VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7938

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

MULTI DATASET CLIMATOLOGICAL FORECASTING USING PSO TUNED RNN WITH MISSING VALUE IMPUTATION

Adhelia Wida Khaidir[1]; Aji Prasetya Wibawa[1]*; Dhia Rafifah Thifal[1]; Adelia Desyana Eka Putri[1]; Adelia Khansa Ristiaputri[1]; Agung Bella Putra Utama[1]

Department of Electrical Engineering and Informatics[1]
Universitas Negeri Malang, Malang, Indonesia[1]
https://um.ac.id/[1]
adhelia.wida.2205356@students.um.ac.id, aji.prasetya.ft@um.ac.id,
dhia.rafifah.2205356@students.um.ac.id, adelia.desyana.2205356@students.um.ac.id,
adelia.khansa.2205356@students.um.ac.id, agung.bella.2305349@students.um.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Missing values in climatological data due to technical glitches or extreme conditions can degrade prediction accuracy and result in biased analysis. This study aims to evaluate six imputation techniques, namely Mean, Mode, Median, Multiple Imputation by Chained Equations (MICE), Last Observation Carried Forward (LOCF), and K-Nearest Neighbors (KNN), on forecasting performance using the Recurrent Neural Network model optimized with Particle Swarm Optimization (RNN-PSO). Evaluation was carried out on three climatological time series datasets with different missing-value characteristics. The research methods include data exploration, imputation of missing values, hyperparameter optimization using PSO, and RNN model training. Model performance was evaluated using MAPE, RMSE, and R^2. The results show that imputation methods play an important role in maintaining the temporal integrity of the time series and consistently improving prediction precision, compared to missing data removal strategies that tend to disrupt temporal patterns and reduce model accuracy. In Dataset 1, the Median was effective in suppressing the relative error (MAPE 0.93374), but LOCF showed a more comprehensive performance by producing the lowest RMSE (3.55282) and the highest R^2 (0.91478), indicating its superiority in preserving temporal structure and improving overall prediction accuracy LOCF also showed strong performance on Dataset 2, especially on RMSE and R^2. Meanwhile, KNN provided the best results in Dataset 3 across all evaluation metrics by utilizing local correlations between data. These findings confirm that the selection of imputation methods needs to be adjusted to the characteristics of missing values, variability, and temporal properties of the dataset. Therefore, the selection of imputation methods should take into account temporal patterns and data distributions. This research provides a framework for selecting appropriate imputation methods to improve the accuracy of forecasting climatological data and support the development of a more reliable weather prediction system.*

Keywords: *Climatology Forecasting, Data Imputation, Missing Values, RNN-PSO, Time Series Forecasting.*

Intisari— *Nilai yang hilang dalam data klimatologis karena gangguan teknis atau kondisi ekstrem dapat menurunkan akurasi prediksi dan menghasilkan analisis yang bias. Penelitian ini bertujuan untuk mengevaluasi enam teknik imputasi, yaitu Mean, Mode, Median, Multiple Imputation by Chained Equations (MICE), Last Observation Carried Forward (LOCF), dan K-Nearest Neighbors (KNN), pada kinerja peramalan menggunakan model Recurrent Neural Network yang dioptimalkan dengan Particle Swarm Optimization (RNN-PSO). Evaluasi dilakukan pada tiga kumpulan data deret waktu klimatologis dengan karakteristik nilai hilang yang berbeda. Metode penelitian meliputi eksplorasi data, imputasi nilai yang hilang, optimasi hiperparameter menggunakan PSO, dan pelatihan model RNN. Kinerja model dievaluasi menggunakan MAPE, RMSE, dan R^2. Hasil penelitian menunjukkan bahwa metode imputasi memainkan peran penting dalam menjaga integritas temporal deret waktu dan secara konsisten meningkatkan presisi prediksi, dibandingkan dengan strategi penghapusan data yang hilang yang cenderung mengganggu pola temporal dan mengurangi*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7938

akurasi model. Pada Dataset 1, Median efektif dalam menekan kesalahan relatif (MAPE 0,93374), tetapi LOCF menunjukkan kinerja yang lebih komprehensif dengan menghasilkan RMSE terendah (3,55282) dan R^2 tertinggi (0,91478), menunjukkan keunggulannya dalam menjaga struktur temporal dan meningkatkan akurasi prediksi secara keseluruhan LOCF juga menunjukkan kinerja yang kuat pada Dataset 2, terutama pada RMSE dan R^2. Sementara itu, KNN memberikan hasil terbaik dalam Dataset 3 di seluruh metrik evaluasi dengan memanfaatkan korelasi lokal antar data. Temuan ini menegaskan bahwa pemilihan metode imputasi perlu disesuaikan dengan karakteristik nilai yang hilang, variabilitas, dan sifat temporal dari kumpulan data. Oleh karena itu, pemilihan metode imputasi harus mempertimbangkan pola temporal dan distribusi data. Penelitian ini memberikan kerangka kerja untuk memilih metode imputasi yang tepat untuk meningkatkan akurasi prakiraan data klimatologi dan mendukung pengembangan sistem prediksi cuaca yang lebih andal.

***Kata Kunci**: Peramalan Klimatologi, Imputasi Data, Nilai Yang Hilang, RNN-PSO, Peramalan Deret Waktu.*

INTRODUCTION

Climatological data plays a critical role in weather modelling, climate change analysis, disaster mitigation, and natural resource management [1]. However, data quality is often hampered by missing values due to technical glitches, gauge failures, or extreme weather conditions [2] Missing data leads to a reduction in available information, weakens statistical power, and has the potential to produce biased or erroneous conclusions if not handled properly [3] Thus, the appropriate handling of missing values is a crucial step to ensure that climatological analysis remains accurate, reliable, and able to support effective decision-making.

A common approach to dealing with missing values is to remove incomplete observations. Although simple, these strategies often omit important information, skew the data distribution, and disrupt the temporal patterns that are crucial for forecasting. Some studies have used simple imputation techniques, such as mean, median, or mode, to fill in missing values. However, these methods are often less flexible for data with complex temporal patterns or high variability [4]. Deep learning models such as Recurrent Neural Networks (RNNs) have also been shown to be sensitive to incomplete inputs, so that missing values can lower prediction accuracy [5].

For example, a previous study applied the Recurrent Neural Network (RNN) model to temperature forecasting and demonstrated its ability to capture temporal dependencies in climatological time series data. The study also improved model performance through Bayesian optimization using the GPyOpt framework for hyperparameter tuning. However, this study emphasizes model architecture and optimization, while the handling of missing data is limited to a basic imputation approach without thorough and systematic comparison. In particular, the impact of various imputation methods on model performance, especially when assessed using metrics such as MAPE, RMSE, and R^2, is still underexplored. Therefore, further investigation is needed to systematically examine how various imputation techniques affect the performance of RNN-based forecasting models [6]. This suggests that, although model optimization has been widely studied, the role of imputation methods in forecasting performance is still poorly explored.

Selecting a more adaptive imputation method aligned with the dataset's characteristics is key to maintaining model quality and improving forecasting performance. To date, few studies have systematically evaluated the influence of various imputation techniques on RNN performance in forecasting climatological data, particularly by considering the mechanisms underlying missing values. Previous research has tended to assess imputation methods solely on statistical metrics, without linking them to forecasting performance. It rarely considers the characteristics of different datasets during the forecasting process.

This study aims to evaluate six imputation techniques: MEAN, MODE, MEDIAN, Multiple Imputation by Chained Equations (MICE), Last Observation Carried Forward (LOCF), and K-Nearest Neighbors (KNN) on three climatology datasets with different characteristics and types of missing values. The forecasting process is carried out using a Recurrent Neural Network combined with Particle Swarm Optimization (RNN-PSO), with parameter and architecture settings adjusted to the temporal patterns, distributions, and local correlations of each dataset. This approach allows imputation and forecasting models to work synergistically to maintain continuity, local patterns, and historical trends, resulting in more accurate and reliable predictions.

In practice, this study provides guidance for researchers and practitioners on choosing the appropriate imputation method to improve the quality of climatological data forecasts. Theoretically, this study contributes to the

literature by examining the impact of missing values and imputation methods on RNN performance in time series forecasting [7]. This contribution opens the door to further research, including multivariate and spatial-temporal approaches to climatological data.

MATERIALS AND METHODS

This section describes a series of research methods carried out systematically, starting with data collection as the basis for analysis. The next stage is data exploration to assess its characteristics, patterns, and completeness. After that, an imputation method is applied to handle missing values, using the one that best suits the dataset's conditions. The entire series became the foundation before entering the final stage, namely forecasting using the RNN model, which was evaluated with MAPE, RMSE, and R^2.

Data Collection

This study uses three time-series datasets obtained from public repositories on the Kaggle platform. Dataset 1, Eighty Years of Canadian Climate Data [8], contains approximately 80 years of daily temperature observations and exhibits a relatively high proportion of missing values. This dataset is relevant for evaluating the robustness of imputation methods under conditions with substantial data loss. Dataset 2, Daily Sunspot Data (1818–2019) [9], records the daily number of observed sunspots as an indicator of solar activity, which acts as an external driver influencing the Earth's climate system through variations in solar irradiance.

Fluctuations in solar radiation contribute to long-term energy balance and are associated with climate variability on decadal scales, while the approximately 11-year solar cycle introduces a clear periodic pattern. Therefore, this dataset is used as a proxy for external forcing in climatological time-series analysis and to evaluate the model's ability to capture long-term temporal dependencies and cyclical behavior. Dataset 3, Sofia Weather Records [10], is the largest and has fewer missing values and a more stable distribution, providing a contrasting scenario with relatively complete observations.

These datasets represent different types of temporal data sources, including regional temperature measurements, astronomical activity indicators, and local weather observations. The use of datasets with diverse statistical properties, including varying levels of missing data, temporal periodicity, and data distribution, enables a more comprehensive evaluation of the proposed model across different variability conditions. Statistical summaries of the three datasets are shown in Table 1.

Table 1. Description of Statistics

Statistics	Dataset 1	Dataset 2	Dataset 3
Number of Instances	27,530	73,718	87,599
Average	-0.42°C	79.25	1.647
Standard deviation	12.87	77.47	9.695
Minimum	-48.1°C	-1	-36.9
Quartile 1 (25%)	-8.4°C	15	-3.6
Median (50%)	1.9°C	58	1.7
Quartile 3 (75%)	10.0°C	125	8.3
Maximum	23.9°C	528	29.1
Missing values	1,691	3,204	50

Source : (Research Results, 2025)

Based on Table 1, the three datasets show different statistical characteristics. Dataset 1 has a value distribution that tends to be low and stable with moderate variability. Dataset 2 shows a very wide, fluctuating distribution of values, including anomalous values, indicating the highest variability among the three. Meanwhile, Dataset 3 has a more centralized and consistent distribution, with the fewest missing values. Overall, Dataset 1 was dominated by low-nuanced values, Dataset 2 showed extreme data dynamics, and Dataset 3 appeared to be the most stable.

The target attributes in each dataset are selected to represent the key variables most relevant to the study's characteristics. Dataset 1 uses *MEAN_TEMPERATURE_WHITEHORSE*, which reflects daily temperatures in Whitehorse, Canada, with a subarctic climate and extreme seasonal variability, making it ideal for testing the model's ability to capture dynamic changes. Dataset 2 uses *the Number of Sunspots* as an indicator of solar activity with a periodic cycle pattern of about 11 years, suitable for assessing the model's performance in studying long-term repeating patterns.

Dataset 3 uses *air_temperature* from Sofia Weather Records that reflect daily air temperatures in a temperate continental climate with gradual changes. Selecting these three attributes enables evaluation of the model's performance across contrasting data patterns, from extreme fluctuations and periodic astronomical phenomena to moderate weather dynamics.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7938

Data Exploration

Data exploration was carried out to understand the structure, quality, and temporal dynamics of each dataset before entering the advanced analysis stage [11]. This process aims to identify the distribution patterns of variables, detect the presence of missing values, and analyze their magnitudes and patterns of occurrence. To quantify these characteristics, the percentage of missing values and gap-length statistics are calculated for each dataset, and the summary results are presented in Table 2.

Table 2. Missingness Magnitude and Gap-Length Statistics of the Datasets

Dataset	Variable	Total Data	Missing (%)	Number of Gaps	Gap Min	Gap Max	Gap Mean
Canadian Climate History	MEAN_TEMPERATURE_WHITEHORSE	28,400	1,691 (6.14%)	125	1	663	6.96
Sunspot Data	Number of Sunspots	73,711	3,204 (4.35%)	1,619	1	52	2.00
Sofia Weather Records	air_temperature	87,649	50 (0.06%)	11	1	24	4.55

Source : (Research Results, 2026)

Table 2 shows that each dataset has distinct characteristics of missingness, both in the proportion of data lost and the length of the temporal gap. This difference results in varied data conditions, ranging from datasets with relatively frequent gaps to those with very few missing values, thus providing diverse test scenarios for evaluating the robustness of the imputation method used in this study.

Table 3. Classification and Characteristics of Missing Value

Aspect	MCAR	MAR	MNAR
Definition	Missingness is independent of both observed and unobserved data [12].	Missingness depends on the observed variables, not on the missing value itself [13].	Missingness depends on the unobserved value itself [14].
Dependence	No dependence on any variable [15].	Depends on other observed variables [15].	Depends on the missing value or unobserved data [15].
Identifiability	Can be statistically tested (Little's MCAR test) [15], [16].	Not directly testable; inferred from relationships with observed variables [16]	Not empirically testable without additional assumptions [16].
Handling Approaches	Simple imputation or drop missing value methods [17].	Model-based or multivariate imputation (MICE, KNN) [18].	Requires explicit modeling of the missingness process or sensitivity analysis [19].
Analytical Impact	Standard analysis remains unbiased under MCAR [20].	Bias can be reduced with appropriate imputation [21].	High risk of bias if the mechanism is ignored [20].

Source : (Research Results, 2025)

The three datasets were analyzed based on the *missing value mechanism* to understand the pattern of absence of values before the imputation process was carried out. The classification of MCAR, MAR, and MNAR in this study is not only based on the level or amount of data loss, but mainly on the indication of the relationship between the data loss mechanism and the characteristics of the observed variables. In Dataset 1, data loss patterns that tend to appear under specific conditions in time series indicate that the missing mechanism is not completely independent of unobserved values. Therefore, this mechanism is suspected to follow an MNAR pattern, because the process of data loss is suspected to be related to the characteristics of extreme values that are not recorded [22]. Dataset 2 does not show a clear relationship between the existence of missing values and available variables or time patterns, so data loss can be assumed to occur randomly. This supports the MCAR classification, where the absence of data is not affected by observed or unobserved variables. [23]. Meanwhile, Dataset 3, although the number of missing is relatively small, there is an indication that the absence pattern is more likely to be influenced by external factors recorded in the data (such as time or measurement conditions), rather than by the value of the main variable itself. This condition is consistent with the MAR mechanism, where missingness depends on observable information [10]. Thus, the difference in missing mechanisms in these three datasets is an important basis for the selection of appropriate imputation strategies, as each assumption has direct implications for the bias and validity of the modeling results.

Data Imputation

Data imputation is the process of filling in missing values using statistical estimates or algorithms to maintain data completeness and integrity, so that analysis can be carried out without removing observations that risk losing important information. The selection of an imputation method must be tailored to the characteristics of the data, the pattern of missing values, and the mechanism underlying their occurrence. The advantages and disadvantages of the imputation method are shown in Table 4.

Table 4. Advantages and Disadvantages of Imputation Methods

Method	Excess	Deficiency
Mean	Fast, efficient	Reduce variability, not fit time series
Mode	Match categorical data	Just categorical, reduce variability
Median	Robust outlier, fast	Reduce variability, not match trends
MICE	Capture uncertainty, fit MAR	Complex, slow, needs assumptions
LOCF	Maintain temporal continuity	Unrealistic constant value assumptions

Source : (Research Results, 2025)

Table 4 shows that each imputation method has distinct characteristics that must be tailored to the analysis's needs. Mean, median, and mode are simple and fast. However, they can reduce data variation and are less suitable for complex patterns. MICE is more flexible and can capture uncertainty, but it requires more complex processes and extensive computations [26]. LOCF is effective at maintaining the continuity of time-series data, although the assumption of fixed values does not always reflect actual conditions [27]. Meanwhile, the KNN method imputes missing values by identifying the most similar data points and using the information to estimate missing entries [13],[14]. Thus, selecting imputation methods should consider the complexity of the data, the analysis purpose, and the trade-off between accuracy and computational efficiency.

Forecasting

The forecasting process in this study is shown in Figure 1. The stages include normalization of imputed data, hyperparameter optimization using PSO, and RNN model training [30]. These pipelines are designed to ensure the model produces accurate predictions based on the imputed data.

Source : (Research Results, 2025)
Figure 1. Forecasting Flow

Normalization aligns the ranges of variables so that scale differences are smaller and do not produce extreme values that can interfere with the model's learning process. One commonly used method is Min-Max normalization, which scales the data to the range 0 to 1 while preserving the original distribution and minimum and maximum values [31]. This method is particularly suitable for neural network models such as RNN that are sensitive to input scale, and the Min-Max normalization formula is shown in Equation (1).

$$x_t^* = \frac{x_t - x_{min}}{x_{max} - x_{min}} \quad (1)$$

In addition, the MinMax Scaler does not rely on any distributional assumptions, so it is flexible to use with different types of data, including time series data with clear upper and lower value limits. In Equation (1), x_t^* represents the normalized value at time t, while x_{min} and x_{max} denote the minimum and maximum values of the original data, respectively [32],[18]. This normalization ensures that all values fall within a uniform range, enabling the RNN model to learn temporal patterns more effectively.

After normalization, forecasting is performed using an RNN, a neural network for sequential data that uses information from the previous timestep via a recurrent connection. The RNN's internal memory allows it to remember past patterns and predict future values, effectively capturing temporal patterns in the time series [19],[34]. Its advantages include the ability to handle variable-length sequences, a simple architecture, efficient parameter sharing, and the ability to learn features directly from raw data. RNNs are suitable for forecasting and time-series classification [35]. The basic Equation is shown in Equations (2) and (3).

$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b_h) \quad (2)$$

$$o_t = W_o h_t + b_o \quad (3)$$

The first step in an RNN model is to compute a new hidden state using an activation function such as tanh, which combines the current input x_t with the previous hidden state h_{t-1}

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7938

[22],[23]. The hidden state h_{t-1}serves as internal memory, carrying information from the previous time step and enabling the network to capture temporal dependencies in sequential data. The imputed dataset is then divided into 80% training data and 20% testing data while maintaining temporal pattern consistency. During the training process, 20% of the training data is further allocated as validation data to monitor model performance and reduce the risk of overfitting. The RNN output is generated from the updated hidden state through weighted connections, producing a predicted value for the next time step. In this study, the forecasting task is formulated as a one-step-ahead prediction, where the model predicts the value at time $t+1$ based on the previous 24 observations using a sliding-window input structure.

The optimal parameters of the RNN model are determined using Particle Swarm Optimization (PSO). This population-based optimization method mimics the social behavior of a herd. In this approach, each particle represents a single candidate hyperparameter combination that moves within the search space, taking into account the best solution the particle has ever achieved (personal best) as well as the best solution found by the entire herd (global best) [38]. In this study, PSO is applied to optimize key hyperparameters, including the number of hidden units, learning rate, batch size, and dropout rate. The corresponding search space is presented in Table 5.

Table 5. Configure Hyperparameter Tuning with PSO

Hyperparameter	Search Space	Notes
Hidden Layers	2 - 10	Explored across multiple architectures
Neurons per Layer	1 - 252	Range explored during optimization
Activation Function	{tanh, ReLU, sigmoid, Linear}	Stable for time-series forecasting
Optimizer	{Adam, RMSprop, SGD}	Selected by PSO for convergence
Loss Function	{MSE, MAE}	Minimized forecasting error
Batch Size	10 - 64	Provides stable training on the dataset
Epochs	10 - 100	Tuned to avoid under-/overfitting
Dropout	0.0 – 0.5	To reduce overfitting
PSO Parameters	swarm=10, generations=5, velocity [0.1], pbest=1.5, gbest=2.0	Process optimization

Source : (Research Results, 2025)

Table 5 presents the predefined ranges for each hyperparameter, enabling PSO to systematically explore the search space and identify the best combination that balances prediction accuracy and training stability while minimizing overfitting [39]. To prevent data leakage, the optimization process is conducted exclusively on the training data, with a portion allocated as a validation set to evaluate candidate solutions generated by the particle swarm. The test data is reserved solely for the final evaluation. Additionally, PSO configurations, such as swarm size and number of generations, are determined to guide the optimization process more effectively [40].

Forecasting experiments were implemented in Python using Google Colab as the development environment. Several libraries were used, including NumPy, Pandas, Matplotlib, and TensorFlow/Keras for data processing, visualization, and RNN model development. To ensure the stability of the results, the experiment was conducted five times independent to reduce the influence of random factors. The results of the reported evaluation performance are the average of the total of the experiment. Overall, this optimization framework results in an optimal hyperparameter configuration, as presented in Tables 6, 7, and 8, which summarize the results of the PSO-based tuning process in each dataset.

Table 6. Dataset 1 Tuning Results

Parameters	Drop Missing Value	KNN	LOCF	MEAN	MEDIAN	MICE	MODE
Hidden Layer	4	6	2	3	7	6	2
Neurons	43	247	125	39	40	43	219
Activation Function	Linear	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Loss Function	MSE	MSE	MSE	MSE	MSE	MSE	MSE
Epochs	18	37	65	67	40	83	96
Batch Size	16	21	20	16	60	50	39

Source : (Research Results, 2025)

Table 7. Dataset 2 Tuning Results

Parameters	Drop Missing Value	KNN	LOCF	MEAN	MEDIAN	MICE	MODE
Hidden Layer	7	4	2	6	5	8	3
Neurons	79	246	33	30	252	174	84
Activation Function	Linear	ReLU	Tanh	ReLU	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	RMSprop	Adam	RMSprop	RMSprop	Adam
Loss Function	MSE	MSE	MSE	MSE	MSE	MSE	MSE
Epochs	45	97	74	33	48	49	13
Batch Size	51	42	51	56	56	22	40

Source : (Research Results, 2025)

Table 8. Dataset 3 Tuning Results

Parameters	Drop Missing Value	KNN	LOCF	MEAN	MEDIAN	MICE	MODE
Hidden Layer	6	4	7	3	4	6	2
Neurons	161	123	154	253	39	43	219
Activation Function	Linear	Tanh	Linear	ReLU	Linear	ReLU	ReLU
Optimizer	RMSprop	Adam	RMSprop	Adam	RMSprop	Adam	Adam
Loss Function	MSE	MSE	MSE	MSE	MSE	MSE	MSE
Epochs	72	12	72	59	25	83	96
Batch Size	39	53	48	34	27	50	39

Source : (Research Results, 2025)

Evaluation Metrics

The performance of the forecasting model in this study was evaluated using three main metrics, namely Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). MAPE measures the relative accuracy of a prediction by calculating the percentage deviation between the actual value and the prediction. The value is calculated as the average absolute error percentage [41]. Formulated in Equation (4).

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{Y_i-\hat{Y}_i}{Y_i}\right| \times 100\% \quad (4)$$

Where m is the total number of observations, Y_i represents the actual value at observation i, and $\hat{Y}_i$ denotes the predicted value at observation i [42]. Thus, MAPE provides a clear picture of the model's relative error rate as a percentage, making it easier to evaluate the model's predictive performance. The RMSE measures the magnitude of the prediction error in the original data unit by assigning a greater penalty to large errors [43]. While R^2 showing the proportion of actual value variations that the model successfully explained. The RMSE formula is shown in Equation (5).

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=i}^{m}\left(Y_i - \hat{Y}_i\right)^2} \quad (5)$$

and the formula R^2 can be seen in Equation (6).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} \quad (6)$$

with $\bar{Y}$ as the average of the actual value and n amount of data. The RMSE provides an average measure of absolute error, while R^2 it assesses the model's ability to explain the overall variation in the data [44]. The combination of these three metrics helps provide a more comprehensive evaluation of the quality and reliability of prediction models.

RESULTS AND DISCUSSION

The results of the evaluation of forecasting performance on three datasets that have undergone imputation and been analyzed using the RNN model are presented to assess the accuracy of the predictions, the magnitude of the errors, and the model's ability to explain the data's variability. Evaluations were conducted using MAPE, RMSE, and R^2, each of which measured relative prediction errors, the magnitude of forecast errors, and the proportion of variance the model could explain. The quantitative results for these metrics are summarized in Tables 9, 10, and 11, enabling a direct comparison of forecasting performance across datasets and imputation methods. These results serve as a basis for analyzing error patterns, prediction consistency, and the influence of dataset characteristics on the RNN model's forecasting performance, thereby enabling a more in-depth interpretation of the model's predictive capabilities.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7938

Table 9. Evaluation of MAPE Results

Imputation Method	DATASET 1	DATASET 2	DATASET 3
Drop Missing Values	2.04318	3.29988	1.43034
Mean	1.39303	7.13235	1.69789
Mode	0.94855	2.47668	1.21057
Median	0.93374	6.29766	1.36252
MICE	1.76192	6.57207	1.07349
LOCF	0.97059	1.08567	1.38796
KNN	1.20815	2.65974	0.66881

Source : (Research Results, 2025)

Table 9 presents the results of the evaluation of forecasting performance based on MAPE for three datasets imputed using various methods. MAPE is used to assess the relative deviation between the predicted and actual values, providing a percentage measure of the model's prediction accuracy. However, MAPE is sensitive to very small or near-zero actual values because they appear in the denominator of the error calculation, leading to very large or unstable error values. In the three datasets used in this study, zero or near-zero values were rare, and most values fell within a stable range. Therefore, the use of MAPE remains appropriate for evaluating the overall prediction accuracy of this dataset. In Dataset 1, which has a seasonal pattern with sharp fluctuations and MNAR missing values, Median (0.9337) and Mode (0.9486) produce the lowest MAPE, followed by LOCF (0.9706). The Mean (1.3930) and MICE (1.7619) methods yielded higher errors, while missing value yielded the highest MAPE (2.0432). These results show that methods that maintain temporal continuity and distribution center values allow the model to capture data peaks and valleys effectively.

In Dataset 2, which has high fluctuations and MCAR missing values, LOCF (1.0857) performed best, followed by Mode (2.4767) and KNN (2.6597). Median (6.2977), MICE (6.5721), and Mean (7.1324) resulted in much higher errors, whereas missing value had a MAPE of 3.2999. These results emphasize the importance of adapting imputation methods to the characteristics of the dataset, where temporal continuity and recognition of local structures can improve prediction accuracy on data with high variation. In Dataset 3, which has a low proportion of missing values and a relatively stable data pattern, KNN produces the lowest MAPE (0.6688), followed by MICE (1.0735) and Mode (1.2106). LOCF (1.3880) is suboptimal because temporal continuity has a limited effect on stable data, whereas deleting missing data yields a MAPE of 1.4303. This suggests that using local correlation can improve prediction accuracy on stable datasets, even when the number of missing values is small.

Table 10. Evaluation of RMSE Results

Imputation Method	Dataset 1	Dataset 2	Dataset 3
Drop Missing Values	3.66956	20.02948	5.62618
Mean	4.30278	67.46023	9.45966
Mode	5.17559	21.72579	2.76761
Median	4.47247	65.48591	4.27161
MICE	4.92567	65.9768	9.52263
LOCF	3.55282	18.93884	10.05797
KNN	12.0311	42.26496	2.21874

Source : (Research Results, 2025)

Table 10 presents the results of the evaluation of forecasting performance based on RMSE for three datasets imputed using various methods. RMSE measures the magnitude of prediction errors in the original units of the data, providing information about the absolute deviation between the predicted and actual values. In general, the results show that the effectiveness of imputation methods varies across datasets and that some imputation methods produce lower errors than drop-missing-values strategies. In Dataset 1, which has a seasonal pattern with fairly sharp fluctuations, the LOCF method produces the lowest RMSE value (3.5528), followed by Drop Missing Values (3.6696) and Mean (4.3028). Meanwhile, the Median (4.4725) and Mode (5.1756) showed a fairly good performance but were slightly higher, while MICE (4.9257) resulted in greater errors. The KNN method yielded the highest RMSE (12.0311), suggesting it is less appropriate for datasets with strong seasonal fluctuations and extreme variation. These results suggest that imputation methods that maintain temporal continuity, such as LOCF, are more effective at preserving the time-series structure, thereby reducing prediction errors.

In Dataset 2, which has high variation and an MCAR missing-value mechanism, the LOCF method again shows the best performance, with the lowest RMSE (18.9388), followed by Drop Missing Values (20.0295) and Mode (21.7258). In contrast, KNN (42.2650) produces much higher errors, while Median (65.4859), MICE (65.9768), and Mean (67.4602) produce the most RMSE values. These results suggest that maintaining temporal continuity is essential in high-variation datasets, as methods that preserve temporal structure help models better track data trends. In Dataset 3, which has a relatively stable data pattern and a small proportion of missing values, the KNN method produces the lowest RMSE (2.2187), followed by Mode (2.7676) and Median (4.2716). The Drop Missing Values method yields an RMSE of 5.6262, while the Mean (9.4597) and MICE (9.5226) methods show higher errors. The LOCF method yields the highest RMSE (10.0580), which indicates

that temporal continuity has a limited influence on datasets with stable patterns. This suggests that methods that can exploit local correlations among values, such as KNN, can improve prediction accuracy on relatively stable datasets.

Table 11. Evaluation of Results R^2

Imputation Method	Dataset 1	Dataset 2	Dataset 3
Drop Missing Values	0.90807	0.89394	0.64314
Mean	0.87181	-0.07806	-0.00948
Mode	0.81905	0.88819	0.91361
Median	0.86149	-0.01588	0.79416
MICE	0.83202	-0.03116	-0.2001
LOCF	0.91478	0.91503	-0.14073
KNN	-0.00219	0.57684	0.94501

Source : (Research Results, 2025)

Table 11 presents the results of the evaluation of forecasting performance, measured by R^2, for three datasets imputed using various methods. R^2 is used to measure the extent to which the model explains variation in actual values through its predictive results. The R^2 value ranges from 0 to 1, where a higher value indicates the model's better ability to capture the relationships among variables. In contrast, a negative value indicates that the model fails to fit the data. In contrast to MAPE and RMSE, which focus on the magnitude of prediction errors, R^2 provides an overall measure of the model's ability to explain the data's variability. In Dataset 1, which has a seasonal pattern with quite sharp fluctuations and an MNAR mechanism for missing values, the LOCF method yields the highest R^2 value (0.91478), followed by Drop Missing Values (0.90807) and Mean (0.87181). Meanwhile, Median (0.86149), MICE (0.83202), and Mode (0.81905) show slightly lower performance. The KNN method yielded a negative R^2 value (-0.00219), indicating that it was less able to capture temporal patterns in datasets with strong seasonal fluctuations. These results suggest that imputation methods that maintain temporal continuity, such as LOCF, are more effective at helping models capture temporal variation in time series with seasonal patterns.

In Dataset 2, which has high variation and an MCAR missing-value mechanism, the LOCF method again shows the best performance, with the highest R^2 value (0.91503), followed by Drop Missing Values (0.89394) and Mode (0.88819). The KNN method yielded an R^2 of 0.57684, indicating moderate ability to explain the data's variation. In contrast, Median (-0.01588), MICE (-0.03116), and Mean (-0.07806) yielded negative R^2 values, suggesting that the model was unable to effectively explain the data's variability when using these imputation methods. This suggests that maintaining temporal continuity is essential in datasets with high variation. In Dataset 3, which had a relatively stable data pattern and a small proportion of missing values, the KNN method yielded the highest R^2 value (0.94501), followed by Mode (0.91361) and Median (0.79416). The Drop Missing Values method yields an R^2 value of 0.64314, while Mean (-0.00948) and MICE (-0.20010) yield negative values, indicating poor model performance. The LOCF method yielded an R^2 value of -0.14073, indicating that a temporal-continuity-based approach is less effective in datasets with relatively stable patterns. These findings suggest that methods that can leverage local correlations among values, such as KNN, can improve the model's ability to explain data variation in stable datasets. The imputation method has been proven to be able to increase the accuracy of prediction compared to the use of data that still contains missing value. Overall, comparative evaluations using MAPE, RMSE, and R^2 showed that no single imputation method consistently outperformed other methods across all evaluation datasets and metrics.

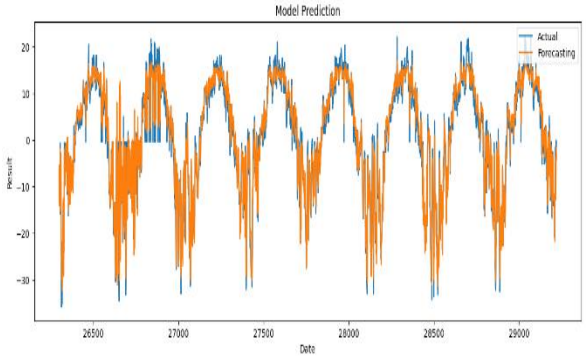
In datasets with high variability and temporal dependency (Datasets 1 and 2), LOCF showed strong performance primarily based on RMSE and R^2, indicating its ability to maintain temporal patterns of data. However, in Dataset 1, the Median method yielded the best MAPE values, suggesting that a central tendency-based approach can provide a lower percentage error under certain conditions. In contrast, KNN showed a more consistent advantage across all metrics on a stable dataset (Dataset 3), highlighting the importance of local correlations under such conditions. These results show that the effectiveness of the imputation method is highly dependent on the characteristics of the selected dataset and evaluation metrics. Therefore, the selection of the right method should be based on a balanced consideration of various performance indicators rather than relying on a single metric or assuming the best approach universally.

Figure 2 illustrates the optimized forecasting performance of the RNN model on Dataset 1 using Median imputation, with precise hyperparameter settings as described earlier in Table 6. The actual observations (solid blue line) and predicted values (solid orange line) are plotted across the temporal index (x-axis: 26,000–29,000) and the target variable values (y-axis: −40 to 25). The predictions closely track the recurring seasonal patterns and periodic fluctuations, aligning with the strong

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7938

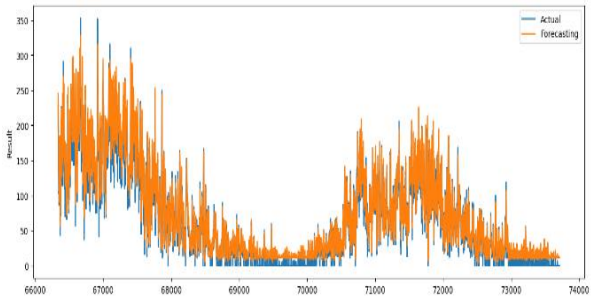
quantitative metrics achieved (RMSE: 3.5528, MAPE: 0.9706, R^2: 0.9148).



Source : (Research Results, 2025)

Figure 2. Comparison Graph of Actual Versus Predicted Values for Dataset 1 using the Median Imputation Method.

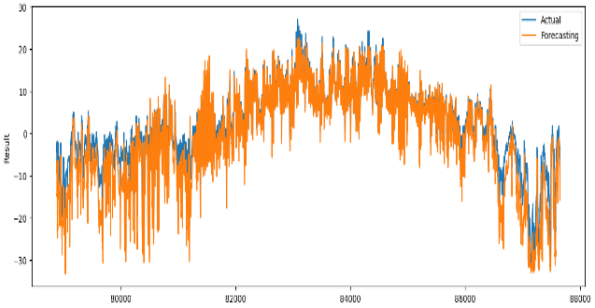
This close overlap confirms that the model effectively learns the underlying temporal structure of the time series. However, an underestimation effect remains visible at extreme points, particularly during sharp declines below −30. In these instances, the predicted values appear smoother than the actual abrupt fluctuations, which contributes to the remaining forecasting errors reflected in the RMSE and MAPE values despite the model's overall strong performance in capturing general temporal dynamics.



Source : (Research Results, 2025)

Figure 2. Comparison Graph of Actual Versus Predicted Values for Dataset 2 using the LOCF Imputation Method.

Figure 3 illustrates the forecasting performance of the optimized RNN model on Dataset 2 using LOCF imputation, utilizing the optimal configuration presented in Table 7. The actual observations (solid blue line) and predicted values (solid orange line) are plotted across the temporal index (x-axis: 66,000–74,000) and the target variable values (y-axis: 0 to 350). The forecasting results successfully follow the overall trend of the actual data, particularly during the sharp decline phase and the subsequent increase after reaching the minimum region. This visual consistency aligns with the strong quantitative metrics achieved (RMSE: 18.9388, MAPE: 1.0857, R^2: 0.9150), indicating that the model effectively explains more than 91% of the variability in Dataset 2. However, deviations remain visible in highly fluctuating regions, especially around the observation ranges of 66,000–68,000 and 71,000–72,000, where both overestimation and underestimation occur. Additionally, the predicted values appear smoother than the actual observations at values close to zero, suggesting reduced sensitivity to sudden local fluctuations and noise. These limitations contribute to the remaining forecasting errors despite the model's overall strong ability to capture the dominant temporal patterns.



Source : (Research Results, 2025)

Figure 3. Comparison Graph of Actual Versus Predicted Values for Dataset 3 using the KNN Imputation Method.

Figure 4 illustrates the forecasting performance of the optimized RNN model on Dataset 3 using KNN imputation, based on the specific network architecture summarized in Table 8. The actual observations (solid blue line) and predicted values (solid orange line) are plotted across the temporal index (x-axis: 79,000–90,000) and the target variable values (y-axis: −35 to 30). The forecasting results closely follow the general trend of the actual data, particularly during the gradual transition from negative to positive values, the peak region, and the subsequent decline toward the end of the observation period. This visual alignment is supported by the strong quantitative metrics achieved (RMSE: 2.2187, MAPE: 0.6688, R^2: 0.9450), indicating very small prediction errors and demonstrating that the model explains approximately 94.5% of the variability in the dataset. The close overlap between the actual and predicted series confirms that the model effectively captures the main temporal structure and trend of relatively stable time-series data. However,

deviations remain visible during extreme drops below −20, where the predicted values appear smoother and less extreme than the actual observations, indicating an underestimation effect. These discrepancies contribute to the remaining forecasting errors despite the model's overall strong performance in representing the dominant temporal dynamics.

All three figures show that RNN performance is influenced by the characteristics of the dataset, particularly the degree of variability and temporal structure, in addition to the imputation method used. Median imputation showed good performance in Dataset 1 by preserving the overall temporal pattern. LOCF was able to follow the temporal continuity and dominant trend changes in Dataset 2, while KNN produced good forecasting results in Dataset 3, which has relatively lower variability and a more stable distribution. The main pattern identified is that increased data variability correlates with an increase in absolute errors, especially reflected in RMSE, although R^2 values can remain relatively comparable. In contrast, datasets with low variability allow for smaller prediction errors and better model suitability. This suggests that the effectiveness of imputation strategies is contextual, relying on the interaction between data variability, temporal dependence, and missing value characteristics.

From a theoretical perspective, this study contributes to the literature by showing that the effectiveness of imputation techniques is not universal but depends on the interaction between the missing data mechanisms (MCAR, MAR, MNAR) and the properties of the dataset. By integrating imputation evaluation with forecasting performance within the RNN-PSO framework, the study offers a more comprehensive perspective on how the treatment of missing data affects deep learning-based time series prediction in climatology. This research has several limitations. First, the imputation approach used is still univariate so that it does not fully capture the complexity of multivariate climate systems. Second, the spatial dependence between observation sites has not been considered, even though approaches such as kriging have the potential to increase imputation accuracy.

The main contribution of this research lies in the provision of an integrated evaluation framework that links imputation strategies, data loss mechanisms, and forecasting performance on various dataset characteristics. This approach provides a more systematic perspective in understanding the relationship between data quality and model performance in environmental time series analysis.

CONCLUSION

This study shows that the effectiveness of imputation methods on RNN-PSO forecasting depends on the characteristics of the dataset, the pattern of missing values, and the evaluation metrics. There is no single imputation method that is consistently best across the entire dataset. LOCF provides strong performance on datasets with high temporal dependencies (Datasets 1 and 2), especially on RMSE and R^2, while Median produces the best MAPE on Dataset 1. In the more stable dataset (Dataset 3), KNN provided the most consistent results across evaluation metrics. These findings confirm that the selection of imputation methods needs to be adjusted to the characteristics of the data to improve the accuracy of climatological time series forecasting. Further research can develop multivariate imputation to capture more complex variable interactions, as well as explore hybrid methods that combine simple approaches and deep learning-based techniques or spatial-temporal models. Additional research directions include determining the effective threshold for missing proportions, developing adaptive imputation based on real-time data, evaluating computational efficiency for large-scale climate data, and validating across various climate regions and extreme conditions to ensure model generalization and support more accurate early warning systems.

REFERENCE

[1] N. T. Luchia, E. Tasia, I. Ramadhani, A. Rahmadeyan, and R. Zahra, "Performance Comparison Between Artificial Neural Network, Recurrent Neural Network and Long Short-Term Memory for Prediction of Extreme Climate Change," *Public Research Journal of Engineering, Data Technology and Computer Science*, no. Issue 2, Jan. 2024, Accessed: Dec. 06, 2025. [Online]. Available: https://journal.irpi.or.id/index.php/predatecs

[2] L. E. Alejo-Sanchez *et al.*, "Missing data imputation of climate time series: A review," Dec. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.mex.2025.103455.

[3] T. M. Pham, N. Pandis, and I. R. White, "Missing data: Issues, concepts, methods," *Semin. Orthod.*, vol. 30, no. 1, pp. 37–44, Feb. 2024, doi: 10.1053/j.sodo.2024.01.007.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7938

[4] L. Munkhdalai, T. Munkhdalai, V. H. Pham, M. Li, K. H. Ryu, and N. Theera-Umpon, "Recurrent Neural Network-Augmented Locally Adaptive Interpretable Regression for Multivariate Time-Series Forecasting," *IEEE Access*, vol. 10, pp. 11871–11885, 2022, doi: 10.1109/ACCESS.2022.3145951.

[5] S. H. Lim, "Understanding Recurrent Neural Networks Using Nonequilibrium Response Theory," 2021. [Online]. Available: http://jmlr.org/papers/v22/20-620.html.

[6] E. A. Nketiah, L. Chenlong, J. Yingchuan, and S. A. Aram, "Recurrent neural network modeling of multivariate time series and its application in temperature forecasting," *PLoS One*, vol. 18, no. 5 May, May 2023, doi: 10.1371/journal.pone.0285713.

[7] M. K. Bhatia and V. Bhatt, "Forecasting Time Series Data using Recurrent Neural Networks: A Systematic Review," *Journal for Research in Applied Sciences and Biotechnology*, vol. 3, no. 6, pp. 184–189, Dec. 2024, doi: 10.55544/jrasb.3.6.22.

[8] A. Turner, "Eighty Years of Canadian Climate Data," Kaggle. Accessed: Oct. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/aturner374/eighty-years-of-canadian-climate-data

[9] Abhinand, "Daily Sunspot Data (1818–2019)," Kaggle. Accessed: Oct. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/abhinand05/daily-sun-spot-data-1818-to-2019

[10] N. Tricky, "Sofia Weather Records," Kaggle. Accessed: Oct. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/nikitricky/sofia-weather-records

[11] M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, "Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review," *BMC Med. Res. Methodol.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12874-024-02310-6.

[12] M. W. Heymans and J. W. R. Twisk, "Handling missing data in clinical research," *J. Clin. Epidemiol.*, vol. 151, pp. 185–188, Nov. 2022, doi: 10.1016/j.jclinepi.2022.08.016.

[13] T. M. Pham, N. Pandis, and I. R. White, "Missing data, part 2. Missing data mechanisms: Missing completely at random, missing at random, missing not at random, and why they matter," Jul. 01, 2022, *Elsevier Inc.* doi: 10.1016/j.ajodo.2022.04.001.

[14] M. Yan, L. Zhou, C. Zhao, C. Shi, and C. Ou, "Comparison of different approaches in handling missing data in longitudinal multiple-item patient-reported outcomes: a simulation study," *Health Qual. Life Outcomes*, vol. 23, no. 1, Dec. 2025, doi: 10.1186/s12955-025-02364-0.

[15] R. J. Little, "Annual Review of Statistics and Its Application Missing Data Assumptions," vol. 27, p. 23, 2026, doi: 10.1146/annurev-statistics-040720.

[16] M. W. , & T. J. W. R. Heymans, "Handling missing data in clinical research.," *J. Clin. Epidemiol.*, 2022, doi: https://doi.org/10.1016/j.jclinepi.2022.04.017.

[17] B. Gomer and K. H. Yuan, "A Realistic Evaluation of Methods for Handling Missing Data When There is a Mixture of MCAR, MAR, and MNAR Mechanisms in the Same Dataset," *Multivariate Behav. Res.*, vol. 58, no. 5, pp. 988–1013, 2023, doi: 10.1080/00273171.2022.2158776.

[18] Y. Zhang and P. J. Thorburn, "Handling missing data in near real-time environmental monitoring: A system and a review of selected methods," *Future Generation Computer Systems*, vol. 128, pp. 63–72, Mar. 2022, doi: 10.1016/j.future.2021.09.033.

[19] A. Gabrio, C. Plumpton, S. Banerjee, and B. Leurent, "Linear mixed models to handle missing at random data in trial-based economic evaluations," *Health Economics (United Kingdom)*, vol. 31, no. 6, pp. 1276–1287, Jun. 2022, doi: 10.1002/hec.4510.

[20] P. Buczak, J. J. Chen, and M. Pauly, "Analyzing the Effect of Imputation on Classification Performance under MCAR and MAR Missing Mechanisms," *Entropy*, vol. 25, no. 3, Mar. 2023, doi: 10.3390/e25030521.

[21] J. H. Li *et al.*, "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Med. Res. Methodol.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12874-024-02173-x.

[22] K. Seu, M.-S. Kang, and H. Lee, "INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION An Intelligent Missing Data Imputation Techniques: A Review," May 2022. [Online]. Available: www.joiv.org/index.php/joiv

[23] P. Buczak, J. J. Chen, and M. Pauly, "Analyzing the Effect of Imputation on Classification

Performance under MCAR and MAR Missing Mechanisms," *Entropy*, vol. 25, no. 3, Mar. 2023, doi: 10.3390/e25030521.

[24] M. Alwateer, E.-S. Atlam, M. M. A. El-Raouf, O. A. Ghoneim, and I. Gad, "Missing Data Imputation: A Comprehensive Review," *Journal of Computer and Communications*, vol. 12, no. 11, pp. 53–75, 2024, doi: 10.4236/jcc.2024.1211004.

[25] J. H. Li *et al.*, "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Med. Res. Methodol.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12874-024-02173-x.

[26] L. O. Joel, W. Doorsamy, and B. S. Paul, "A comparative study of imputation techniques for missing values in healthcare diagnostic datasets," *Int. J. Data Sci. Anal.*, vol. 20, no. 7, pp. 6357–6373, Nov. 2025, doi: 10.1007/s41060-025-00825-9.

[27] Y. Wang *et al.*, "Research on Missing Value Imputation to Improve the Validity of Air Quality Data Evaluation on the Qinghai-Tibetan Plateau," *Atmosphere (Basel).*, vol. 14, no. 12, Dec. 2023, doi: 10.3390/atmos14121821.

[28] S. M. Memon, R. Wamala, and I. H. Kabano, "A comparison of imputation methods for categorical data," *Inform. Med. Unlocked*, vol. 42, Jan. 2023, doi: 10.1016/j.imu.2023.101382.

[29] F. A. Tyas, M. Setianama, R. F. Fajriyah, and A. Ilham, "Implementation of Particle Swarm Optimization (PSO) to Improve Neural Network Performance in Univariate Time Series Prediction," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, no. No. 4, Nov. 2021.

[30] M. L. Lin, C. W. Tsai, and C. K. Chen, "Daily maximum temperature forecasting in changing climate using a hybrid of Multi-dimensional Complementary Ensemble Empirical Mode Decomposition and Radial Basis Function Neural Network," *J. Hydrol. Reg. Stud.*, vol. 38, Dec. 2021, doi: 10.1016/j.ejrh.2021.100923.

[31] M. M. Hasan, M. J. Hasan, and P. B. Rahman, "Comparison of RNN-LSTM, TFDF and stacking model approach for weather forecasting in Bangladesh using historical data from 1963 to 2022," *PLoS One*, vol. 19, no. 9 September, Sep. 2024, doi: 10.1371/journal.pone.0310446.

[32] Peshawa J. Muhammad Ali, "Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements," 2021, Accessed: Dec. 06, 2025. [Online]. Available: (doi:10.14500/aro.10955)

[33] F. Masri, D. Saepudin, and D. Adytia, "Forecasting of Sea Level Time Series using Deep Learning RNN, LSTM, and BiLSTM, Case Study in Jakarta Bay, Indonesia."

[34] H. Song and H. Choi, "Forecasting Stock Market Indices Using the Recurrent Neural Network Based Hybrid Models: CNN-LSTM, GRU-CNN, and Ensemble Models," *Applied Sciences (Switzerland)*, vol. 13, no. 7, Apr. 2023, doi: 10.3390/app13074644.

[35] L. Larasati, S. Saadah, and P. E. Yunanto, "Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) Methods to Forecast Daily Turnover at BM Motor Ngawi," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 7, no. 1, p. 141, Jan. 2024, doi: 10.24014/ijaidm.v7i1.27643.

[36] M. K. Bhatia and V. Bhatt, "Forecasting Time Series Data using Recurrent Neural Networks: A Systematic Review," *Journal for Research in Applied Sciences and Biotechnology*, vol. 3, no. 6, pp. 184–189, Dec. 2024, doi: 10.55544/jrasb.3.6.22.

[37] A. M. Assaf, H. Haron, H. N. Abdull Hamed, F. A. Ghaleb, S. N. Qasem, and A. M. Albarrak, "A Review on Neural Network Based Models for Short Term Solar Irradiance Forecasting," Jul. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app13148332.

[38] S. Khan *et al.*, "Optimizing load demand forecasting in educational buildings using quantum-inspired particle swarm optimization (QPSO) with recurrent neural networks (RNNs):a seasonal approach," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-04301-z.

[39] D. Kilichev and W. Kim, "Hyperparameter Optimization for 1D-CNN-Based Network Intrusion Detection Using GA and PSO," *Mathematics*, vol. 11, no. 17, Sep. 2023, doi: 10.3390/math11173724.

[40] H. S. Salem, M. A. Mead, and G. S. El-Taweel, "Particle Swarm Optimization-Based Hyperparameters Tuning of Machine Learning Models for Big COVID-19 Data Analysis," *Journal of Computer and*

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7938

Communications, vol. 12, no. 03, pp. 160–183, 2024, doi: 10.4236/jcc.2024.123010.

[41] A. Yunita *et al.*, "Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models," Dec. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.mex.2025.103462.

[42] W. H. Lin, P. Wang, K. M. Chao, H. C. Lin, Z. Y. Yang, and Y. H. Lai, "Wind power forecasting with deep learning networks: Time-series forecasting†," *Applied Sciences (Switzerland)*, vol. 11, no. 21, Nov. 2021, doi: 10.3390/app112110335.

[43] Y. Zhang and P. J. Thorburn, "Handling missing data in near real-time environmental monitoring: A system and a review of selected methods," *Future Generation Computer Systems*, vol. 128, pp. 63–72, Mar. 2022, doi: 10.1016/j.future.2021.09.033.

[44] V. Kramar and V. Alchakov, "Time-Series Forecasting of Seasonal Data Using Machine Learning Methods," *Algorithms*, vol. 16, no. 5, May 2023, doi: 10.3390/a16050248.