Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Assessment Of IDW And ANN on Daily Rainfall Data Imputation in Semarang Central Java

Eko Taufiq Suharmanto[1)*], Aji Supriyanto[2)]
[1,2)]Faculty of Information Technology, STIKUBANK University, Semarang, Indonesia
[1)]ekotaufiq0011@mhs.unisbank.ac.id, [2)]ajisup@edu.unisbank.ac.id

Submitted :Jan 10, 2025 | **Accepted** : Jan 22, 2025 | **Published** : Feb 10, 2025

Abstract: Rainfall plays a critical role in the global water and energy cycle, influencing surface water availability and recharge processes both spatially and temporally. Traditional rainfall data collection using ombrometers provides accurate live data, but often faces the challenge of missing data due to equipment failure or transmission, especially in agencies such as BMKG. This study analyzes the effectiveness of Artificial Neural Network (ANN) and Inverse Distance Weighting (IDW) methods in imputing missing rainfall data in Semarang City, using data from 31 observation stations over a 34-year period. Of the total 177,093 data, there are 7,159 data that require imputation. The ANN model showed superior performance with an RMSE of 1.2231 mm and R^2 of 0.9961 in the wet season, and an RMSE of 0.9489 mm and R^2 of 0.9926 in the dry season. Meanwhile, the IDW method showed limitations with an RMSE of 18.8206 mm and R^2 of 0.0084 in the wet season, and an RMSE of 10.9974 mm and R^2 of 0.0019 in the dry season. Although IDW achieved perfect recall (1.0000) in rainfall event classification, its low precision resulted in a suboptimal F1-Score (0.6460 wet season; 0.3124 dry season). The results indicate that ANN is superior in capturing non-linear patterns of rainfall and adapting to seasonal variations, while IDW has significant limitations in explaining rainfall variability. These findings make an important contribution to the development of more reliable rainfall data imputation methods for climatology and hydrology applications in urban areas.

Keywords: Rainfall; Imputation; IDW; ANN; Missing Data

INTRODUCTION

Precipitation is crucial to Earth's water and energy cycles, as it regulates climate patterns and affects the distribution of surface water and groundwater recharge across space and time (Navarro Céspedes et al., 2022). As a key element in regional water balance, rainfall directly shapes both ecological systems and economic activities. Traditionally, data on rainfall is collected using rain gauges or ombrometers. These devices are recognized as a reliable source of reference data, as they are able to provide direct measurements of the rainfall that occurs at a given location. The importance of rainfall in the ecosystem and economy of a region cannot be overstated. By influencing ecological and economic processes, rainfall can determine crop growth patterns, the availability of water resources, and even affect socio-economic activities that depend on the availability of water.

The use of an ombrometer as a traditional measuring instrument remains the primary choice as it provides the most accurate data at the point of direct measurement. This helps researchers and policy makers to get a clearer picture of rainfall patterns, which in turn can be used in water governance planning and extreme weather-related disaster mitigation.

The Indonesia Meteorology, Climatology and Geophysics Agency (BMKG) is an institution that focuses on providing crucial weather and climate information services. It processes a large amount of complex weather and climate data, which requires the application of advanced artificial intelligence for various needs, such as earthquake forecasting, fire prediction, and wind strength estimation. However, BMKG faces challenges related to data loss due to equipment failure or transmission failure. This has an impact on the sustainability of the data collected, including the very important rainfall data (Oktaviani & Putrada, 2022). Incomplete rainfall data can reduce the quality of analysis in hydrological research, as these data are vital elements in hydrological modeling (Chiu et al., 2021; Jahan et al., 2019). The issue of missing data, particularly in the context of rainfall, has significant implications for scientific analysis, given its essential role in supporting hydrological studies and related modeling.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Therefore, resolving this issue is important for improving the accuracy and reliability of the information provided by BMKG. Through the use of advanced artificial intelligence technology, improvements in the data collection and processing system are expected to ensure that the accuracy and completeness of the data is maintained. Thus, improving the analysis and resolution of data loss problems at BMKG will not only strengthen the quality of weather and climate information services, but also support better decision-making in disaster mitigation and sustainable development planning. BMKG's role in providing high-quality data is becoming increasingly relevant, along with the increasing need for accurate and timely climate information in the current era of climate change.

Missing data in ground-based rainfall measurements occurs when no readings are recorded at specific times. These gaps can distort statistical analyses by affecting both wet/dry condition probabilities and extreme weather assessments. For instance, data gaps during rainy periods might conceal peak rainfall amounts within timeframes like water years (Demetris Koutsoyiannis, 2021). Meanwhile, limited data availability during dry spells or across extended timeframes can impact the evaluation of meteorological droughts, particularly when determining the average and peak duration and intensity of extreme weather events (Costa et al., 2021). Rainfall databases often face challenges, such as the presence of noise, inconsistencies and missing data. Therefore, a data recovery process is required through imputation to fill in missing data as well as filtering to remove irrelevant data (Varada Rajkumar & Subrahmanyam, 2021)Various methods have been developed to handle missing rainfall data in previous studies. The selection of the best method depends on the location, as each region has different rainfall patterns and spatial distributions (Azman et al., 2021). Methods for filling in missing rainfall data are designed to supplement the data set with model-based estimates. In general, conventional methods such as the normal ratio (NR) method, linear interpolation (LI), regression-based techniques, and the arithmetic average (AA) method are quite effective for dealing with relatively small data gaps. (Wangwongchai et al., 2023). However, more efficient methods are needed to interpolate rainfall data in the context of geospatial systems (Sahoo & Ghose, 2022). Thus, it is important to select and apply appropriate methods for dealing with missing data so that statistical analysis on rainfall time series can be performed accurately and reliably. The right approach will help in better describing rainfall dynamics, thus supporting more informative research and decision-making.

In recent years, as the need for a comprehensive and precise rainfall data network has increased, various approaches have been proposed to address missing data in recorded rainfall time series. These new approaches show great progress compared to conventional methods in overcoming the missing data problem (Miró et al., 2017). This study aims to assess the accuracy of several rainfall data imputation techniques such as Inverse Distance Weighting (IDW) and Artificial Neural Network (ANN). The main focus of this study is to improve the quality of missing or empty rainfall data at the Meteorology, Climatology and Geophysics Agency (BMKG). It is expected that this study can offer an effective and efficient imputation method, make an academic contribution to the field of meteorology and climatology, and offer alternative imputation methods suitable for various timeframes. The results of this research are expected to improve the quality of weather and climate predictions with more complete and accurate data.

LITERATURE REVIEW

In an effort to deal with the problem of missing data, two main approaches that are often used are conventional methods and imputation methods. The first approach, called conventional methods, includes actions such as ignoring missing data, deleting it, assigning a value of zero, or using the average as an estimate (Qin and Wang, 2023). These methods can work effectively if the percentage of missing data is less than 1%. However, their effectiveness decreases drastically if the level of missing data exceeds 5%. The second approach is the filling method, which aims to replace missing data with estimates based on the distribution of available data in a time series, taking into account an acceptable error rate. This method offers flexibility for the user to choose the technique that best suits the specific needs of the case at hand (Mohamad et al., 2022). Research related to data filling has been conducted in various contexts, through both soft computing techniques and statistical methods (Li et al., 2023). There are two main approaches to data analysis: soft computing and statistical methods. Soft computing uses AI-based techniques like Artificial Neural Networks (ANN), BPCA, KNN, and SVM. Statistical approaches rely on models like SARIMA, IDW, MICE, MLR, and spatial interpolation. A comparison study in Southwest Colombia evaluated different methods for estimating monthly rainfall. It found that among conventional statistical approaches (Non-linear Regression, Principal Component Regression, and Partial Least Squares Regression), Non-linear Regression performed worst, showing higher error rates. Artificial Neural Networks proved most accurate, with significantly lower error measurements (RMSE between 0.01-0.29, MAE between 0.01-0.22, and ABIAS between 0.01-0.22).(Castillo-Gómez et al., 2023).

In the methodology of calculating missing daily rainfall, the IDW is often considered as one of the most intensive and efficient techniques. This method works by collecting available data from weather stations, which are then used to estimate rainfall values at measurement points by means of interpolation. This interpolation process involves a weighting factor that is calculated based on the distance between the weather station and the measurement point (Wuthiwongyothin et al., 2021). IDW relies on the basic assumption that closer values have a

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

greater influence than farther ones. As such, data from weather stations located closer to the measurement point will be given a higher weight in the calculation process. This approach enables IDW to provide more accurate estimates by considering spatial variations of rainfall. In practice, the use of IDW contributes significantly to improving the precision of meteorological data, especially in the context of areas that have an uneven distribution of measurement stations. By effectively integrating spatial data, this method offers a reliable and robust solution to the challenges of analyzing rainfall data in complex geographical terrains. Therefore, the application of IDW is an important part of modern hydrometeorological studies, contributing to better climate modeling and prediction. Likewise, since IDW is relatively easy to implement and can be adapted to various time and space scales, it has become a popular choice for researchers and practitioners in the field of geospatial and information technology. Periodic evaluation and improvement of the IDW algorithm is also ongoing to adapt to the increasingly complex data needs in dynamic scientific studies.

Another researcher conducted an evaluation of rainfall data fills in Ghana. The analysis of data subjected to deficiencies or incompleteness showed that the IDW (Inverse Distance Weighting), OK (Ordinary Kriging), GC (Gaussian Copula), missForest, RE (Regression Estimation), and PPCA (Probabilistic Principal Component Analysis) methods provided superior performance at the various levels of data incompleteness studied. However, there are some differences in the results achieved by each method. The IDW and OK methods tend to overestimate the number of predicted dry periods. On the other hand, the GC method has difficulty in providing accurate estimates of dry periods (Addi et al., 2022).

METHOD

The research was conducted through three main phases: Input, Process, and Output, each of which consists of a series of systematic and structured steps. The research stages are shown in Fig. 1.

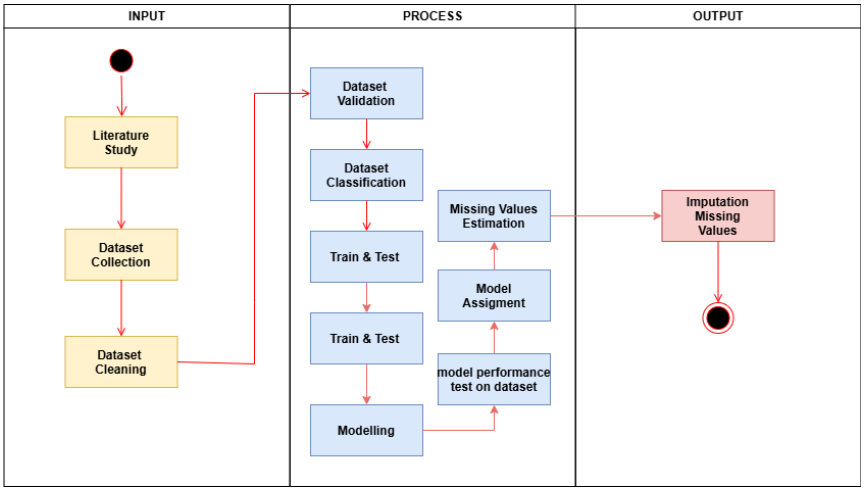

Fig. 1 Research Stage

This study uses primary data collected from the Central Java Climatology Station. The data source included 31 cooperative rainfall station observation points spread across the Semarang City area, providing comprehensive geographical coverage for regional level analysis. The observation period was determined from 1991 to 2024, providing a significant time span of 34 years, in accordance with the recommendation of the World Meteorological Organization, which suggests using at least 30 years of long-term climate data for climate evaluation (Sa'adi et al., 2023). The focus of the analysis was placed on daily rainfall data, which is an important meteorological parameter in climatology and hydrology studies. The presence of missing values is a common characteristic in multivariate time series that can be caused by various factors, including errors during the data collection process. This can affect the performance of advanced analytical applications on such time series (Zhang et al., 2021). Accurately estimating missing rainfall values is a major challenge, given that rainfall has high variability, both spatially and temporally (Mital et al., 2020). The choice of data sources, spatial coverage and time span allowed this study to capture long-term climate variability as well as specific patterns of rainfall. The use of daily rainfall data supports a more detailed analysis of precipitation phenomena, including the intensity, frequency and temporal distribution of rainfall. This data framework not only enriches the understanding of rainfall dynamics in the region, but also provides a strong foundation for further applications in hydrological modeling, prediction of drought conditions or analysis of long-term climate trends. With this, the research is able to provide deep insights that can be used in a variety of scientific and practical contexts. The sensor location shown in Fig.2.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

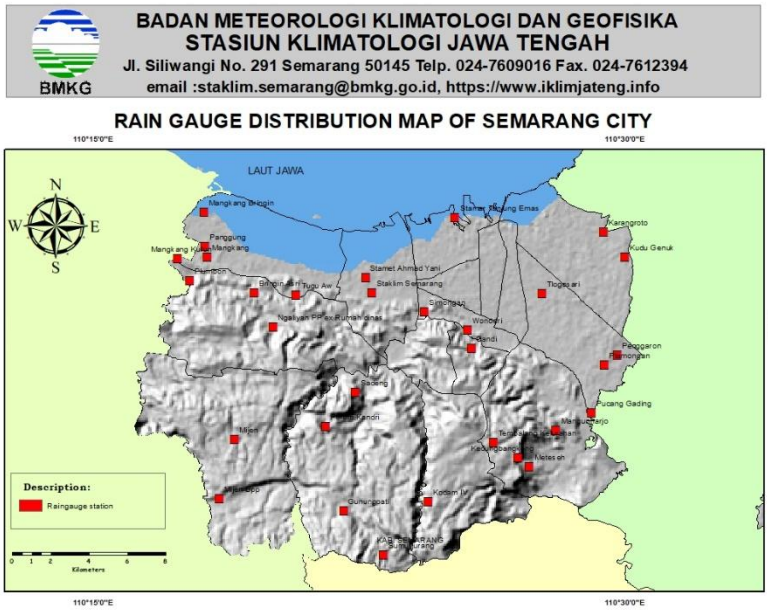

Fig. 2 Map of Rain Gauge Distribution in Semarang City

The initial stage in data preprocessing is crucial to ensure the data is ready to be used in further analysis. The first step in this process is profiling and cleaning the data. This step aims to address issues that are often found in raw data, such as missing or unnatural values, so as to make the data cleaner and ready for further processing. One of the main challenges in this stage is handling missing data effectively. For example, in certain datasets, values such as 8888 may indicate unmeasured rainfall data, while 9999 may indicate faulty observation equipment. These values should be converted to NaN (Not a Number) so that they can be imputed in the next step. At this stage, it is important to understand the data loss pattern and choose the right strategy to deal with the problem (Chiu et al., 2019). In addition, data preprocessing also includes steps such as normalization, transformation, and feature engineering. These processes aim to improve the quality of model inputs so that analysis and prediction can be performed more accurately. Normalization reduces the scale of data that varies too much, transformation transforms data into a format that is more suitable or optimal for analysis, and feature engineering helps in extracting more relevant information from existing data. By designing and using relevant and informative features, we can drastically improve model performance. This includes identifying key variables that may not be apparent from the underlying data, as well as creating new features that can better capture the complexity of the data. For example, combining time with temperature or humidity data can provide deeper insights that help in predictive analysis. (Abdelouahed et al., 2024).

Table 1. Feature Engineering

Transformasi	*Feature Engineering*
Siklik	*df['month_sin'] = np.sin(2 * np.pi * df['month']/12)*
Interaksi Spasial	*df['elevation_lat'] = df['elevation'] * df['latitude']*
Statistik Rolling	*df[f'rainfall_{window}d_mean'] = df.groupby('station_name')['rainfall'].transform(lambda x: x.rolling(window, min_periods=1).mean())*
Lag Features	*df[f'rainfall_lag_{lag}'] = df.groupby('station_name')['rainfall'].shift(lag)*

A three-layered Artificial Neural Network (ANN) uses feed-forward backpropagation to optimize connection weights between neurons. The algorithm minimizes the gap between predicted and actual outputs by continuously adjusting these weights during training until reaching minimal error (Agrawal, 2023). The model is evaluated using a specific equation that can be further explained :

$$P\chi\ (t) = f[p1(t), p2(t), p3(t), p4(t), \ldots, pn(t)] \qquad (1)$$

where p1, p2, p3, p4,...pn, denotes rainfall stations. The data used is normalized between 0 and 1 based on the log-sigmoidal function. Normalization is done using the following equation:

$$\ddot{X} = \frac{(X - X_{min})}{X_{max} - X_{min}} \qquad (2)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Where Ẍ is the normalized input value, X_max and X_min are the values of the actual value, across all observations and X is the original data set.

The basic model implemented contains three hidden layers, each with a different number of neurons: the first layer have 64 neurons, the second contains 32 neurons, and the third has 16 neurons. Each layer adopts the ReLU (Rectified Linear Unit) activation function, which serves to assist the neural network in recognizing complex and non-linear data patterns. For the output layer, a standard linear activation function suitable for regression tasks is used, thus facilitating rainfall prediction in a continuous range of values.

Some of the parameters used in the ANN model are described in Table 2.

Table 2. ANN Parameters

Category	*Parameters*	*Range*
Total Layer	*Hidden Layers*	*3*
Neuron Layer 1	*neurons_1*	*32,64,128*
Neuron Layer 2	*neurons_2*	*16,32,64*
Neuron Layer 3	*Fixed*	*16*
L1 Regularization	*l1*	*1.00E-05*
L2 Regularization	*l2*	*1.00E-04*
Dropout Rate	*dropout_rate*	*0.1,0.2,0.3,*
Learning Rate	*learning_rate*	*0.0001,0.001,0.01*
Optimizer	*-*	*Adam*
Batch Size	*batch_size*	*16,32,64*
Epochs	*epochs*	*50,100*
Hidden Layers	*Activation*	*ReLU*
Output Layer	*Activation*	*Linear*
.Early Stopping	*Patience*	*15*
Metric Monitoring	*-*	*Validation Loss*
Drop Factor	*factor*	*0.2*
Drop Patience	*patience*	*5*
Minimum LR	*min_lr*	*0.0001*

To overcome the overfitting problem and improve generalization ability, the model makes use of various sophisticated regularization techniques. One of the techniques used is elastic net, which is a combination of L1 and L2 regularization, applied to each layer by utilizing a kernel regularizer. This helps in controlling the complexity of the model weights. In addition, by applying a dropout rate between 0.2 to 0.3, a number of neurons are randomly deactivated during training, which promotes the formation of a more resilient model representation. Batch normalization is also applied after each hidden layer, which serves to normalize activations, stabilize the training process, and accelerate model convergence. The optimization process is performed using the Adam algorithm (Adaptive Moment Estimation). This algorithm is a sophisticated optimization method, combining the advantages of RMSprop and momentum. The learning rate used is adaptive with the application of the ReduceLROnPlateau mechanism, which automatically lowers the learning rate when there is no improvement in model performance. The learning rate ranges from 0.0001 to 0.01, allowing for a thorough exploration of the parameter space. The model has a unique approach to data handling, building separate models for the wet and dry seasons. This is done to accommodate the significant differences in rainfall patterns between the seasons. The process of finding the best parameters is done through random sampling of hyperparameter combinations, using time series cross-validation at each iteration. The model's training is stopped early by tracking the validation loss, allowing up to 15 epochs without improvement before termination.. This mechanism allows automatic termination of training when signs of overfitting appear, and restores the best weights obtained during training. The batch size used varies from 16 to 64, allowing the model to adapt to the computational and statistical characteristics of the dataset.

One of the main advantages of this artificial neural network (ANN) model lies in its ability to integrate complex spatial-temporal features. Through advanced feature engineering processes, such as cyclic transformation, rolling statistics, and lag features, the model is able to capture complex and non-linear rainfall dynamics.

The Inverse Distance Weighting (IDW) method is implemented to solve the problem of missing rainfall data through a series of systematic stages. The initial stage begins with data preprocessing which includes the removal of anomalous values onverted to blank values (NaN). Temporal transformation was performed on the timestamp data to generate more detailed time components, such as month, day, and season classifications, which are essential for rainfall pattern analysis. To improve prediction accuracy, comprehensive feature engineering was performed. Cyclic transformations are applied to temporal variables to maintain data continuity, while spatial interactions are formed by combining elevation, latitude and longitude parameters. Rolling statistics and lag features are also

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

integrated to capture the temporal patterns of historical data, providing a richer context for the imputation process. The IDW imputation process is implemented by dividing the dataset into training data and imputation target data. Comprehensive visualizations were generated to facilitate analysis of the results, including the distribution of missing values per station, temporal and spatial patterns of rainfall, and time series for each observation station. All imputation results and associated metadata are systematically documented in a structured format. This implementation demonstrates excellence in terms of considering the seasonal variability and spatial characteristics of rainfall data. Extensive use of feature engineering and seasonal-based parameter optimization resulted in an imputation model that is more adaptive to the data characteristics. The applied methodology not only produces missing value estimates, but also provides a comprehensive evaluation framework to validate the quality of imputation results. One such problem is the estimation of missing values at a base station, often denoted as Pm. This estimation process involves using the values observed at other stations as well as considering the distance between the base station and the other stations. All this is calculated by a formula (Djerbouai, 2022) :

$$P_m = \frac{\sum_{i=1}^{n} P_i d_{mi}^{-k}}{\sum_{i=1}^{n} d_{mi}^{-k}} \qquad (3)$$

P_m represents the rainfall measured at base station m, while P_i indicates the rainfall recorded at station i. The variable n denotes the total number of stations, dmi represents the distance between station m and station i, and k is defined as the friction distance. This approach is particularly relevant for applications in the context of climatology and hydrology, where the accuracy of rainfall data has significant implications for subsequent analysis and decision-making. Some parameters used in IDW can be seen in Table 3.

Table 3. IDW Parameters

Parameter Category	*Parameter*
Temporal	month, day, year, day_of_year, week_of_year
	season
Feature Engineering	mean, std, median, skewness (window size: 3, 7, 15, 30 days)
	lag (1, 2, 3, 7, 14 days)
Spatial	latitude, longitude, elevation
	elevation_lat, elevation_long, lat_long
IDW	power (1, 2, 3)
Validation	mask (20% of training data)
	RMSE, MAE, MAPE, R^2
Other	random seed (42)

In the implementation of the IDW method for rainfall data imputation, there are several key parameters used to optimize the imputation results. These parameters can be grouped into several main categories based on their function in the imputation process. Temporal parameters are an important component in the analysis of rainfall data. The code uses time parameters such as month, day, year, day_of_year, and week_of_year. Seasonal determination is also an important parameter, where the period November to April is categorized as the rainy season (season=1), while May to October as the dry season (season=2). In the context of feature engineering, several rolling statistical parameters are used with varying window sizes (3, 7, 15, and 30 days). These parameters include mean, standard deviation (std), median, and skewness. Lag parameters are also applied with periods of 1, 2, 3, 7, and 14 days to capture the temporal patterns of historical data. Spatial parameters are fundamental components in the IDW method, including latitude, longitude, and elevation. Interactions between spatial parameters are also formed through elevation_lat (elevation × latitude), elevation_long (elevation × longitude), and lat_long (latitude × longitude) to enrich spatial information in the imputation process. The most critical parameter in the IDW method is the power value that determines how quickly the influence of the observation point decreases with distance. In this implementation, power optimization is performed by trying values of 1, 2, and 3, The optimal value is chosen based on the lowest RMSE score obtained from the validation dataset. This optimization process is performed separately for each season to accommodate differences in rainfall characteristics. For model validation, a dataset sharing parameter is used where 20% of the training data is used as validation (mask = np.random.rand(len(train_season)) < 0.2). Evaluation parameters include RMSE, MAE,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

MAPE, and R^2 which are used to assess the accuracy of imputation results in each season (Marcelino et al., 2022; Norazizi & Deni, 2019; Saputra et al., 2021).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (5)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{|y_i - \hat{y}_i|}{y_i} \right) \quad (6)$$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (7)$$

This implementation also uses a random seed parameter to ensure reproducibility of results. This parameter is important in the context of scientific research as it enables consistent and replicable results.

RESULT

The spatial distribution of rainfall in Semarang City utilizes data from 31 observation stations spread across the city. The data, collected over more than three decades, provides deep insights and is visually displayed in Figure 3. The observation stations are located between 110.30° to 110.50° East Longitude and 6.94° to 7.11° South Latitude, covering the entire city of Semarang. Geographically, these observation stations are denser in the northern part of the city, especially in the latitude range between -6.94° and -7.00°. In contrast, in the southern part of the city, where latitudes range between -7.05° and -7.11°. This difference in density may reflect variations in data needs in different areas or practical constraints in station placement. Visual data shows that rainfall intensity in Semarang City is variable, with values ranging from 0 to 400 millimeters. Most of the observation points show relatively low intensity, indicated by the dominance of dark purple on the visualization graph. This suggests a certain pattern in rainfall distribution, which may have implications for drainage planning and flood anticipation in urban areas. The distribution of rainfall across space/area shown in Fig. 3.

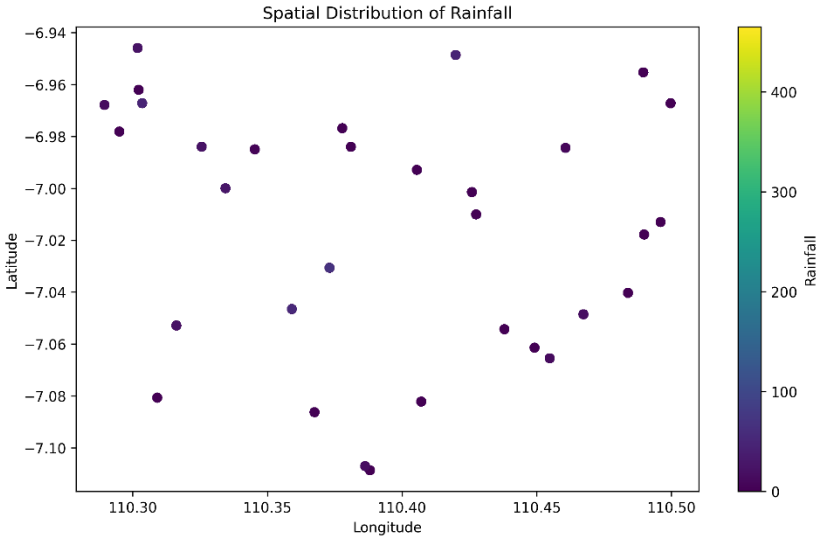

Fig.3 Spatial Distribution of Rainfall in Semarang City

Analysis of average monthly rainfall in Semarang City based on data from 31 observation stations for more than 30 years shows a dynamic and variable pattern as shown in Fig. 4. The data recorded rainfall ranges from 0 to 25 millimeters per month, with significant fluctuations but shows a consistent cyclic pattern throughout the observation period. The highest recorded rainfall peaks of around 24 millimeters were observed during the 150th and 250th months, while dry periods were characterized by rainfall close to 0 millimeters.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

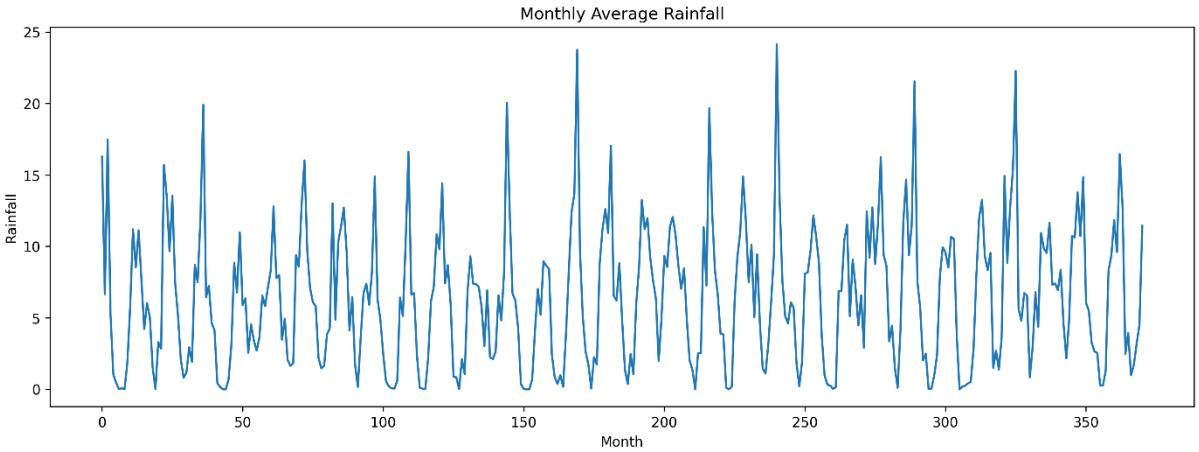

Fig. 4 Average Monthly Rainfall

A comprehensive evaluation of the Artificial Neural Network (ANN) model in rainfall data imputation showed very satisfactory results with consistent and accurate prediction characteristics. The model successfully produced imputed values within a realistic range, with a minimum value of 0.0 mm and a maximum of 465.0 mm, without producing meteorologically invalid negative values. In the wet season period, the ANN model was configured with an architecture consisting of two hidden layers (32 and 64 neurons) and a dropout rate of 0.3 to prevent overfitting. With a learning rate of 0.01 and a batch size of 64 for 100 epochs, the model achieved excellent performance with an RMSE of 1.2231 and MAE of 0.6543. The very high coefficient of determination (R^2) of 0.9961 indicates that the model is able to explain 99.61% of the variability in the data. The MAPE of 3.8580% indicates a relatively low prediction error rate. Furthermore, the precision value of 0.9987 and recall of 0.8294 resulted in an F1-score of 0.9062, demonstrating a good balance between accuracy and completeness of prediction. For the dry season period, the model was optimized with a different configuration, using 128 neurons in the first layer and 32 neurons in the second layer, as well as a lower dropout rate (0.1). With a smaller batch size (32) but the same learning rate, the model performed even better in some aspects. The lower RMSE and MAE (0.9489 and 0.3124) indicate higher prediction accuracy, albeit with a slightly lower R^2 (0.9926). The MAPE of 3.9501% still indicates an acceptable error rate. The precision value of 0.9891 and recall of 0.7495 resulted in an F1-score of 0.8528, indicating a solid albeit slightly lower classification performance compared to the wet season. A comparison of model performance between the two seasons revealed that the ANN performed better in terms of prediction accuracy (RMSE and MAE) in the dry season, but had slightly better generalization ability (R^2) and classification metrics in the wet season. This may be due to the lower variability of rainfall during the dry season, allowing the model to make more precise predictions, albeit with simpler pattern complexity. The results of this evaluation demonstrate that the optimized ANN model is capable of imputing rainfall data with a very high degree of accuracy, while maintaining realistic physical characteristics of the data. Consistent model performance across both seasons, with minimal error and robust evaluation metrics, indicates that this approach is well suited for rainfall data imputation applications in the context of climatology and hydrology.

Table 4. The Evaluation Results Of The ANN Model

Season	RMSE	MAE	MAPE (%)	R^2	PRECISION	RECALL	F1 SCORE
Wet	1.2231	0,6543	3.8580	0,9961	0.9987	0.8294	0.9062
Dry	0.9489	0.3124	3.9501	0,9926	0.9891	0.7495	0.8528

The data imputation process using the Inverse Distance Weighting (IDW) method over a 30-year period aims to correct missing or incomplete data, which in this context involves 7,159 data out of a total of 169,934 data used as training data. The IDW interpolation technique produced different results and parameters during wet and dry seasons. Analysis identified an optimal power parameter of 1 for both seasons, confirming a linear relationship between distance and rainfall estimation in the study area. During the wet season, the model exhibited substantial deviations with an RMSE of 18.8206 mm and MAE of 12.0882 mm. The MAPE value of 1.7821 indicates an average predictive deviation of 178.21% from actual observations. The minimal coefficient of determination (R^2) of 0.0084 suggests limited model capability in explaining rainfall variability. However, in terms of rainfall event classification, the model achieved perfect recall (1.0000) with moderate precision (0.4771), yielding an F1-Score of 0.6460, reflecting an adequate balance between these metrics. The dry season demonstrated enhanced numerical

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

accuracy with an RMSE of 10.9974 mm and MAE of 5.1438 mm, significantly lower than the wet season. A MAPE of 0.8694 indicates an average deviation of 86.94%, showing relative improvement compared to the wet season. Nevertheless, the extremely low R^2 (0.0019) continues to indicate limitations in the model's predictive capabilities. In classification context, while the model maintained perfect recall (1.0000), very low precision (0.1851) resulted in a suboptimal F1-Score (0.3124).

Performance divergence between seasons can be attributed to intrinsic precipitation characteristics. The wet season, characterized by higher temporal and spatial variability, produces more substantial errors, while the dry season's more stable patterns demonstrate minimal errors. Perfect recall in both periods indicates high model sensitivity in detecting rainfall events, though low precision reflects significant false positive prevalence, particularly during the dry season. These findings underscore the inherent complexity in rainfall data imputation and the significance of considering seasonal characteristics in model development. While the IDW model demonstrates superior capability in detecting rainfall events, low precision and R^2 values indicate the necessity for further development, potentially through the incorporation of additional predictors or implementation of more sophisticated methodologies to optimize quantitative estimation accuracy.

Table 5. The Evaluation Results Of The IDW Model

Season	RMSE	MAE	MAPE (%)	R^2	PRECISION	RECALL	F1 SCORE
Rainy	18.8206	12.0882	178.21	0.0084	0.4771	1	0.6460
Dry	10.9974	5.1438	86.94	0.0019	0.1851	1	0.3124

DISCUSSIONS

As shown in Figures 5 and 6, ANN (Artificial Neural Network) and IDW (Inverse Distance Weighting) methods demonstrated varying effectiveness in filling missing daily rainfall data from 31 weather stations across Semarang City over a 34-year period. The ANN method managed to reconstruct rainfall patterns well at most stations, especially at stations with high data variability. The ANN was able to recognize seasonal patterns and overcome uncertainties arising from significant data gaps. Visualization of the results shows that ANN imputation results in a more uniform distribution, reflecting the ability of the model to learn from temporal and spatial correlations among observing stations. In contrast, the IDW method, which is based on spatial interpolation, produces imputation results that depend on the geographical proximity between stations. This approach is effective for stations with close proximity and similar climate characteristics, but tends to be less accurate for stations with significant elevation differences or microclimate characteristics. Some small anomalies were observed in the IDW imputation results, especially at stations with very sporadic or discontinuous rainfall data. In terms of comparison, ANN is superior in modeling complex rainfall data because it is able to capture nonlinear relationships in the data.

Artificial Neural Network (ANN) and Inverse Distance Weighting (IDW) methods in imputing rainfall data in Semarang City show some interesting findings to be discussed. A comparison of the performance of these two methods reveals significant differences in data imputation capabilities, especially when analyzed based on seasonal characteristics. The ANN method shows consistent superiority in various evaluation metrics. In the wet season, the ANN model achieved a very high level of accuracy with an R^2 of 0.9961 and an RMSE of 1.2231 mm, indicating the model's superior ability to capture the complexity of rainfall patterns. The model performance even improved in the dry season with a lower RMSE (0.9489 mm), albeit with a slight decrease in R^2 (0.9926). This success can be attributed to the specifically optimized model architecture for each season, with different neuron configurations and dropout rates, demonstrating the importance of parameter adjustment based on the temporal characteristics of the data. On the other hand, the IDW method showed significant limitations in the imputation of rainfall data. Despite using the same optimal power parameter (1) for both seasons, the model produced substantial deviations with RMSE reaching 18.8206 mm in the wet season and 10.9974 mm in the dry season. The very low R^2 values (0.0084 and 0.0019) indicate the inability of the model to explain rainfall variability. However, it is worth noting that IDW achieved perfect recall (1.0000) in the classification of rainfall events, albeit with relatively low precision. The striking difference in performance between the two methods can be explained by several factors. Firstly, ANNs have the ability to learn complex non-linear patterns in rainfall data, while IDW relies on simple linear assumptions about the relationship between distance and rainfall values. Secondly, ANN can adapt its parameters based on seasonal characteristics, while IDW has limited flexibility in this regard. These findings have important implications for the development of rainfall data imputation systems. Although IDW has advantages in simplicity of implementation and interpretation, the results suggest that this method may not be optimal for areas with high rainfall variability such as Semarang City. In contrast, ANN shows excellent potential as a more reliable solution for rainfall data imputation, especially when optimized by considering seasonal characteristics. A limitation in this study lies in the uneven spatial distribution of observation stations, with higher

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14452
e-ISSN : 2541-2019
p-ISSN : 2541-044X

density in the northern area of the city. This may affect the performance of both models, especially IDW which is highly dependent on the distance between stations. Future research could explore the effect of station spatial distribution on imputation accuracy and the possibility of developing a hybrid model that combines the advantages of both methods. Overall, the results of this study contribute significantly to the understanding of the relative performance of rainfall data imputation methods and the importance of considering seasonal characteristics in model development. These findings can serve as a basis for the development of more reliable data imputation systems for climatology and hydrology applications in urban areas.

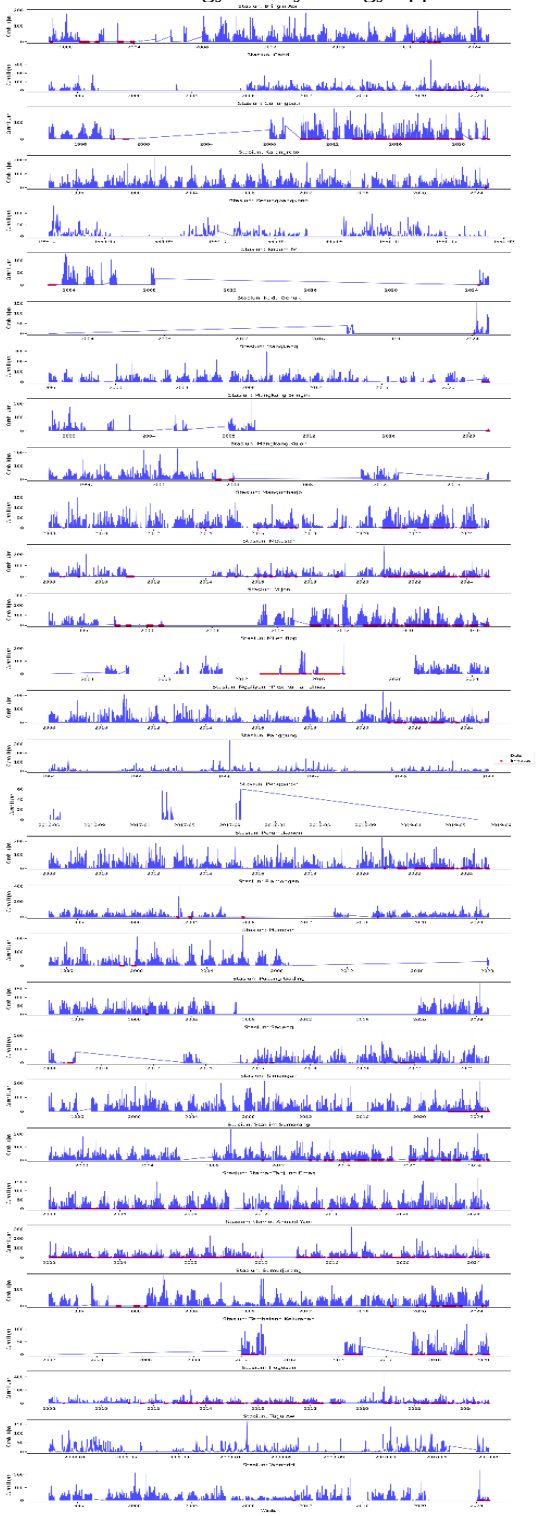

Fig. 5 Time series imputation with ANN

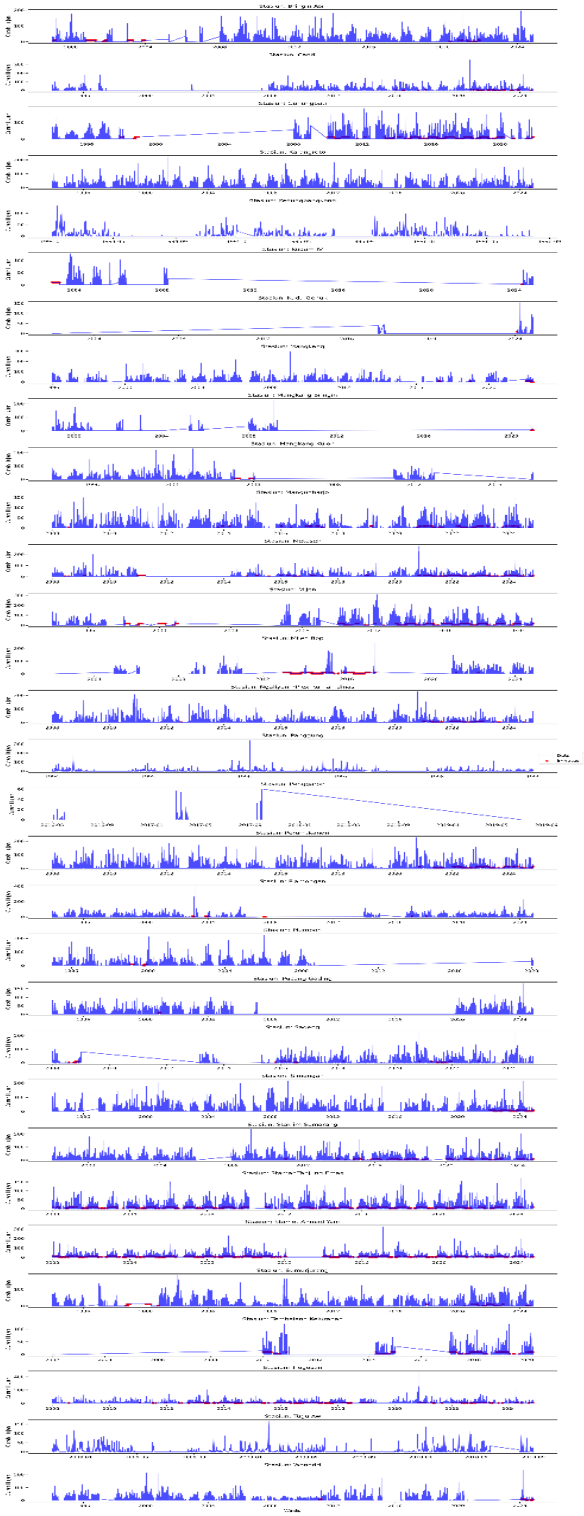

Fig. 6 Time series imputation with IDW

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

CONCLUSION

This study has successfully conducted a comprehensive evaluation of the spatial distribution of rainfall and data imputation methods in Semarang City using data from 31 observation stations for more than three decades. Some important findings can be summarized as follows: The spatial distribution of rainfall observation stations in Semarang City shows different densities between the northern and southern regions, with higher concentrations in the northern region (latitude -6.94° to -7.00°). Rainfall intensity varies between 0 and 400 millimeters, with the predominance of low intensity visualized through the graph. Monthly rainfall patterns show significant but consistent fluctuations, with the highest rainfall peak reaching 24 millimeters. The Artificial Neural Network (ANN) method showed excellent performance in imputing rainfall data. In the wet season, the model achieved high accuracy with RMSE 1.2231, MAE 0.6543, and R^2 0.9961, and F1-score 0.9062. The model performance in the dry season was even better in terms of prediction accuracy with an RMSE of 0.9489 and MAE of 0.3124, albeit with a slightly lower R^2 (0.9926). The optimal architecture of the model is different for the two seasons, indicating the importance of parameter adjustment based on seasonal characteristics. Meanwhile, the Inverse Distance Weighting (IDW) method showed mixed results. In the wet season, the model produced an RMSE of 18.8206 mm and MAE of 12.0882 mm with a very low R^2 (0.0084). The dry season showed better numerical accuracy with an RMSE of 10.9974 mm and MAE of 5.1438 mm, but still with a very low R^2 (0.0019). Despite having a perfect recall (1.0000), the low precision resulted in a less than optimal F1-Score, especially in the dry season. Based on the results of this study, it can be concluded that the ANN method is superior to IDW in imputing rainfall data in Semarang City. The ANN model is able to maintain realistic physical characteristics of the data while providing high prediction accuracy in both seasons. However, the difference in performance between the wet and dry seasons indicates the importance of considering seasonal variability in the development of rainfall data imputation models. This research makes a significant contribution to the understanding of rainfall spatial distribution and the development of reliable data imputation methods for climatology and hydrology applications in urban areas.

Based on the results and findings in this study, several recommendations can be made for future research development: First, the development of hybrid models that integrate the advantages of ANN and IDW methods needs to be further explored. The hybrid model can be designed to combine the ANN's ability to capture non-linear patterns with the IDW's ability to consider spatial aspects. This approach has the potential to produce a more robust and accurate imputation method. Secondly, optimizing the spatial distribution of rainfall observation stations needs special attention. Further research can focus on analyzing the effect of density and distribution of observation stations on imputation accuracy, as well as developing a methodology to determine the optimal location for placing new stations. Thirdly, the integration of satellite data and surface observation data could be a promising research direction. The use of satellite data can help overcome the limited spatial coverage of surface observation stations and potentially improve imputation accuracy, especially in areas with low station density. Fourth, the development of more complex deep learning models such as Convolutional Neural Networks (CNN) or CNN-LSTM combinations that can consider both spatial and temporal aspects simultaneously. These models have the potential to capture more complex rainfall patterns and produce more accurate imputations. Fifth, research on the influence of climate change on rainfall patterns and its implications for imputation methods needs to be carried out. This includes analysis of long-term trends and changes in rainfall characteristics, as well as adaptation of imputation models to these changes. Sixth, development of a more comprehensive framework for cross-validation, including model testing at different time scales and under different climate conditions. This will help understand the robustness and generalizability of the model under various conditions. Seventh, further investigation of the influence of other meteorological variables such as temperature, humidity and air pressure on the accuracy of rainfall data imputation. Integration of these variables into the model can improve understanding of the factors affecting rainfall and potentially improve imputation accuracy. Finally, the development of an early warning system based on rainfall imputation results needs to be explored. The implementation of these recommendations is expected to contribute significantly to the development of more accurate and reliable rainfall data imputation methods, as well as improve the understanding of rainfall dynamics in urban areas. This in turn will support better decision-making in water resources management and hydrometeorological disaster mitigation.

REFERENCES

Abdelouahed, S. M., Abla, R., Asmae, E., & Abdellah, A. (2024). Harnessing feature engineering to improve machine learning: A review of different data processing techniques. *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 1–6. https://doi.org/10.1109/ISCV60512.2024.10620105

Addi, M., Gyasi-Agyei, Y., Obuobie, E., & Amekudzi, L. K. (2022). Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. *Hydrological Sciences Journal*, *67*(4). https://doi.org/10.1080/02626667.2022.2030868

Agrawal, J. Das. (2023). ANN in forecasting Missing Rainfall Data. *E3S Web of Conferences*, *405*, 04017. https://doi.org/10.1051/e3sconf/202340504017

*name of corresponding author

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

Castillo-Gómez, J. S. Del, Canchala, T., Torres-López, W. A., Carvajal-Escobar, Y., & Ocampo-Marulanda, C. (2023). Estimation of monthly rainfall missing data in Southwestern Colombia: comparing different methods. *RBRH*, *28*. https://doi.org/10.1590/2318-0331.282320230008

Chiu, P. C., Selamat, A., Krejcar, O., & Kuok, K. K. (2019). Missing rainfall data estimation using artificial neural network and nearest neighbor imputation. *Frontiers in Artificial Intelligence and Applications*, *318*. https://doi.org/10.3233/FAIA190044

Chiu, P. C., Selamat, A., Krejcar, O., Kuok, K. K., Herrera-Viedma, E., & Fenza, G. (2021). Imputation of rainfall data using the sine cosine function fitting neural network. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(7). https://doi.org/10.9781/ijimai.2021.08.013

Costa, R. L., Barros Gomes, H., Cavalcante Pinto, D. D., da Rocha Júnior, R. L., dos Santos Silva, F. D., Barros Gomes, H., Lemos da Silva, M. C., & Luís Herdies, D. (2021). Gap Filling and Quality Control Applied to Meteorological Variables Measured in the Northeast Region of Brazil. *Atmosphere*, *12*(10), 1278. https://doi.org/10.3390/atmos12101278

Demetris Koutsoyiannis. (2021). *Advances in stochastics of hydroclimatic extremes*. https://doi.org/10.13140/RG.2.2.30655.05282/1

Djerbouai, S. (2022). Missing Precipitation Data Estimation Using Long Short-Term Memory Deep Neural Networks. *Journal of Ecological Engineering*, *23*(5), 216–225. https://doi.org/10.12911/22998993/147322

Jahan, F., Sinha, N. C., Rahman, M. M., Rahman, M. M., Mondal, M. S. H., & Islam, M. A. (2019). Comparison of missing value estimation techniques in rainfall data of Bangladesh. *Theoretical and Applied Climatology*, *136*(3–4). https://doi.org/10.1007/s00704-018-2537-y

Li, C., Ren, X., & Zhao, G. (2023). Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data. *Algorithms*, *16*(9), 422. https://doi.org/10.3390/a16090422

Marcelino, C. G., Leite, G. M. C., Celes, P., & Pedreira, C. E. (2022). Missing Data Analysis in Regression. *Applied Artificial Intelligence*, *36*(1), 2032925. https://doi.org/10.1080/08839514.2022.2032925

Miró, J. J., Caselles, V., & Estrela, M. J. (2017). Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric Research*, *197*. https://doi.org/10.1016/j.atmosres.2017.07.016

Mital, U., Dwivedi, D., Brown, J. B., Faybishenko, B., Painter, S. L., & Steefel, C. I. (2020). Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests. *Frontiers in Water*, *2*. https://doi.org/10.3389/frwa.2020.00020

Mohamad, N. B., Lai, A.-C., & Lim, B.-H. (2022). A case study in the tropical region to evaluate univariate imputation methods for solar irradiance data with different weather types. *Sustainable Energy Technologies and Assessments*, *50*, 101764. https://doi.org/10.1016/j.seta.2021.101764

Navarro Céspedes, J. M., Hernández, J. H., Alcántara Concepción, P. C., Morales Martínez, J. L., Carreño Aguilera, G., & Padilla Benítez, F. (2022). A comparison of missing values imputation methods applied to precipitation of two semi-arid and humid regions of México. *ATMÓSFERA*. https://doi.org/10.20937/ATM.53095

Norazizi, N. A. A., & Deni, S. M. (2019). Comparison of Artificial Neural Network (ANN) and Other Imputation Methods in Estimating Missing Rainfall Data at Kuantan Station. In M. W. Berry, B. W. Yap, A. Mohamed, & M. Köppen (Eds.), *Soft Computing in Data Science* (pp. 298–306). Springer. https://doi.org/10.1007/978-981-15-0399-3_24

Oktaviani, I. D., & Putrada, A. G. (2022). KNN imputation to missing values of regression-based rain duration prediction on BMKG data. *JURNAL INFOTEL*, *14*(4), 249–254. https://doi.org/10.20895/infotel.v14i4.840

Sa'adi, Z., Yusop, Z., Alias, N. E., Chow, M. F., Muhammad, M. K. I., Ramli, M. W. A., Iqbal, Z., Shiru, M. S., Rohmat, F. I. W., Mohamad, N. A., & Ahmad, M. F. (2023). Evaluating Imputation Methods for rainfall data under high variability in Johor River Basin, Malaysia. *Applied Computing and Geosciences*, *20*, 100145. https://doi.org/10.1016/j.acags.2023.100145

Sahoo, A., & Ghose, D. K. (2022). RETRACTED ARTICLE: Imputation of missing precipitation data using KNN, SOM, RF, and FNN. *Soft Computing*, *26*(12), 5919–5936. https://doi.org/10.1007/s00500-022-07029-4

Saputra, M. D., Hadi, A. F., Riski, A., & Anggraeni, D. (2021). Principal Component Regression in Statistical Downscaling with Missing Value for Daily Rainfall Forecasting. *International Journal of Quantitative Research and Modeling*, *2*(3), 139–146. https://doi.org/10.46336/ijqrm.v2i3.151

Varada Rajkumar, K., & Subrahmanyam, D. K. (2021). A Novel Method for Rainfall Prediction and Classification using Neural Networks. *International Journal of Advanced Computer Science and Applications*, *12*(7). https://doi.org/10.14569/IJACSA.2021.0120760

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Wangwongchai, A., Waqas, M., Dechpichai, P., Hlaing, P. T., Ahmad, S., & Humphries, U. W. (2023). Imputation of missing daily rainfall data; A comparison between artificial intelligence and statistical techniques. *MethodsX*, *11*, 102459. https://doi.org/10.1016/j.mex.2023.102459

Wuthiwongyothin, S., Kalkan, C., & Panyavaraporn, J. (2021). Evaluating Inverse Distance Weighting and Correlation Coefficient Weighting Infilling Methods on Daily Rainfall Time Series. *SNRU Journal of Science and Technology*, *13*(2).

Zhang, Y., Zhou, B., Cai, X., Guo, W., Ding, X., & Yuan, X. (2021). Missing value imputation in multivariate time series with end-to-end generative adversarial networks. *Information Sciences*, *551*. https://doi.org/10.1016/j.ins.2020.11.035

*name of corresponding author