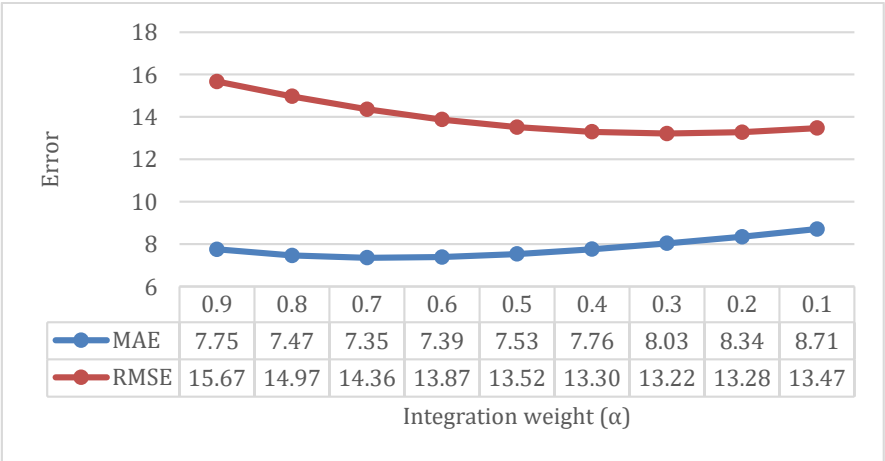


Figure 2. Performance of the KNN-SVD integration method

Figure 2 presents the results of integration testing with the best parameters, namely $k = 10$ and $l = 5$, which yielded better imputation performance compared to the use of a single method. From the experiments, variations in integration weights were shown to significantly affect model performance. The lowest MAE was obtained when the weight leaned more toward SVD ($\alpha = 0{,}7$), while the lowest RMSE was achieved when KNN had a greater dominance ($\alpha = 0{,}3$). These results indicate that the two methods complement each other: KNN is more effective in maintaining the stability of the average error, whereas SVD contributes to reducing extreme errors.

From a mathematical perspective, the weight parameter(α) is used as a control to balance bias and variance. A larger weight on KNN tends to reduce error variance by preserving the accuracy of local patterns, while a larger weight on SVD reduces bias by utilizing global information from matrix decomposition. Thus, the flexibility in choosing α allows the model to adapt to the characteristics of the analyzed data. This view is consistent with (Hastie et al., 2006), who emphasized that weighting-based combinations often improve predictive performance by balancing the strengths and weaknesses of each algorithm. Based on these results, it can be concluded that the integration of KNN-SVD through linear weighting is proven to produce more accurate and stable rainfall data imputation. The advantage of this integration lies not only in reducing error values but also in its ability to maintain the reliability of imputation results against data variations, thereby supporting climatological analysis and data-driven decision-making more effectively. After obtaining the optimal parameters for KNN and SVD, as well as their integration results, the next step is to evaluate the method's performance under various training and testing data proportions. This evaluation aims to assess the effect of the testing data proportion on the imputation error values. The performance evaluation results of KNN, SVD, and the KNN–SVD integration are presented in Figure 3.

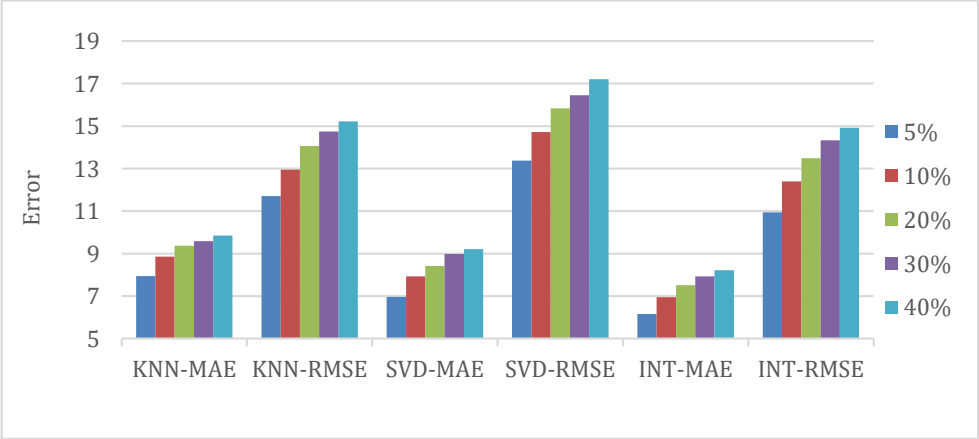


Figure 3. Imputation performance based on proportion

Figure 3 illustrates the performance evaluation results of the imputation methods based on variations in training and testing data proportions, showing a consistent pattern: the larger the proportion of data used for training, the smaller the resulting error values. At the 95:5 proportion, both MAE and RMSE reached their lowest values compared to other proportions, indicating that a larger availability of training data provides a more comprehensive representation of rainfall patterns. Conversely, as the testing data proportion increases, the

error gradually rises, reflecting the model's limitation in constructing estimations with the available information.

Mathematically, this phenomenon can be explained through the principle of the bias-variance trade-off, while the integration method plays a role in balancing the bias and variance inherent in each approach. This finding is consistent with the literature, which states that the more data available for the estimation process, the more stable the imputation results tend to be (Richards et al., 1989). Therefore, the KNN-SVD integration is proven not only to excel in terms of numerical performance but also to be adaptive to variations in data proportions, making it a more robust method for rainfall data imputation in West Java. The model's performance needs to be analyzed based on each observation station, since rainfall characteristics at different locations exhibit distinct patterns and variability, which may result in varying error values. Thus, station-based evaluation can provide a more detailed picture of the consistency of the imputation method, while also identifying which stations tend to produce the smallest error values. The model performance error values by station can be seen in Figure 4.

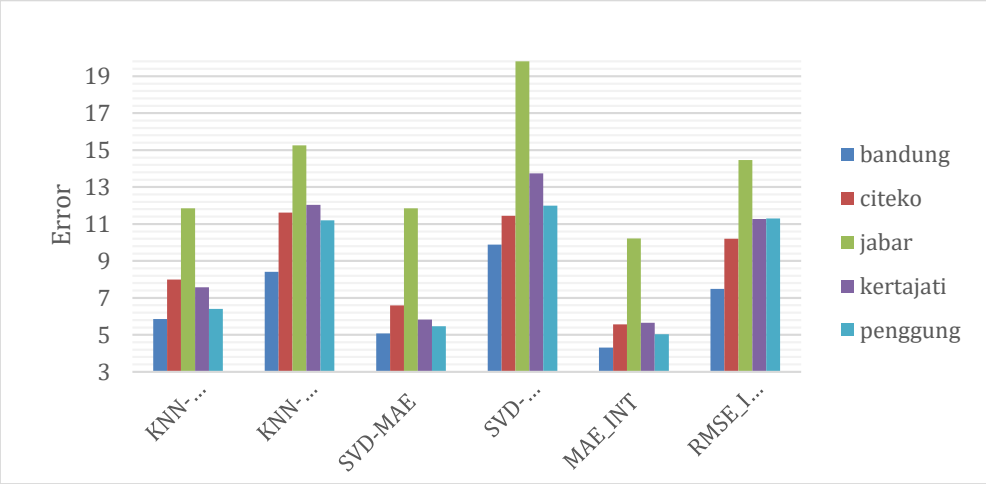


Figure 4. Imputation performance by station

Figure 4 shows the evaluation results indicating considerable variation in error across stations. The Bandung station recorded the lowest error values, both in MAE and RMSE, indicating a more stable rainfall pattern that is easier for the imputation method to learn. In contrast, the Jabar station recorded the highest error, particularly in RMSE, reflecting high rainfall fluctuations and data complexity that are more difficult to handle using either the KNN or SVD approach. From a methodological perspective, the KNN–SVD integration consistently produced lower errors compared to single methods across all stations. This result demonstrates that combining the two approaches can balance bias and variance, as well as improve prediction accuracy, in line with the theory that integration methods can enhance performance (Doz et al., 2023). Overall, the KNN–SVD integration proved to be more stable and accurate, even with data exhibiting diverse spatial characteristics, making it robust for climatological analysis in West Java. The integrated KNN–SVD method outperformed single methods, with the appropriate selection of parameters k in KNN and the parameter l in SVD, significantly reducing errors. Furthermore, although errors tend to increase with higher

proportions of missing values and differences in station characteristics, the integration method consistently provided the best results, making it a reliable imputation approach worthy of recommendation.

5. Discussion

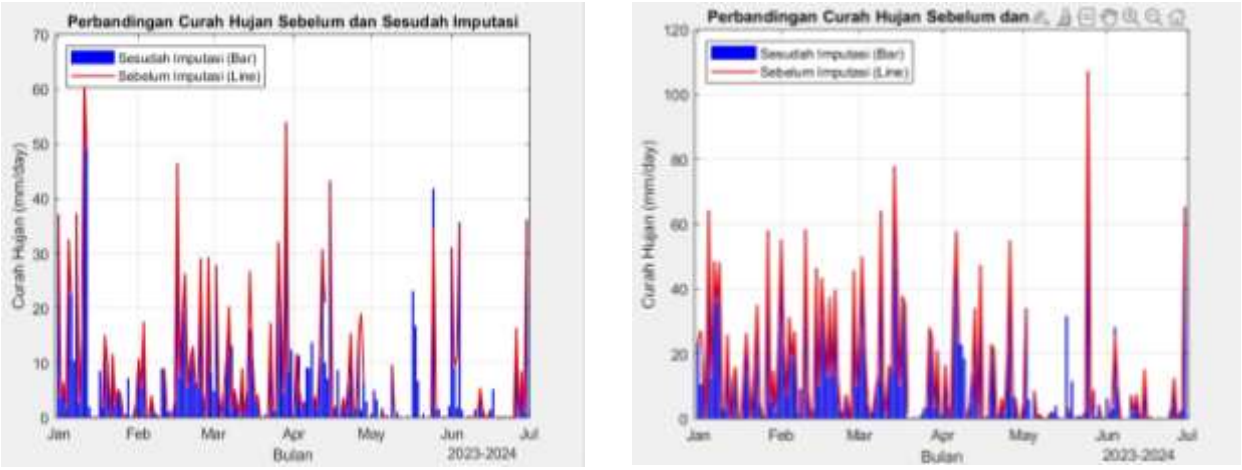
The superior performance of the integrated KNN–SVD model observed in the results can be directly attributed to the complementary roles of the two constituent methods. The KNN component effectively captures local rainfall similarity among neighboring stations, which is crucial for representing spatially correlated precipitation behavior, while the SVD component preserves dominant seasonal and spatial rainfall patterns through low-rank approximation. By combining these local and global sources of information using a weighted linear combination scheme, the proposed model is able to mitigate the limitations of each individual method, resulting in more robust and stable imputation performance for rainfall data characterized by high variability.

The results of this study are consistent with previous research showing that KNN-based imputation performs well in environmental and rainfall datasets due to its ability to exploit local similarity among observation points. Similarly, earlier studies have reported that SVD-based or low-rank approaches are effective in capturing dominant temporal and spatial patterns in climatic data. However, most existing studies apply these methods independently. In contrast, this study demonstrates that integrating KNN and SVD through a weighted linear combination framework provides a more balanced utilization of local and global information, leading to more stable and accurate rainfall imputation across multiple stations. This highlights the academic contribution of the proposed approach in extending existing imputation strategies for spatio-temporal rainfall data.

This study provides a significant contribution to the development of rainfall data imputation methods, particularly through the integration of KNN and SVD using the Weighted *Linear Combination* (WLC) approach. The results indicate that the combination of optimal parameters, namely $k = 10$ in KNN and $l = 5$ In SVD, it produces lower MAE and RMSE values compared to single methods. This suggests that the integration approach can enhance both the accuracy and stability of imputation results for data with complex characteristics. Nevertheless, this study has certain limitations, including the data coverage being restricted to the West Java region, meaning that the generalization of results needs to be tested in areas with different climatic conditions. Additionally, aspects of temporal variability and extreme climate events have not been fully explored, thus providing opportunities for further research. Future model development could focus on integrating other imputation methods or ensemble learning techniques to improve model reliability and resilience against more diverse data dynamics.

The visualization of imputation results shows a comparison between the original data and the filled-in data, where the imputed values tend to follow the actual rainfall values. The prediction line from the integrated model is able to represent seasonal patterns, including periods of high and low rainfall, with relatively small differences compared to the observed data. Overall, these results confirm that the KNN–SVD integration through WLC outperforms single methods, as it reduces prediction errors while maintaining data consistency. Therefore, the proposed method can serve as a reliable solution for addressing missing values in rainfall

data, particularly in the context of climate and hydrological analysis in West Java. A comparison of rainfall values before and after imputation can be seen in Figure 5.



(a) (b)

Figure 5. Comparison of rainfall before and after imputation: (a) Bandung station; (b) Citeko station

Figure 5 presents the results of rainfall data imputation, displayed as a graph to compare the original and filled-in data. The visualization was performed for several sample stations, such as Bandung (Figure 5a) and Citeko (Figure 5b). In the graph, the X-axis represents the daily observation time, while the Y-axis indicates rainfall (mm/day). Observed data before imputation are shown with a red line, whereas the imputed data are visualized with blue bars. The previously empty areas caused by missing values were successfully filled by the model, preserving the seasonal rainfall patterns.

At the Bandung station (Figure 5a), it can be observed that the imputation model accurately follows the actual rainfall patterns, with relatively small differences, both during periods of high and low rainfall. Similar results are observed at the Citeko station (Figure 5b), where the KNN–SVD integration with WLC consistently preserves temporal values and reduces deviations from the observed data. This visualization confirms that the proposed integration method successfully reduces prediction errors while improving data quality. Therefore, this method can be relied upon to correct missing values in rainfall data at stations in West Java, supporting more accurate climate and hydrological analyses.

From a practical perspective, the proposed KNN–SVD imputation framework can support the development of more reliable rainfall datasets, which are essential for hydrological modeling, climate analysis, and disaster risk management in regions with incomplete observations. Despite its promising performance, this study is subject to several limitations. The analysis is based on a limited number of observation stations and assumes a random missing data mechanism, which may not fully represent real-world missing patterns. Additionally, only rainfall data were considered in this study. Future research may extend the proposed framework by incorporating a larger spatial network, addressing non-random missing mechanisms, and integrating additional climatic variables to further enhance imputation accuracy and robustness.

The results obtained in this study are consistent with and support findings from previous research on missing value imputation in environmental and rainfall datasets. Several studies

have reported that KNN-based imputation performs effectively by exploiting local the similarity among absorption points, particularly in climate and hydrological data, which aligns with the strong performance of the KNN model observed in this study (Fadlil et al., 2022; Nida et al., 2023; Saeipourdizaj et al., 2021). Similarly, prior research has demonstrated that SVD-based or low rank approximation methods are capable of capturing dominant spatial and temporal patterns in climatic data, although their performance may be limited when used as standalone approaches (Arciniegas-Alarcón et al., 2020; Yuan et al., 2019). Importantly, this study extends previous findings by demonstrating that integrating KNN and SVD through a weighted linear combination framework produces more accurate and stable imputation results than either method alone. This outcome is in line with earlier studies showing that integrated or hybrid imputation approaches outperform single-method techniques by balancing local and global information (Carter & Rinner, 2014; Sanjar et al., 2020). Therefore, the results of this study not only support existing literature but also contribute additional empirical evidence highlighting the effectiveness of hybrid imputation strategies for regional rainfall data characterized by spatial heterogeneity, such as those in West Java.

D. CONCLUSION AND SUGGESTIONS

The results of this study indicate that the integrated KNN-SVD model consistently produced the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values compared to standalone KNN and SVD models across all evaluated training and testing data proportions, with the best performance achieved using 95% training data. Quantitatively, the integrated KNN-SVD model consistently achieved lower error values, with MAE ranging from 7,35 to 8,71 and RMSE ranging from 13,22 to 15,67 across all experimental configurations, outperforming the standalone KNN and SVD methods. This confirms that the integrated KNN-SVD model is generally superior and capable of providing consistent imputation results across different data conditions and observation station characteristics in West Java. However, these results are limited to station-based rainfall data with similar spatial and temporal characteristics and should not be directly generalized to other regions or climatic variables without further validation. From a practical perspective, the improved imputation accuracy enhances the reliability of rainfall datasets for hydrological analysis, flood risk management, and regional environmental policy support, while future research is recommended to test the proposed approach on additional stations, longer observation periods, and alternative integrated imputation frameworks to further extend its applicability.

REFERENCES

Afshar, M., & Usefi, H. (2021). Dimensionality reduction using singular vectors. *Scientific Reports*, *11*(1), 1–13. https://doi.org/10.1038/s41598-021-83150-y

Agrawal, J. Das. (2023). ANN in forecasting Missing Rainfall Data. *E3S Web of Conferences*, *405(2)*, 1–10. https://doi.org/10.1051/e3sconf/202340504017

Arciniegas-Alarcón, S., García-Peña, M., & Krzanowski, W. J. (2020). Imputation using the singular value decomposition: Variants of existing methods, proposed and assessed. *International Journal of Innovative Computing, Information and Control*, *16*(5), 1681–1696. https://doi.org/10.24507/ijicic.16.05.1681

Arciniegas-Alarcón, S., García-Peña, M., Krzanowski, W. J., & Rengifo, C. (2023). Missing value imputation in a data matrix using the regularised singular value decomposition. *MethodsX*, *11*(July), 1-8. https://doi.org/10.1016/j.mex.2023.102289

Azman, A. H., Tukimat, N. N. A., & Malek, M. A. (2021). Comparison of Missing Rainfall Data Treatment Analysis at Kenyir Lake. *IOP Conference Series: Materials Science and Engineering*, *1144*(1), 012046. https://doi.org/10.1088/1757-899x/1144/1/012046

Brand, M. (2002). Incremental singular value decomposition of uncertain data with missing values. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2350(1)*, 707–720. https://doi.org/10.1007/3-540-47969-4_47

Carter, B., & Rinner, C. (2014). Locally weighted linear combination in a vector geographic information system. *Journal of Geographical Systems*, *16*(3), 343–361. https://doi.org/10.1007/s10109-013-0194-3

Cihan, P., & Ozger, Z. B. (2019). A new heuristic approach for treating missing value: ABCIMP. *Elektronika Ir Elektrotechnika*, *25*(6), 48–54. https://doi.org/10.5755/j01.eie.25.6.24826

Duarte, L. V., Formiga, K. T. M., & Costa, V. A. F. (2022). Comparison of Methods for Filling Daily and Monthly Rainfall Missing Data: Statistical Models or Imputation of Satellite Retrievals? *Water (Switzerland)*, *14*(19), 1-20. https://doi.org/10.3390/w14193144

Erichson, N. B., Voronin, S., Brunton, S. L., & Kutz, J. N. (2019). Randomized matrix decompositions using R. *Journal of Statistical Software*, *89*(11), 1-48. https://doi.org/10.18637/jss.v089.i11

Fadlil, A., Herman, & Praseptian M, D. (2022). K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(4), 570–576. https://doi.org/10.29207/resti.v6i4.4173

Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, *41*(12), 3692–3705. https://doi.org/10.1016/j.patcog.2008.05.019

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(1), 0–36. https://doi.org/10.1186/s41044-016-0014-0

Ha, J., Kambe, M., & Pe, J. (2011). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*, Morgan Kaufman. https://doi.org/10.1016/C2009-0-61819-5

Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, *53*(2), 217–288. https://doi.org/10.1137/090771806

Hariyanto, D. (2020). *Optimization of Missing Value Data Imputation Automatic Dependent Surveillance* Broadcasting *(ADS-B) Based on K-Nearest Neighbor and Genetic Algorithm*. *9*(12), 327–331. https://doi.org/10.7753/ijcatr0912.1003

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., & Botstein, D. (2006). Imputing missing data for gene expression arrays. *Stanford University Statistics Department Technical Report Httpwwwstat Stanford Edu HastiePapersmissing Pdf Cll Qxd*, *3*(March 2013), 27. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.9789&rep=rep1&type=pdf

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Kim, T., Ko, W., & Kim, J. (2019). Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. *Applied Sciences (Switzerland)*, *9*(1), 1-18. https://doi.org/10.3390/app9010204

Lee, B., Lee, H., & Ahn, H. (2020). Improving load forecasting of electric vehicle charging stations through missing data imputation. *Energies*, *13*(18), 1–15. https://doi.org/10.3390/en13184893

Liu, C. H., Tsai, C. F., Sue, K. L., & Huang, M. W. (2020). The feature selection effect on missing value imputation of medical datasets. *Applied Sciences (Switzerland)*, *10*(7), 1–12. https://doi.org/10.3390/app10072344

Malczewski, J. (2000). On the use of weighted linear combination method in GIS: Common and best practice approaches. *Transactions in GIS*, *4*(1), 5–22. https://doi.org/10.1111/1467-9671.00035

Moatadid, I., Abnane, I., & Idri, A. (2023). Comparing Ensemble and Single Classifiers Using KNN Imputation for Incomplete Heart Disease Datasets. *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings*, *1*(Ic3k), 379–386. https://doi.org/10.5220/0012208300003598

Mohammed, M. B., Zulkafli, H. S., Adam, M. B., Ali, N., & Baba, I. A. (2021). Comparison of five imputation methods in handling missing data in a continuous frequency table. *AIP Conference Proceedings*, *2355*(April 2022). https://doi.org/10.1063/5.0053286

Nida, H., Kashif, M., Khan, M. I., & Ghamkhar, M. (2023). Comparison of missing data imputation methods using weather data. *Pakistan Journal of Agricultural Sciences*, *60*(2), 327–336. https://doi.org/10.21162/PAKJAS/23.228

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*, *13*(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

Ochieng' Odhiambo, F. (2020). Comparative Study of Various Methods of Handling Missing Data. *Mathematical Modelling and Applications*, *5*(2), 87. https://doi.org/10.11648/j.mma.20200502.14

Pertiwi, A. G., Bachtiar, N., Kusumaningrum, R., Waspada, I., & Wibowo, A. (2020). Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm without smote in diagnosis of diabetes disease in balanced data. *Journal of Physics: Conference Series*, *1524*(1), 1-8. https://doi.org/10.1088/1742-6596/1524/1/012048

Purnamasari, I., Wahyu Saputra, T., & Ristiyana, S. (2021). Pola Spasial Kekeringan di Jawa Barat Pada Kondisi El Nino Berbasis Metode Palmer Drought Severity Index (PDSI). *Jurnal Teknik Pengairan*, *12*(1), 16–29. https://doi.org/10.21776/ub.pengairan.2021.012.01.02

Radjab, F., Akib, H., Jasruddin, J., Rifdan, R., & Umar, F. (2020). Three Parties Partnership Between BMKG, Government Institution and General Public on Management of Rainfall Observations Networks in South Sulawesi. *SSRN Electronic Journal*, *August*, 7(2), 28–30. https://doi.org/10.2139/ssrn.3528943

Rafhida, S. A., Nurdiati, S., Budiarti, R., & Najib, M. K. (2024). Bias Correction of Lake Toba Rainfall Data Using Quantile Delta Mapping. *CAUCHY: Jurnal Matematika Murni Dan Aplikasi*, *9*(2), 297–309. https://doi.org/10.18860/ca.v9i2.29124

Richards, L. E., Little, R. J. A., & Rubin, D. B. (1989). Statistical Analysis with Missing Data. In *Journal of Marketing Research* (Vol. 26, Issue 3), 1-449. https://doi.org/10.2307/3172915

Saeipourdizaj, P., Sarbakhsh, P., & Gholampour, A. (2021). Application of imputation methods for missing values of pm10 and o3 data: Interpolation, moving average and k-nearest neighbor methods. *Environmental Health Engineering and Management*, *8*(3), 215–226. https://doi.org/10.34172/EHEM.2021.25

Sanjar, K., Bekhzod, O., Kim, J., Paul, A., & Kim, J. (2020). *Missing Data Imputation for Geolocation-based Price Prediction Using KNN – MCF Method*, 9(227), 1-13. https://doi:10.3390/ijgi9040227

Septiawan, P., Nurdiati, S., & Sopaheluwakan, A. (2019). Numerical Analysis using Empirical Orthogonal Function Based on Multivariate Singular Value Decomposition on Indonesian Forest Fire Signal. *IOP Conference Series: Earth and Environmental Science*, *303*(1), 1-10. https://doi.org/10.1088/1755-1315/303/1/012053

Shantal, M., Othman, Z., & Bakar, A. A. (2023). Impact of Missing Data on Correlation Coefficient Values: Deletion and Imputation Methods for Data Preparation. *Malaysian Journal of Fundamental and Applied Sciences*, *19*(6), 1052–1067. https://doi.org/10.11113/mjfas.v19n6.3098

Sriwahyuni, L., Nurdiati, S., Nugrahani, E. H., Sukmana, I., & Najib, M. K. (2025). Imputation of Missing Daily Rainfall Data Using Convolutional Neural Networks (Cnn) With Spatial Interpolation. *Barekeng*, *19*(4), 2921–2936. https://doi.org/10.30598/barekengvol19iss4pp2921-2936

Steinbach, M., & Tan, P. N. (2009). kNN: k-Nearest Neighbors. In *The Top Ten Algorithms in Data Mining* (pp. 151–161). https://doi.org/10.1201/9781420089653-15

Syauqi, R. M., Sabrina, P. N., & Santikarama, I. (2023). K-Means Clustering with KNN and Mean Imputation on CPU Benchmark Compilation Data. *Journal of Applied Informatics and Computing*, *7*(2), 231–239. https://doi.org/10.30871/jaic.v7i2.6491

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Yuan, X., Han, L., Qian, S., Xu, G., & Yan, H. (2019). Singular value decomposition based recommendation using imputed data. *Knowledge-Based Systems*, *163(1)*, 485–494. https://doi.org/10.1016/j.knosys.2018.09.011

Zhang, Z. W., Tian, H. P., Yan, L. Z., Martin, A., & Zhou, K. (2022). Learning a Credal Classifier With Optimized and Adaptive Multiestimation for Missing Data Imputation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *52*(7), 4092–4104. https://doi.org/10.1109/TSMC.2021.3090210