Journal of King Saud University - Computer and Information Sciences 35 (2023) 101816

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

1D Convolutional LSTM-based wind power prediction integrated with *PkNN* data imputation technique

Farah Shahid [a,b], Atif Mehmood [a,b,*], Rizwan Khan [a], Ahmad AL Smadi [c], Muhammad Yaqub [d], Mutasem K. Alsmadi [e], Zhonglong Zheng [a,*]

[a] *Department of Computer Science and Technology, Zhejiang Normal University, Jinhua 321002, China*
[b] *Zhejiang Institute of Photoelectronics & Zhejiang Institute for Advanced Light, Source, Zhejiang Normal University, Jinhua, Zhejiang 321004, China*
[c] *Department of Data Science and Artificial Intelligence, Zarqa University, Zarqa 13100, Jordan*
[d] *Faculty of Information Technology, Beijing University of Technology, 100024, Beijing, China*
[e] *Department of Management Information Systems, College of Applied Studies and Community Service, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia*

ARTICLE INFO

Keywords:
Missing data imputation
K values selection
Meterological variables
Patterned observation
KNN
LSTM

ABSTRACT

Various supervised machine-learning algorithms for wind power forecasting have been developed in recent years to manage wind power fluctuations and effectively correlate to energy consumption; Meanwhile, the performance of the model does suffer from missing values. To address the issue of missing values in wind power forecast, this paper proposes two methods: Clue-based missing at random (*CMAR*) and patterned k-nearest Neighbor (*PkNN*). In addition, a hybrid wind energy forecasting system has been created that is built on 1D-Convolutional neural networks, which are used to extract features from raw input, and Long Short-Term Memory, which employs time series data internal representation learning to improve the accuracy of month-wise wind power forecasting. The efficacy of the proposed model on generated datasets is also compared to the classic machine learning model to check the generalization ability. The strength of the Convolutional LSTM has been estimated in terms of different performance metrics such as mean absolute error (MAE), root mean squared error (RMSE), R^2, and explained variance score (EVS). The experimental results show that the use of the *PkNN* algorithm for data imputation; integrated with regression-based Convolutional LSTM is much more efficient in prediction over other deep neural network models.

1. Introduction

Wind power as a renewable source of energy is one of the rapidly developing techniques for providing electricity worldwide. It is an environmentally friendly method of increasing economic profit and sustainability efficiency (Liao et al., 2021). Wind turbines, unlike traditional energy sources such as fossil fuels, coal, and natural gas, provide pollution-free electricity; hence, wind power is always accessible. Furthermore, the fluctuation of wind energy output has a significant influence on the operation of power systems (Sun et al., 2016), and reliable forecasts are essential for making predictions. Wind farm power generation systems are vulnerable to fluctuations and abnormalities, which complicates the administration and planning of the power system. (Tawn et al., 2020). To ensure the safety and stability of the electricity system, historical data collected by the supervisory control and data acquisition (SCADA) system must be utilized to accurately predict the output power of the wind farms. However, the gathered data might be missing certain information since the SCADA system is frequently disrupted by a variety of occurrences, including sensor breakdown, cyber-attacks, and transmission latency (Sun et al., 2021). Consequently, the

Abbreviations: ARIMA, Autoregressive integrated moving average; SCADA, Supervisory control and data acquisition; CNN, Convolutional neural network; GRU, Gated recurrent unit; NN, Neural network; SVM, Support vector machine; ML, Machine learning; MNAR, Missing not at random; LSTM, Long short-term memory; NLP, Natural language processing; SAE-CD, sparse autoencoder and a coordinate descendant optimization; AR, Auto regression; MAR, Missing at random; MCAR, Autoregressive; MLP, Multilayer perceptron; KNN, K-Nearest Neighbor; R^2, coefficient of determination; MSE, Mean squared error; EVS, Explained variance score; MAE, Mean absolute error; RMSE, Root mean square error; ANN, Artificial neural network.

Peer review under responsibility of King Saud University.

* Corresponding authors.
E-mail addresses: atifmeh@zjnu.edu.cn (A. Mehmood), zhonglong@zjnu.edu.cn (Z. Zheng).

https://doi.org/10.1016/j.jksuci.2023.101816
Received 13 July 2023; Received in revised form 8 October 2023; Accepted 26 October 2023
Available online 2 November 2023

imputation of missing data for wind farms is crucial for wind power forecasts.

The challenges such as management, maintenance, and energy usage can be resolved by carefully doing data analysis on the information that has been acquired. Hence, data analytics is employed to predict the demand for electricity at different times, including short-term, medium-term (Al-Musaylh et al., 2018), and long-term (Heinermann and Kramer, 2016). These assessments may also improve policy development and assist in comprehending electric power loads and energy-saving performance (Bigerna et al., 2019; Wang et al., 2018). Three distinct missing data frameworks—MCAR, MAR, and MNAR—are commonly employed in data analysis (Doretti et al., 2018). In the case of MCAR, the likelihood of a data record being missing is independent of any other factors in the dataset. Here, the remaining complete information mirrors the original population's distribution without any missing data, ensuring unbiased model estimation. For MAR data, the absence of one variable depends on the value of another, but not on its specific value. On the other hand, MNAR missing patterns are contingent on the value required to produce a missing point. Neglecting this gap in a model study can lead to skewed results, as the distribution of values in the remaining observed cases differs from that in the missing data points. Among these, MCAR can be more useful for wind power missing values, although MAR missing information is caused by implied terms such infrastructure issues or poor communication, both of which are environmental. Missing measurements are one of the challenges in acquiring temperature data from sensors (Wang et al., 2021). Several studies have shown that when there are missing values in the data, the performance of the model in making predictions suffers.

There are several approaches to handling missing data, with different levels of effectiveness depending on the type of missing data. However, data imputation methods for wind farm meteorological data have recently been used in research. Mittel *et* al. (Mittal et al., 2019) implemented K-Nearest Neighbor (KNN) interpolation for missing data in diagnostic analysis, whereas Jing *et* al. used average interpolation for missing data in hydro-meteorology time series modeling (Jing et al., 2022). Iterative imputation was employed to analyze the data collected on the Cimandiri River for water quality monitoring (Sudriani et al., 2019). Another study demonstrated that zero restoration has been superior to KNN imputation, average imputation, and recursive imputation when it came to the prediction of cardiovascular disease, hypertension, and diabetes (Zhang et al., 2014). In general, there is a research opportunity to show that the KNN imputation strategy is better than other imputation techniques for estimating the duration of the rain (Oktaviani and Putrada, 2022). There are numerous examples of wind power datasets that highlight the data imputation in the literature, including meter reading data (Norazian et al., 2008), electrical energy data, photovoltaic data (Jung et al., 2020), and air quality data (Peppanen et al., 2016). Imputation of missing values for temporal data is a topic covered in some articles similar to the one above. Statistical techniques like autoregressive integrated moving average (ARIMA), linear extrapolation, and regression, as well as machine learning algorithms like k-nearest neighbor (K-NN), multilayer perceptron (MLP), and support vector regression (SVR), are all used in these studies as lost values imputation techniques (Yeh et al., 2019). Whereas, Zheng *et* al. examined two related regression techniques, examining K-NN and linear extrapolation in particular. However, the problem of missing data imputation has not been extensively investigated in the area of electric wind power data (Zheng et al., 2019).

Over recent years, massive amounts of temporal data have been collected using tools like sensors and electronic meters. These observations are gathered at regular intervals, such as hourly, daily, weekly, monthly, or yearly, which is known as a time series. Numerous sectors, including finance, environment, energy, and meteorology, have given time series prediction a lot of attention. In the same way, accurate wind power prediction is necessary for the design and reliable operation of power grids to provide a continuous supply (Shahid et al., 2021). Furthermore, machine learning algorithms are gaining popularity in a variety of academic sectors, including computer vision and image recognition (Fujiyoshi et al., 2019), as well as health care (Beam and Kohane, 2018), natural language processing (Otter et al., 2020), and time series forecasting (Kusiak et al., 2009). Due to the flexible architecture and effective feature-learning techniques of deep learning algorithms. In the energy sector, deep learning algorithms have demonstrated great performance, particularly in the areas of energy consumption and wind power forecasting. When compared to conventional methods, their ability to capture nonlinear time sequences yields results that improve forecast accuracy for wind and solar generation (Dairi et al., 2020). Popular deep learning models like LSTM and GRU are used often in the fields of load forecasting (Farah et al., 2022), fault detection, and the prediction of power consumption (Liu et al., 2023).

The primary challenge of deep learning is that determining a global optimum becomes harder as neural network complexity grows. CNN can extract wind power data's key properties using the weight-sharing approach, which has fewer parameters. CNN is used in image processing and computer vision due to its performance (Huang et al., 2022). that employ convolution layers to handle spatial information in images, whereas fully connected layers retain time-series data. Artificial intelligence and time-series forecasting issues differ only in the data provided to the model, which is an image vector for computer vision and a one-dimensional array for forecasting (Zhang et al., 2022). The initial input data may be seen by the observation series as a 1D array that the CNN model can access and interpret and may be utilized for time-series analysis. Several researchers are interested in FFNN using a hierarchical framework due to deep learning breakthroughs in image recognition, computer vision, and natural language processing during the last decade (Zhang et al., 2020) (Chu et al., 2018). CNN-based univariate, multivariate, multi-step, and multivariate multi-step models can solve time-series data (Zhang et al., 2020).

In this study, a patterned (year-wise) wind power dataset from the European Center for Medium-range Weather Forecasts (ECMWF) (Zameer et al., 2017) containing some missing meteorological values is taken into consideration. The first step is based on the preparation of historical wind power data by averaging the sequential data to minimize the number of input observations while capturing the characteristics of weather pattern changes. In the second phase, two different imputation methods have been suggested which address the missing value problem in wind power prediction by using a patterned K-nearest neighbor (*PkNN*) method and Clue-based missing at random (*CMAR*). Following that the use of the convolutional neural network as a productive method of weight sharing, which means it has fewer parameters and is capable of extracting inherent features. In addition, deep learning time series algorithms like LSTM are evaluated using an average of meteorological data to evaluate the performance of forecast wind energy for a public dataset that includes 24-hour annual data from seven wind farms located in climatic zones that are similar to one another. The performance of hourly energy generation forecast employing one step ahead of several standard methodologies is examined in this work, which is also distinctive in that it uses the same dataset. The following are the main contributions of this work:

- K-nearest neighbors are used to fill in the gaps in the data and boost the accuracy of the predictions. The suggested technique estimates the missing values by taking into account the local data patterns and developing a customized populated imputation.
- The effectiveness of clue-based missing at random (*CMAR*) imputation relies on the assumption that missing values are correlated with the available information. This assumption may not always be met in real-world scenarios.
- To improve prediction accuracy, a Convolutional neural network is employed to explore the hidden temporal features of wind power.
- For time series analysis, long short-term memory is used to predict the one-hour ahead forecast.

- Comparative analysis is conducted with other wind farms and with traditional machine learning models.

Following this introduction, the study is divided into three sections. The second section reveals the related work. The third section enlightens the architecture of the prediction algorithms used, the dataset examined, and the averaging of features used as input parameters. Section four presents the outcomes of implementing the algorithms to wind power in a supervised way to offer both short- and long-term projections of wind energy power generation and also addresses the significance of the findings.

2. Related work

Although incomplete data may be found in almost every area of research, it is crucial when discussing data about energy systems because most energy system models need complete data. These models may be used for various purposes, such as developing a plan for a future energy system, discovering strategies to cut CO2 emissions, or calculating investment costs (Emmanuel et al., 2021). To deal with values missing in data time series, deep learning-based imputation approaches have received significant attention in published research. The researcher created a complicated framework using a sparse autoencoder and a coordinate descendant optimization technique (SAE-CD) (Liu and Zhang, 2021). They used this strategy to deal with missing information from the supervisory control and data acquisition (SCADA) systems of fifteen different turbines that generated electricity. Experiments conducted with many different imputation algorithms based on machine learning proved that this method is preferable (Waqas Khan et al., 2020).

Moreover, Regression and classification are two of the most common applications of machine learning, which are influenced by data and create a mapping among independent and dependent variables and contain several types of neural networks, such as neural networks with feed-forward propagation and logistic regression. Shabbir *et* al. employed a support vector machine (SVM) method to predict the generation of wind energy for the following days (Shabbir et al., 2019). The authors observed that the suggested methods yielded superior forecasting outcomes, as evidenced by the lowest root mean square error (RMSE) values. Traditional predictive machine learning methods may encounter difficulties in efficiently collecting spatial information and generating precise predictions for intricate and unpredictable renewable energy data, as stated in (Voyant et al., 2017). Similarly, Ruggles *et* al. demonstrate that the findings of a power system model that only considers one region can differ by as much as 5 % when employing one of two advanced data imputation methodologies, even after the gaps in the electricity demand time series have been filled in this study (Rinaldi et al., 2021).

Since energy system models cannot function properly with incomplete data, an imputing option must be taken. Therefore, Shuker *et* al. suggest a hybrid model that utilizes Auto regression (AR) and neural network techniques to address the issue of missing values in daily wind speed time series (Shukur and Lee, 2015). The findings are contrasted with alternative methodologies, namely Linear Regression and Nearest Neighbors, evincing a marginal ascendancy of the suggested technique. Identifying a precise latent manifold is a crucial and fundamental concern in backward imputation. The K Nearest Neighbor (KNN) is recognized for its strong nonlinearity and notable regularization mechanism, which allows for the extraction of features. Several studies have shown its usefulness in identifying useful characteristics for various classification tasks including troubleshooting, finding anomalies, graphic and event categories, and regression-related tasks such as forecasting helpful life and time-series data. KNN exhibits significant potential in enhancing missing data imputations by creating a suitable feature space that identifies the latent manifold. Yu *et* al. proposed various algorithms such as the Support Vector Machine (SVM) and the Multilayer Perceptron (MLP) to estimate the missing values in the wind velocity time series. Based on the findings, it has been determined that the SVM obtains the highest accuracy for the 40-m wind speed series impute (MAE: 0.25) (Yu, 2018).

Through the literature, the results produced by utilizing simple data imputation techniques approaches are typical of poor quality. Models that predict the capacity and production of electricity generation over the long term highlight the shortcomings of these measures. Differences in predicted missing values might result in various planning techniques because supply security must always be maintained, and power demand must be satisfied. However, this might be a challenge even for predicting models useful for the short term. Regarding wind farms, short-term planning is essential to strike a healthy balance between the supply and demand of power (Emblemsvåg, 2022). However, these methodologies necessitate up-to-date reports of power and wind velocity from conceivably multiple wind farms. The observations above are susceptible to experiencing data loss, which may occur due to various factors such as maintenance, mistakes in communication, or delays. Akçay and Filik *et* al. utilized a dataset containing wind speed measurements from five distinct places in Turkey (Akcay and Filik, 2017). Their findings indicate that the average percentage of missing values dataset was 2.17 %. The PJM data and the openness platform ENTSO-E have commonly utilized data sources for conducting energy system analysis, with the former being utilized in the United States and the latter in Europe. These sources are widely recognized and frequently referenced in academic literature.

Moreover, Wu *et* al. employed a hybrid approach that integrates convolutional neural networks and LSTM to address the issue of values missing in time-series air-conditioning appliance data. Optimizing the previously mentioned structure was crucial and required adjusting hyperparameters. Furthermore, the training dataset accounted for 67.7 % of the total data utilized. In contrast to the individual CNN and LSTM models, the blended methodology exhibited superior performance (Wu et al., 2021). The LIME-RNN technique was investigated by Maarif et al. to determine missing values in various time-series datasets, such as power consumption data. The LIME-RNN is a straight recall of the vector recurrent neural networks. Despite employing identical model architectures for distinct data streams, the dependence of the model on a substantial volume of training data persisted, specifically 70 % of the complete dataset (Maarif et al., 2023). The algorithm demonstrated higher accuracy than alternative statistical and machine learning-related imputation techniques. A significant body of literature is dedicated to addressing the issue of imputing missing data in time series. Basic techniques such as mean imputation, interpolation, or last notice performed forward are commonly used for handling missing data. However, these techniques must include more complex time dependencies.

In addition, Autoregressive techniques such as ARIMA or GARCH models, eliminate the trend component of a time series and calculate the interdependencies among every moment step. The utilization of Kalman filters is expanded by incorporating state algorithms in conjunction with ARIMA, as indicated by reference (Antoniou et al., 2008). The machine and deep learning progress have facilitated the proliferation of imputation techniques. A commonly employed methodology involves utilizing sequence models such as recurrent neural networks (RNNs or LSTMs) for imputing time series data (Farah et al., 2022). A sophisticated neural network built on bi-directional LSTMs is proposed as the solution to the problem of recovering values missing from sensor data in the method outlined by Li *et* al. (Li et al., 2021). When the model was applied to data on the marine quality of the great barrier Reef, the results showed an improvement of as much as 70 % on assessment measures compared to selected baseline systems. Several contemporary methodologies utilize an attention-based framework to acquire knowledge of interdependencies in multifaceted time series. As proposed by Hahn *et* al. the self-attention system has emerged as the leading approach in various domains of machine learning, particularly in the realm of NLP.

Table 1
Raw wind power dataset containing missing values.

	$A_{ttribute\_1}$	$A_{ttribute\_2}$	$A_{ttribute\_3}$	$A_{ttribute\_4}$	$T_{arget\_outcome}$
X_{sample_1}	3.15	4.73	33.66	158.31	0.536
X_{sample_2}	2.92	4.51	32.91	117.68	0.662
X_{sample_3}	?	?	?	?	?
X_{sample_4}	?	?	?	?	?
X_{sample_5}	?	?	?	?	?
X_{sample_6}	2.66	4.18	32.5	108.59	0.817

Nevertheless, the methodology can be extended to artificial intelligence or sequential imitation domains (Hahn, 2020).

The following inferences are drawn from the review of the relevant literature: (1) It has been demonstrated that hybrid models are, in general, more accurate and useful than individual models when it comes to forecasting wind power. (2) The use of 1D Convolutional-LSTM networks has led to the development of a limited number of contemporary methods for wind power prediction. (3) Most experiments in the literature are not conducted on missing datasets.

3. Materials and methods

The proposed 1D Convolutional LSTM-based model to forecast wind power comprises three phases. Initially, two methods are employed to impute the missing data samples that are taken from three wind farms (WDF_1 – WDF_3); Table 1 depicts specific data patterns of missing records of multiple variables. The first method is patterned K-nearest neighbor (PkNN), in which the K-nearest neighbor depends on the best k value to fill the missing data values during the forecasts.

Another method is the customized populated imputation to interpolate missing values using a technique called clue-based missing at random (*CMAR*), which will increase the original size of the dataset. After that, a convolutional neural network is used to evaluate the fully imputed dataset and provide the feature extraction capability to select the most information-containing, non-redundant features, which are then given to the LSTM block. Finally, the regression-based LSTM predictors are trained by using these exclusive features for the final prediction. Fig. 1 depicts the suggested workflow diagram of the methodology, while the relevant details of intermediate processes are provided in the following sub-sections.

3.1. Wind power data imputation using K-nearest neighbor (PkNN)

The *PkNN* algorithm classifies information based on the highest weight of its neighbors, and it is allocated to the class with the highest frequency of occurrence among all its neighbors. The k-nearest neighbor technique is a filling method that works on measuring similarity and first finds the k samples that are nearest to the missing information based on the weighted distance that exists among the sample scores, and then it estimates the missing dataset of the data based on the "distance" that exists among the k observations and the missing information. Fig. 2 illustrates the overall framework of missing data transformed into imputed records.

Recently, several theories on measuring distance using K-nearest neighbor have been published (Heinermann and Kramer, 2016; Doretti et al., 2018; Wang et al., 2021; Zhang et al., 2020; Zameer et al., 2017). The Euclidian distance formulae are used to calculate the distance measure. Consider that the z^{th} imputation characteristic of X_{sample_z} is missing from Table 1. The sample set picked its K nearest neighbors from the training subset after computing the distance from X_{sample_z} to all the training samples, as shown in Eq. (1).

$$A_{X_{sample}} = \left\{X_{sample\_z}\right\}_{z=1}^{z} \tag{1}$$

The Eq. (1) set $\left\{X_{sample\_1}, X_{sample\_2}, X_{sample\_3}, \ldots, X_{sample\_z}\right\}$ represents the number of observations collected from WD_1. Hence, evaluating the distance among the non-missing values in the incomplete feature that needed to be imputed allowed for the selection of the K examples that were the closest. During the selection of its K nearest neighbors, the value that could not be determined was approximated using the feature values of $A'_{X_{sample\_z}}$ up to the z^{th} position. The assigned value $\left\{X_{sample\_3k}, X_{sample\_4k}, X_{sample\_5k}\right\}$ has been acquired through the average values of their nearest k neighbors, only in the case when the z^{th} values did not include any nulls. One major change has been assigning more weight to observations that were located at less distance X_{sample} to account for their relative weight, as seen in Eq. (2).

$$A'_{X_{sample\_z}} = {}^{1}/_{kw} \sum_{k=1}^{k} \left(w_k X_{sample\_zk}\right) \tag{2}$$

3.2. Data imputation for wind power using CMAR

There are three types of missing sequences in missing datasets. i) Missing completely at random (MCAR), the missing input has no dependency on any other input. Thus, the risk of missing information is proportional to the number of units (Farah et al., 2022); (Huang et al., 2022). ii) Missing at random (MAR), the missing values are determined by the data that is accessible. The existing data is used to estimate the lost data. iii) Missing not at random (MNAR), the interpretation of missing values from data is dependent on other missing values, making the missing values unpredictable. The wind power dataset observations as shown in Table 1 are made up of patterned missing data that can be

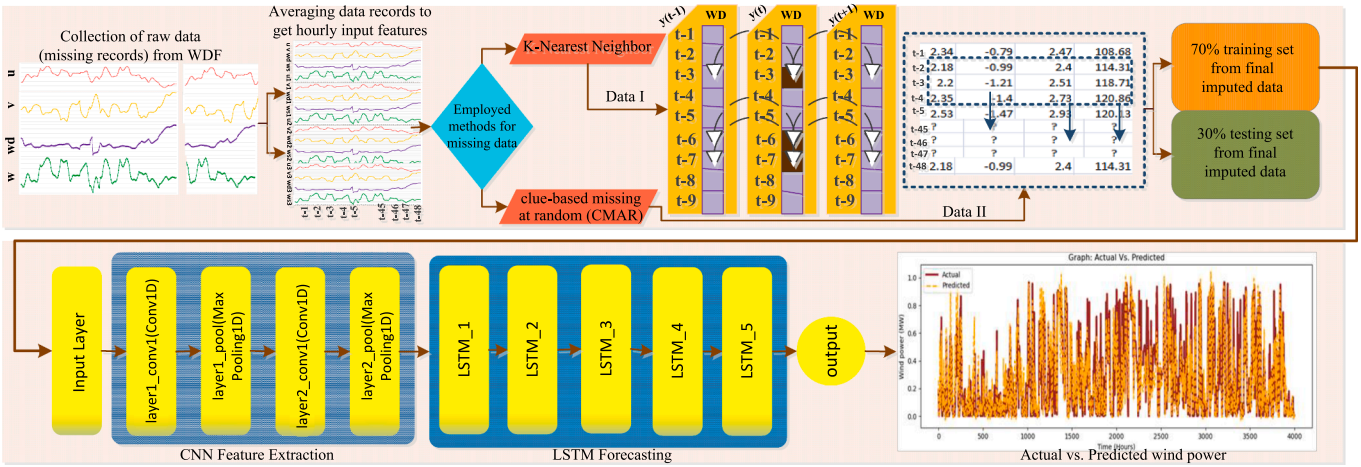


Fig. 1. Schematic diagram of proposed 1D Convolutional LSTM integrated of PkNN and CMAR to predict the wind power.

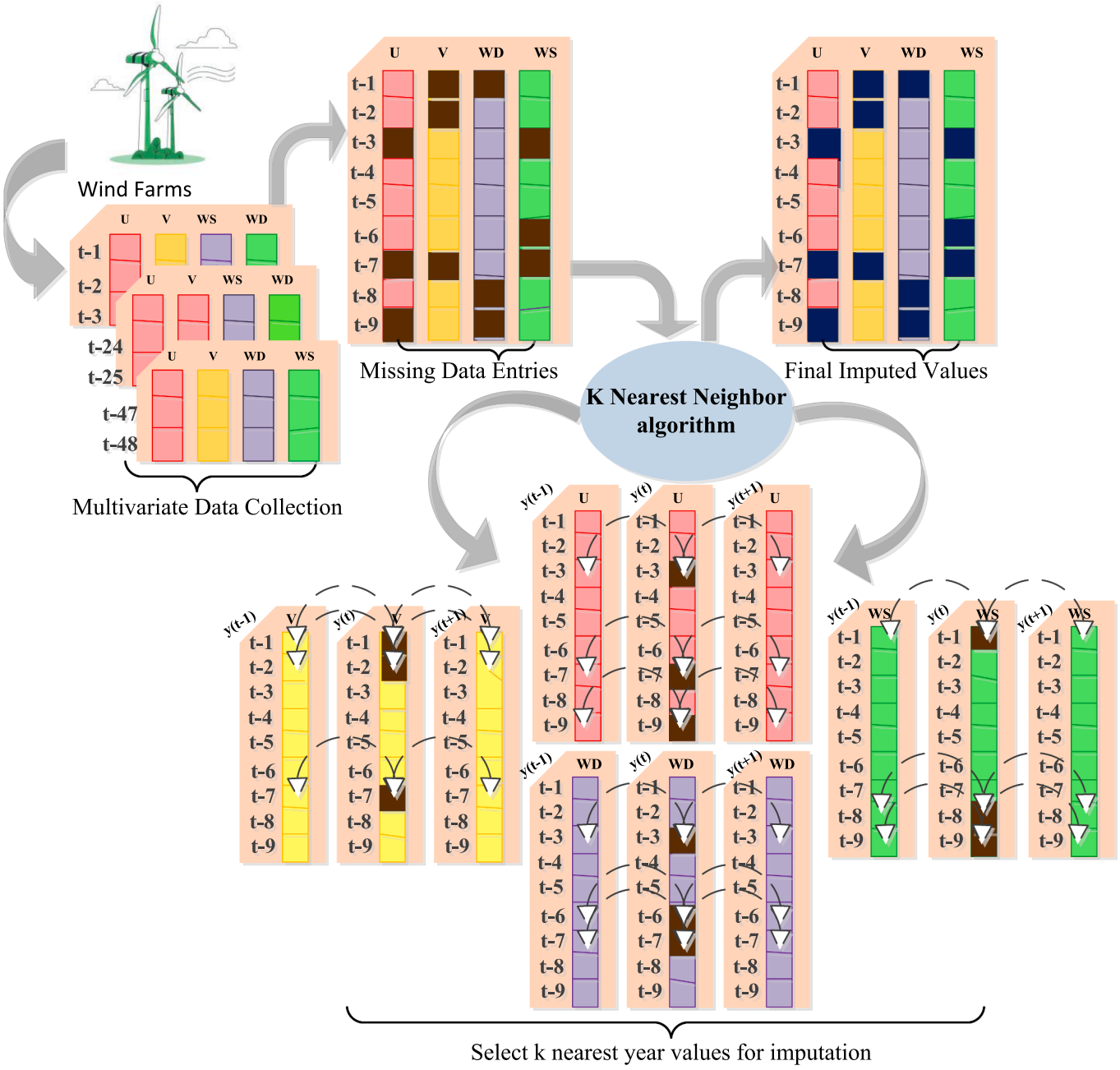


Fig. 2. Stepwise process of the *PkNN* method to fill the missing records in the wind power dataset.

used to interpolate missing values using a technique known as clue-based missing at random (*CMAR*). It entails replacing missing data with values taken from an observation or sample that is comparable.

The publicly available dataset with missing values spans three years (2009–2012) with observations made at 48-hour intervals. The four intervals are indicated as 1 January 2011 to 3 January 2011, 4 January 2011 to 6 January 2011, January 8 to 10 January 2011, and January 15 to January 17, 2011. Missing data has a distinct pattern that is often two days out. The research uses this technique *CMAR*, which takes into consideration the values of the first year to fill in the values of subsequent years. Several distinct data sets are produced and are all individually examined. It was shown that when compared to all other approaches for addressing missing values, the multiple imputation methodology can produce less biased results.

3.3. Feature extraction using a 1D Convolutional neural network

Convolutional neural networks, one of the numerous artificial neural network architectures, are frequently employed for the recognition of images, classification, and recognition of faces. CNNs are additionally enhanced for sequential interpretation by considering time intervals as an element for the convolutes. It can also be used for time series regression predictions. This deep learning framework has four distinct layers: a convolutional layer, a max-pooling layer, a fully connected layer, and a regression layer. The architecture contains neurons alongside adjustable weights and differences, which can enhance the abstract-level data characteristics (Heinermann and Kramer, 2016). As the depth of the network increases, low-level features are merged into multi-level features, guiding future models to acquire and modify these features. The structure is depicted in Fig. 3.

The CNN framework increases the efficiency of the looking forward

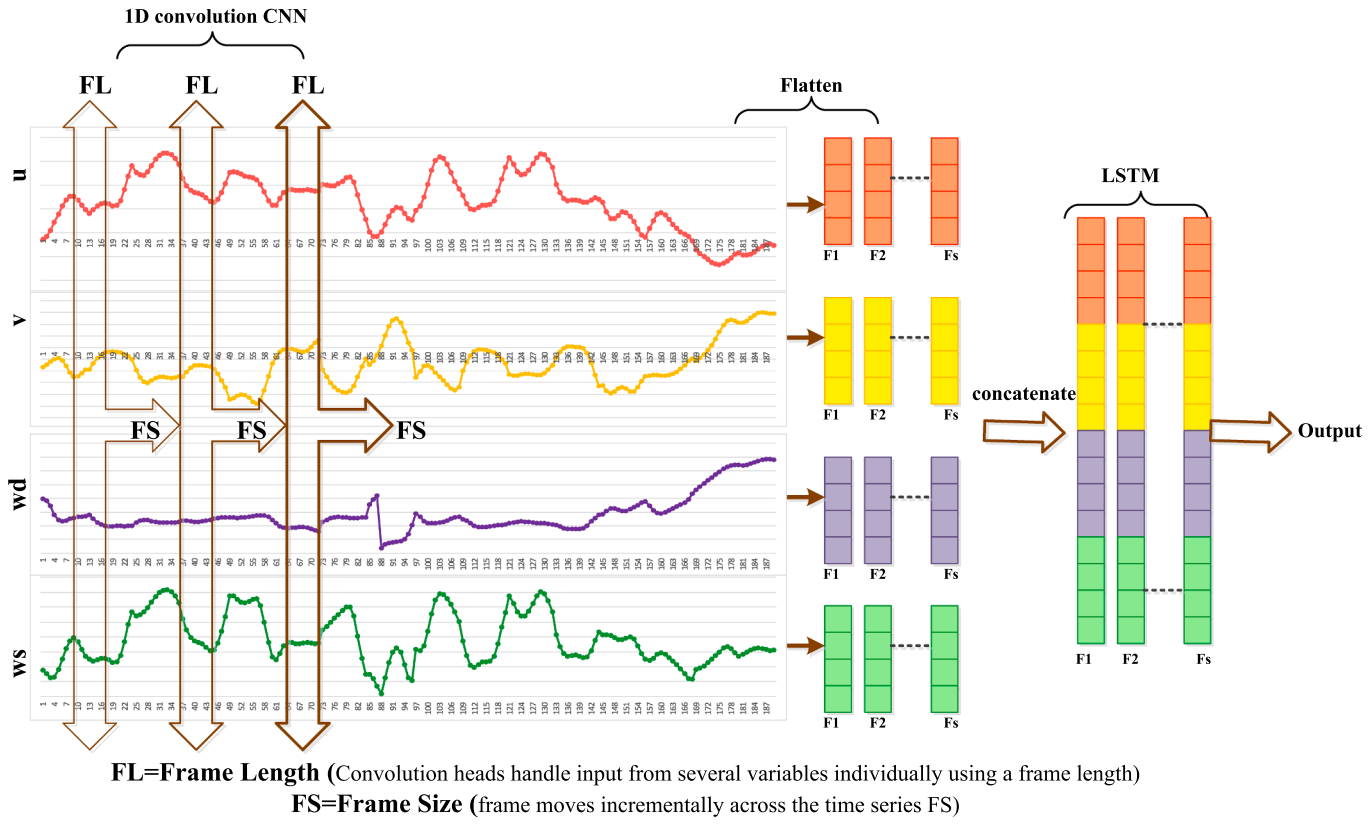


Fig. 3. Illustration of 1D Convolutional neural network with Frame Length (FL), and Frame Size (FS) to extract features (F1. F2… FS); concatenate and pass to the Convolution LSTM for prediction.

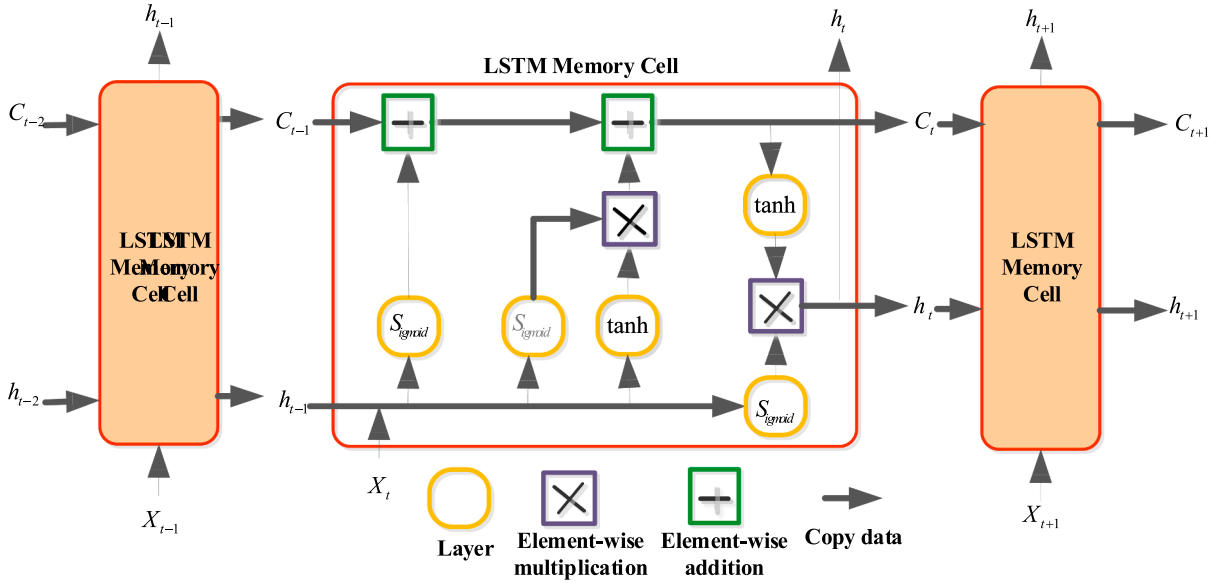


Fig. 4. Representation of previous, current, and next layers of a single LSTM memory cell.

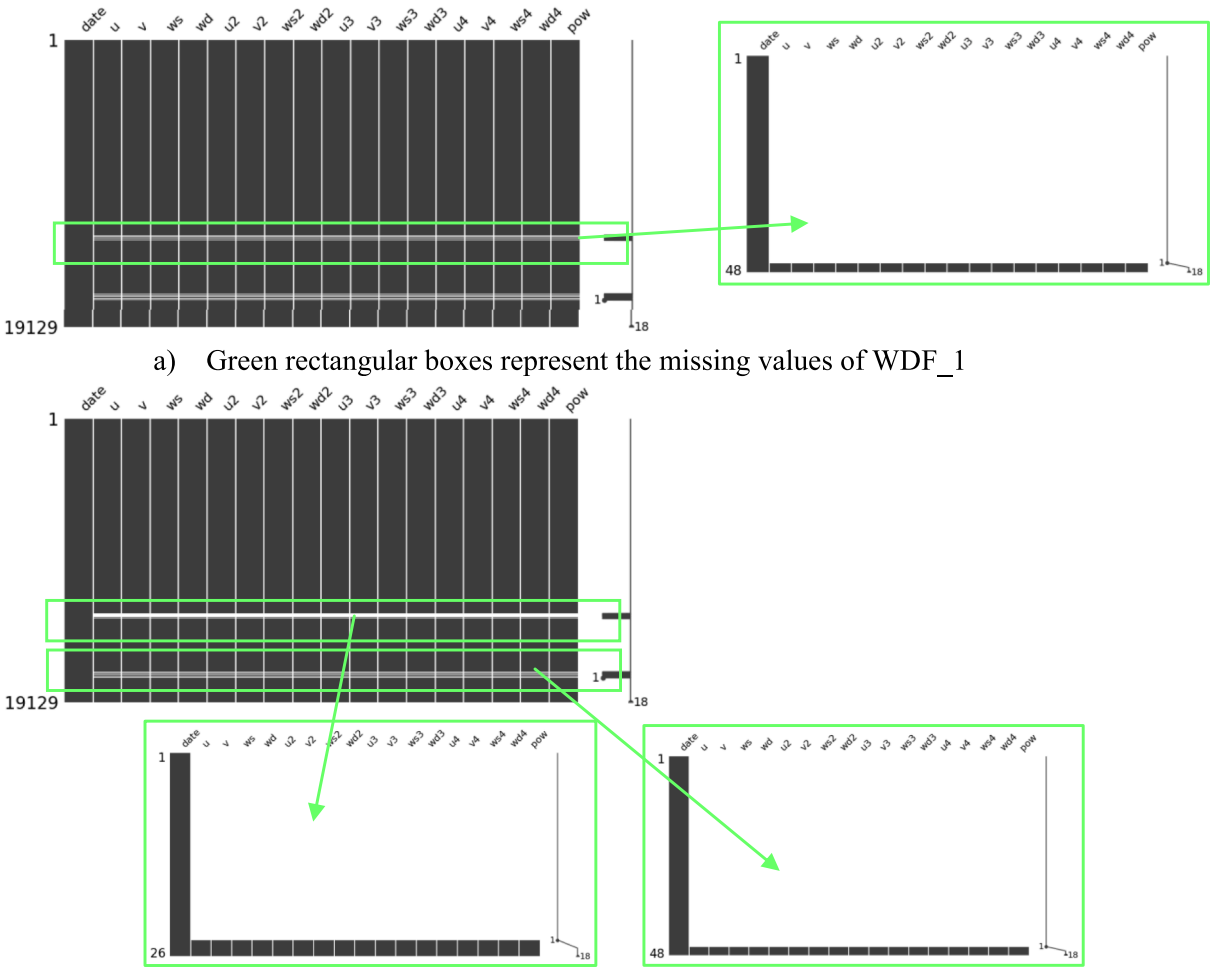
transfer function by incorporating particular characteristics into the convolutional framework and reducing the number of network parameters. The output of the 1D convolutional layer, which can extract features from the time axis, is obtained as in Eq. (3).

$$y_i = \tanh\left(\sum_{j=1}^{n} w_j x_{i-j+n} + b\right) \tag{3}$$

Here, x_i is the temporal input, and w_j is known as the convolutional kernel weight matrix, n which is the number of convolution filters and b denotes the Bias values.

3.4. Interval-wise wind power forecasting based on LSTM

The time series-based architecture of LSTM was proposed by Sepp Hochreiter (Emmanuel et al., 2021) which is an improved version of the



a) Green rectangular boxes represent the missing values of WDF_1

b) Green rectangular boxes represent the missing values of WDF_2

Fig. 5. (a) and (b) demonstrates the missing values of WDF_1 and WDF_2.

recurrent neural network (RNN). Long-short-term memory refers to the capacity to reconcile either long or very short loopbacks for associated tasks. LSTM is composed of the following gates: the forget gate (f_t), the update gate (u_t), the input gate (i_t), and the output gate (o_t), as well as a unit of memory named a cell ($\widetilde{C}_t$). Using these four gates, it is feasible to keep, write, or obtain information into or from a cell. Fig. 4 depicts the working of the LSTM sequences in detail. (4)– (9) are the controlling equations (Peppanen et al., 2016; Yeh et al., 2019). Consider that both W; and b, $X(f,i,c,o)$ are the control gates of current input X_t preceding output o_t, weight matrix, and bias, respectively.

$$f_t = S_{igmoid}\left(W_f{}^{\circ}[h_{t-1}, X_t] + b_f\right) \tag{4}$$

Here, Eq. (4) represents the element-wise multiplication of previous information and current input which depend upon the current value of the forget gate. Nonzero and zero values of the forget gate mean to pass and throw away the information, correspondingly. On the other hand, input carries out the information and stores it in the memory cell. After that i_t (the input gate) decides on the sigmoid kernel function which information is transferred and forget from the memory cell. The input gate generates a near-zero output to prevent cell updates from new data input. Thus, Cells can store data for later use.

$$i_t = S_{igmoid}(W_i{}^{\circ}[h_{t-1}, X_t] + b_i) \tag{5}$$

$$\widetilde{C}_t = \tanh(W_C{}^{\circ}[h_{t-1}, X_t] + b_C) \tag{6}$$

Finally, the newly created cell memory combines with the output gate for determining the present value of the LSTM, where the output gate uses sigmoid activation to decide which conditions in the present cell will serve as outcomes and the new memory cell uses tanh to assign output values.

$$o_t = S_{igmoid}(W_o{}^{\circ}[h_{t-1}, X_t] + b_o) \tag{7}$$

$$h_t = o_t{}^{\circ}\tanh(C_t) \tag{8}$$

Among the challenges encountered in this field is the ability to tackle tasks with long-term dependencies. LSTM has emerged as a robust algorithm for accurate time series forecasting. Despite dealing with issues such as vanishing and expanding gradients, LSTM is generally used for applications that rely heavily on previous data.

3.5. Wind power dataset normalization

The raw data of wind power is subsequently standardized using the following expression shown in Eq. (9).

$$scale\left(X_{sample\_z}\right) = \frac{X_{sample\_z} - X_{\min,sample\_z}}{X_{\max,sample\_z} - X_{\min,sample\_z}} \tag{9}$$

Here, X_{sample_z} represents the raw wind power data; the scale parameter shows the normalization of the wind dataset, and $X_{\max,sample\_z}$ $X_{\min,sample\_z}$ corresponds to the maximum and minimum values of wind power.

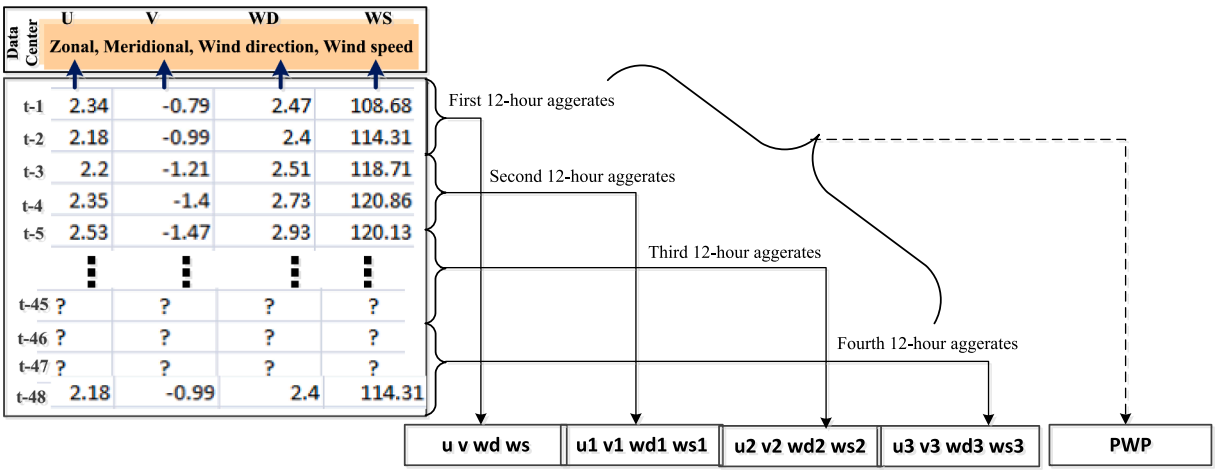


Fig. 6. The transformation of averaging previous 48-hour WDF data that comprise feature-selected inputs on every per-hour data entry.

3.6. Evaluation metrics of wind power prediction

Within Equation (10–14), metrics for performance such as MSE, MAE, RMSE and R^2, and Explained Variance Score (EVS) have been used:

$$MSE = 1/N \sum_{n=1}^{N} \left(X_{actual} - X_{forecast}\right)^2 \tag{10}$$

$$MAE = 1/N \sum_{n=1}^{N} \left|X_{actual} - X_{forecast}\right| \tag{11}$$

$$RMSE = \sqrt{1/N \sum_{n=1}^{N} \left(X_{actual} - X_{forecast}\right)^2} \tag{12}$$

$$R^2 = 1 - \sum \left|X_{actual} - X_{forecast}\right| / \sum \left|X_{actual} - X_{mean}\right| \tag{13}$$

$$EVS = 1 - Var(X_{actual} - X_{forecast}) / Var(X_{actual}) \tag{14}$$

Here, X_{actual}, and $X_{forecast}$ represent the forecasted power values, X_{mean} shows the average values of error. $X_{actual} - X_{forecast}$ illustrate the mean values and desired value of performance metrics for optimal predictions should be zero for MSE, RMSE, MAE, and one in the case of R^2, EVS.

4. Simulation results

4.1. Wind powers data curation

The Global Energy Forecasting Competition 2012 (GEFCom2012) dataset from the Kaggle competition platform is taken for this study. However, the entire dataset comprised seven wind farms (selected three wind farms: WDF_1 to WDF_3) evaluated for simulations) in the form of CSV (Comma-Separated Values) spreadsheet format. The observations are conducted at 48-hour intervals containing missing values. The four missing intervals are noted as 1 Jan 2011 to 3 Jan 2011 at time intervals 1:00 and 00:00; 4 Jan 2011 to 6 Jan 2011 at intervals 13:00 and 12:00, respectively. Additionally, 8 Jan 2011 to 10 Jan 2011 at time intervals 01:00 and 00:00; 15 Jan 2011 to 17 Jan 2011 at time intervals 01:00 and 00:00, respectively. The missing values of WDF_1 and WDF_2 is shown in the subfigure of Fig. 5.

4.2. Preprocessing and normalization of raw wind power data

The most fundamental wind farm (WDF) dataset comprised four features that were measured at certain heights above the surface of the ground. One of these elevations is commonly 10 m over the ground surface. These are the variables for the WDF_1 to WDF_3. The raw dataset includes the variables; velocity vector of zonal (u)_(ms^{-1}) in the direction of the east meridional; and the velocity vector (v)_(ms^{-1}) in the direction of north. These two factors are derived from the main variables of wind direction (wd)_(^{0}) and wind speed (ws)_(ms^{-1}). Weather

Table 2
Summary of descriptive analysis of all features of WDF_1.

Variable	Min	Max	Mean	Std. Dev	25th percentile	50th percentile	75th percentile
u	−8.66	12.14	1.6098	3.00570	-0.3700	1.9000	3.5300
v	−10.74	10.02	-0.1507	2.89877	−2.2800	-0.2500	2.0200
wd	0.12	359.97	145.6069	92.96057	75.2000	124.3500	194.7600
ws	0.02	13.91	4.0705	1.86568	2.7600	3.7000	5.1200
u1	−8.79	12.36	1.6111	3.02820	-0.4500	1.9150	3.5400
v1	−10.32	10.12	-0.13.03	2.89177	−2.2600	-0.2200	2.0300
wd1	0.03	359.97	145.8024	93.15087	74.2600	123.5550	196.9775
ws1	0.01	12.50	4.0784	1.87368	2.7600	3.6800	5.1000
u2	−9.07	13.08	1.6511	3.03107	-0.4100	1.9600	3.6000
v2	−10.79	10.71	-0.1399	2.88786	−2.3000	-0.1900	2.0100
wd2	0.02	359.94	144.7328	92.61005	73.2325	122.6400	195.3175
ws2	0.06	13.35	4.0913	1.87978	2.7800	3.6900	5.1600
u3	−9.24	13.27	1.6703	3.05517	-0.4100	1.9600	3.6400
v3	−10.85	10.34	-0.1866	2.90108	−2.3400	-0.2700	1.9700
wd3	0.02	359.92	145.2106	92.02760	74.6525	123.5850	194.9500
ws3	0.06	13.89	4.1170	1.90398	2.8100	3.7100	5.1500
Pwp	0.000	0.947	0.25286	0.246393	0.05000	0.18000	0.38600

Table 3
Patterned K-Nearest Neighbor (*PkNN*) neighborhood steps.

```
# Phase 1: Data Imputation
  Initialize missing datasets: WDF_1_imputed, WDF_2_imputed, WDF_3_imputed
  # Method 1: Patterned K-nearest neighbor (PkNN) imputation
  for each wind farm (WDF) in [WDF_1, WDF_2, WDF_3]:
  for each missing data sample in WDF:
  best_k = find_best_k(WDF, missing_sample)
  imputed_value = PkNN_impute(WDF, missing_sample, best_k)
  add imputed_value to WDF_imputed
  # Method 2: Customized populated imputation (CMAR)
  for each wind farm (WDF) in [WDF_1, WDF_2, WDF_3]:
  WDF_imputed = CMAR_impute(WDF)
# Phase 2: Feature Extraction with Convolutional Neural Network (CNN)
  Initialize empty dataset: fully_imputed_dataset
  for each wind farm (WDF) in [WDF_1_imputed, WDF_2_imputed, WDF_3_imputed]:
  features = extract_features_with_CNN(WDF)
  add features to fully_imputed_dataset
# Phase 3: LSTM-Based Wind Power Forecasting
  Initialize LSTM-based base regressor model
  for each wind farm (WDF) in fully_imputed_dataset:
  split WDF into training_data and testing_data
  train LSTM_base_regressor on training_data
  make_predictions = LSTM_base_regressor.predict(testing_data)
# Final Wind Power Forecasting
  Combine predictions from all wind farms to obtain the final forecast.
```

forecasts are frequently issued at regular intervals, with projections of u, v, ws, and wd for the next twenty-four hours. WDF dataset is considered here comprised of 12-hour predictions given every twelve hours.

This implies that for each 48-hour, each WDF can compile 48 hourly data variables. These data variables are obtained from four distinct sets of 12-hour forecasts provided at 12-hour intervals before each power measurement. The forecast model takes the 12-hourly WDF predictions and averages them rather than discarding the majority of the previous WDF data. Using this approach, feature values are aggregated such that each one contributes equally to the total value. Calculating the average value of the wind velocity and direction has a significant impact on wind power forecasting accuracy. The 12-hour WDF data patterns are utilized as input variables for this study. Power measurements collected each hour disclose sixteen meteorological feature averages. Sixteen input features are listed as {u, v, wd, ws, u_1, v_1, wd_1, ws_1 u_2, v_2, wd_2, ws_2, u_3, v_3, wd_3, ws_3}.

Fig. 6 depicts the way WDF data records are calculated and assembled on an intermittent hourly basis. Moreover, the distributions of costumed variables of WDF_1 are displayed in the form of mean, standard deviation, and 25, 50 75 percentile values in Table 2 which indicate the zonal, meridional, wind direction, and wind speed features at the interval of averaged 12-hours. This table provides a detailed summary of the descriptive statistics for each variable in the dataset WDF_1, including the minimum, maximum, mean, standard deviation, and percentiles. These statistics offer valuable insights into the distribution and characteristics of the data.

In order to prevent scaling concerns between the datasets for WF1–WF3, the power prediction (Pwp) numbers that are provided for each wind farm in the dataset are presented on a normalized scale that ranges from 0 to 1. As a result of compiling this information, 19,129 data records have been created for each wind farm (WDF_1-WDF_3), with each record including seventeen variable values, consisting of sixteen independent variables and one dependent variable (Pwp). Table 2 presents a statistical summary of the distributions of the variable values. This summary is for WF1, which is representative of the three wind farms that were taken into consideration.

4.3. Wind power forecast results based on 1D Convolutional_LSTM

The original dataset comprised 18,720 samples, after analyzing the missing observations as described in Fig. 6 (section 4.2). Therefore, data processing follows the 19,129 data observations for each wind farm that consists of 16 input features. Missing data is handled through two proposed techniques; the first one is *PkNN* which employs the K nearest neighbor to impute the missing observations and the second is *CMAR* (clue-based missing at random); according to the data patterns, missing observations are filled by the previous data observations. The processed *PkNN* and *CMAR* dataset is called here as data-I and data-II, respectively. Sections 3.1 and 3.2 previously described the proposed techniques in detail and selected features are shown in Fig. 6. A time series DL algorithm 1D Convolutional_LSTM is proposed to evaluate the hourly data observations of these three wind farms. For simulations, the normalized data is partitioned into 70 % (13390) and 30 % (5739) for the train and test set, which is passed to the proposed model and the remaining for prediction to evaluate the efficacy of the model, respectively. Fig. 1 and Table 3 depict the execution steps of the proposed hybrid model with *PkNN* and *CMAR* data imputation techniques.

The important hyper-parameters are the number of hidden neurons, the optimizer, and the number of filters essential to be decided. Parameter optimization plays a very important role in achieving higher accuracy, better efficiency, and good convergence. The Grad student descent/(babysitting AKA) trial and error rule to optimize each hyper-parameter as listed in Table 4, keeping the rest constant and then selecting the best parameter values as a trade-off between accuracy and time consumption. The experimental results for parameter optimization used different numbers k nearest neighbors for *PkNN*, which is achieved by computing the best score of the validation set, based on the score number k nearest neighbors selected. The system specifications to conduct the simulations for this work are described as GPU: precision

Table 4
Control and parameters configuration used in *PkNN* and Convolutional-LSTM models.

Proposed Models	Parameters	Value
K Nearest Neighbor (*PkNN*)	Weighted distance	Uniform
	N-neighbors	{2, 5, 8,.., 35}
	Metric	Euclidean_distance
Convolutional Neural Network	Kernel size	3
	No. of filters	32
	Activation function	'Relu'
	No. of Layers	2
LSTM	Activation function	'Relu'
	Layer	5
	Dropout	0.25
	No. of neurons	{128, 80, 64, 32, 16}
	Learning rate	0.001
	Optimizer	Adam
	Batch size	128
	Epochs	64

Table 5
Performance evaluation of proposed 1D Convolutional_LSTM for data-I &II for all WDF.

	Error Metrics	data-I (*PkNN* data imputation)	data-II (*CMAR* data imputation)
WDF-1	MSE	0.03037	0.03044
	MAE	0.13463	0.13575
	RMSE	0.17775	0.18504
	R^2	0.57991	0.52004
	EVS	0.56846	0.56468
WDF-2	MSE	0.03024	0.03145
	MAE	0.12258	0.12348
	RMSE	0.16975	0.17736
	R^2	0.67049	0.67029
	EVS	0.67989	0.67748
WDF-3	MSE	0.03026	0.03067
	MAE	0.13067	0.13453
	RMSE	0.18312	0.18418
	R^2	0.69811	0.69534
	EVS	0.68877	0.68045

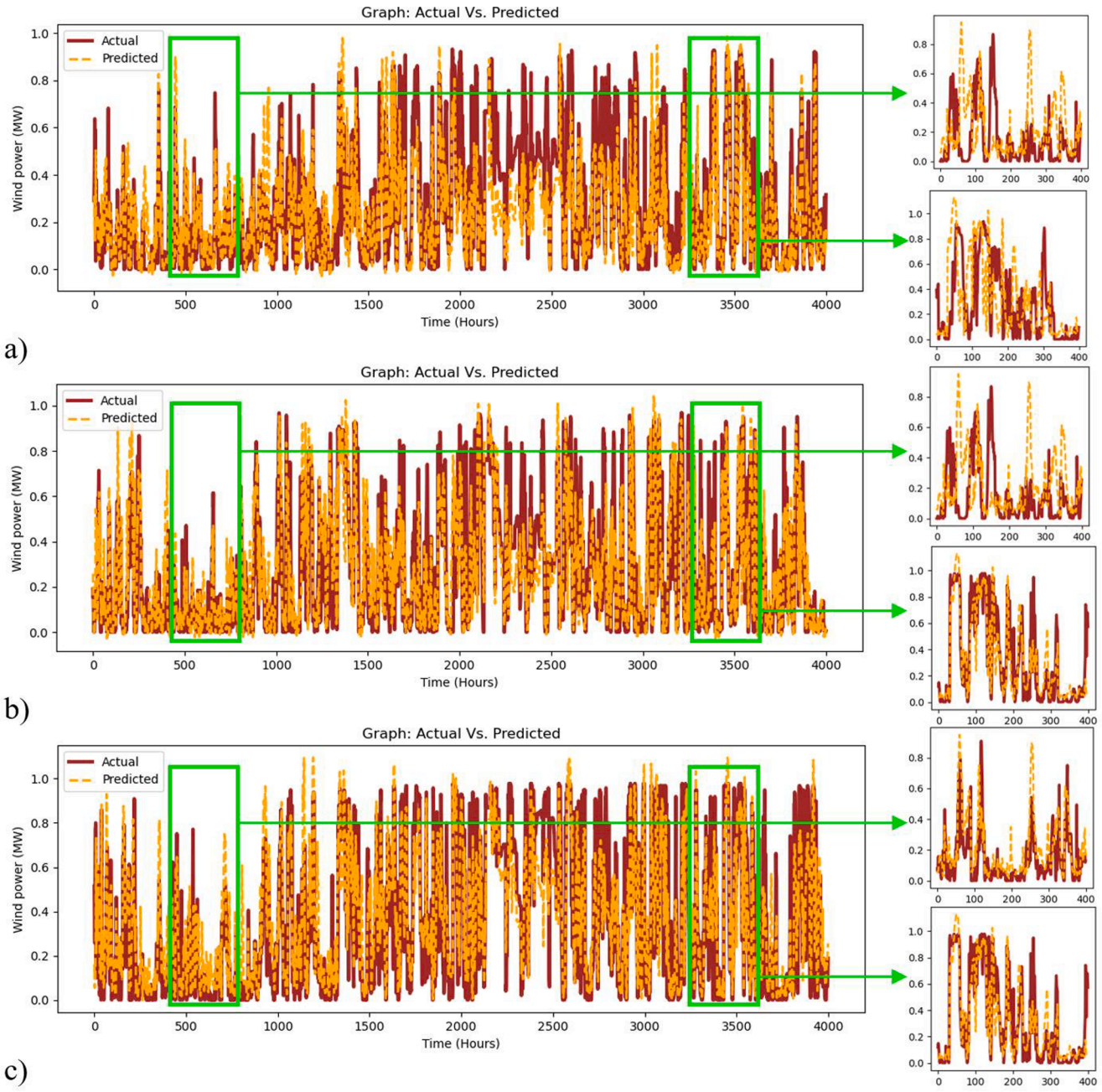


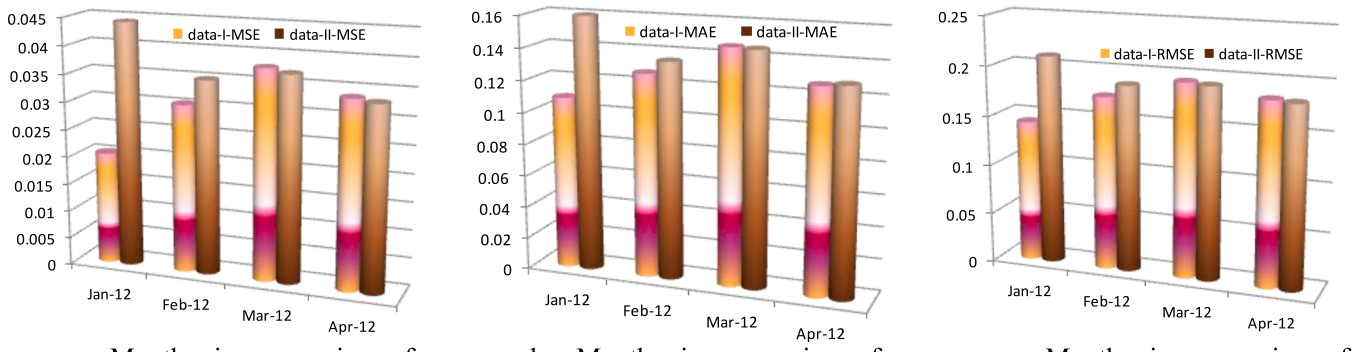
Fig. 7. Actual vs. forecasted wind power plots of data-I for WDF_1, 2, 3 as a, b, and c, respectively.

DEL 7670 workstations; RAM: 64 GB DDR4; graphics card: 2000; SSD drive: 1 TB.

However, data-I and II are trained through the proposed method and forecasted results are demonstrated in terms of five error measures in Table 5. Fig. 7 illustrates the actual and predicted curves of wind power for WDF_1 to WDF_3. It can be seen from these curves that the proposed method intimately matches the actual forecast values for all three WDFs. This illustrates that the suggested strategy operates consistently across many wind farms in the same location. However, this approach satisfies the general pattern of actual power. To assess the viability of the suggested imputation approaches, this study used three popular regression models S_LSTM, Attention_RNN, and CNN to determine accurate performance and average training accuracy as a comparative indication across both datasets (data-I and data-II), evaluate them using five common error measures as shown in Table 10, and 11 respectively.

The dataset undergoes partitioning, with 70 % allocated for training and validation, and the remaining 30 % earmarked for testing for both missing and imputed datasets (data-I, data-II). The parameters for 1D Convolutional LSTM are meticulously fine-tuned to ensure optimal learning, a process meticulously detailed in Table 4. Subsequently, the obtained results, in terms of various performance measures, are meticulously laid out in Table 5.

It is evident that *PkNN*-Conv-LSTM (data-I) performs exceptionally well across three wind farms over *CMAR*-Conv-LSTM (data-II), excelling across error metrics. It results in lower values of MSE, MAE, and RMSE for the WDF_2, while producing greater values for R^2 and EVS for the WDF_1, and WDF_3. For easier visualization, these error indices are presented as bar charts in each of the Fig. 8 subfigures. Another comparison is carried out with proposed hybrid imputed models using standard models such as S_LSTM, Attention_RNN, and CNN for all wind farms. The corresponding performance metrics are illustrated in Tables 10 and 11, and it can be observed that all these metrics have better



a. Month-wise comparison of data-I & II for WDF_2

b. Month-wise comparison of data-I & II for WDF_2

c. Month-wise comparison of data-I & II for WDF_2

Fig. 8. Bar chart analysis of monthly wise comparisons of WDF_2 between proposed imputed hybrid techniques for data-I and data-II (a-c) in terms of MSE, MAE, and RMSE.

Table 6
Performance measures of data-I of 1D Convolutional_LSTM to predict wind power forecast (Jan 2012 to June 2012) for WDF_2.

Month	MSE	MAE	RMSE	R^2	EVS
Jan	0.02043	0.10887	0.14295	0.71471	0.72297
Feb	0.03044	0.12745	0.17447	0.62610	0.63285
Mar	0.03785	0.14634	0.19455	0.64341	0.64596
Apr	0.03375	0.12746	0.18373	0.64007	0.64241
May	0.10747	0.14993	0.12783	0.57596	0.57506
June	0.06601	0.19443	0.24693	0.48486	0.48408

Table 7
Performance measures of data-I of 1D Convolutional_LSTM to predict wind power forecast (Jan 2012 to June 2012) for WDF_3.

Month	MSE	MAE	RMSE	R^2	EVS
Jan	0.02430	0.10186	0.15590	0.70295	0.71829
Feb	0.05151	0.16035	0.22697	0.55317	0.64750
Mar	0.04098	0.13369	0.20245	0.70655	0.72236
Apr	0.04252	0.14204	0.20622	0.65518	0.69419
May	0.05141	0.10613	0.24912	0.51021	0.51748
June	0.05499	0.16660	0.23450	0.64703	0.69485

Table 8
Performance analysis of data-II of proposed Convolutional_LSTM to predict wind power forecast (Jan 2012 to June 2012) for WDF_1.

Month	MSE	MAE	RMSE	R^2	EVS
Jan	0.06097	0.19288	0.24692	0.44267	0.49016
Feb	0.03677	0.14294	0.19177	0.60306	0.60370
Mar	0.04442	0.15955	0.21078	0.47616	0.50907
Apr	0.04454	0.15432	0.21104	0.59598	0.73269
May	0.03483	0.14458	0.18664	0.46879	0.50438
June	0.06242	0.18559	0.24985	0.50304	0.51260

Table 9
Performance analysis of data-II of proposed Convolutional_LSTM to predict wind power forecast (Jan 2012 to June 2012) for WDF_2.

Month	MSE	MAE	RMSE	R^2	EVS
Jan	0.04430	0.15999	0.21048	0.49236	0.57291
Feb	0.03505	0.13550	0.18722	0.57001	0.58460
Mar	0.03706	0.14575	0.19253	0.65184	0.65269
Apr	0.03319	0.12864	0.18218	0.64725	0.65201
May	0.10567	0.24711	0.32507	0.55381	0.42515
June	0.06281	0.19016	0.25061	0.56599	0.52391

values with Convolutional LSTM as opposed to compared models. However, computation time and test score on validation are comparable.

Furthermore, to check the adaptability and flexibility of the proposed model, data-I and II are again trained on one-year data (Jan 2011-Dec 012) for WDF_2, WDF_3, and WDF_1, WDF_2, respectively, to predict one –step ahead wind power of Jan 2012- June 2012. The proposed Convolutional LSTM has better-forecasted values of data-I for WDF_2 with respect to MSE 0.02043, MAE 0.10887, RMSE 0.14295, (that is near to zero) and R^2 0.71471, EVS 0.72297 (that is close to 1) in comparison with WDF_3. Based on the monthly predictions generated through different datasets (I, and II), It can also be seen from Tables 6 and 7 that the error values for January, March, and April are the lowest among all months, concisely revealing the Convolutional_LSTM best for data-I.

Moreover, comparing the error values of the proposed datasets I and II for different months, Tables 8 and 9 display the error levels for 1D Convolutional LSTM under various evaluation indices. It can be noted from these tables the error values of data-II in the month of April in terms of MSE 0.03319, MAE 0.12864, RMSE 0.18218 and R^2 0.64725, and EVS 0.65201. Overall, the forecasted results for WDF_2 are slightly lower values in the error comparison in the case of six months (Jan – June 2012). The proposed approach may generally be utilized all year long to forecast wind power.

The analysis of monthly wind power generation for wind farm WDF_2 has been extensively presented in Tables 6 and 9, employing the proposed technique on both data sets, I and II depicted in Fig. 8. Notably, it can be observed that data-I exhibits lower values in terms of MSE, MAE, and RMSE for January and February. However, comparable results were observed for March and April in the year 2012. In the case of data set II, a notable performance improvement is observed, particularly during June and July.

This implies that the proposed technique performs particularly well in capturing wind power generation trends, especially in the initial months of the year. This suggests that the proposed technique demonstrates enhanced effectiveness in accurately forecasting wind power generation during this period. This improvement may be attributed to factors such as enhanced data quality or seasonal patterns that align more favorably with the model's capabilities.

Table 5 exhibits the simulation results produced by 1D Convolutional_LSTM in the case of data-I and data-II with respect to all error metrics. In addition, Tables 10 and 11 demonstrate a comparison with three deep learning methodologies: S_LSTM, Atttention_RNN, and CNN for three wind farms. Since datasets vary due to region-wise wind farms, their methods have been applied to the current dataset for comparison purposes. Using the proposed algorithm, the value of each error measure can be seen to be in good agreement with the other techniques. It can be observed from Table 10 which has shown the results of data-I, produced

Table 10
Comparison results of the proposed technique with standard models of data-I w.r.t error measures.

	Models	MSE	MAE	RMSE	R^2	EVS
WDF_1	Convolutional_LSTM	0.03037	0.13463	0.17775	0.57991	0.56846
	S_LSTM	0.03159	0.13556	0.18461	0.52045	0.52751
	Attention_RNN	0.03547	0.14713	0.18835	0.51647	0.53069
	CNN	0.03173	0.13724	0.17813	0.51836	0.48274
WDF_2	Convolutional_LSTM	0.03024	0.12258	0.16975	0.67049	0.67989
	S_LSTM	0.03181	0.12642	0.17391	0.62885	0.63174
	Attention_RNN	0.03644	0.15601	0.19091	0.54351	0.61521
	CNN	0.03146	0.13772	0.17737	0.59476	0.61081
WDF_3	Convolutional_LSTM	0.03026	0.13067	0.18312	0.6981	0.68045
	S_LSTM	0.03278	0.13759	0.18105	0.67279	0.67884
	Attention_RNN	0.03754	0.15641	0.19377	0.66386	0.67962
	CNN	0.03367	0.13565	0.18551	0.68444	0.68476

Table 11
Comparison results of the proposed technique and standard models on data-II w.r.t error measures.

	Models	MSE	MAE	RMSE	R^2	EVS
WDF_1	Convolutional_LSTM	0.03044	0.13575	0.18504	0.52004	0.56468
	S_LSTM	0.03187	0.13273	0.18854	0.52514	0.53801
	Attention_RNN	0.03831	0.14587	0.19574	0.51646	0.54803
	CNN	0.03262	0.13692	0.18662	0.51398	0.52383
WDF_2	Convolutional_LSTM	0.03145	0.12348	0.17736	0.67029	0.67748
	S_LSTM	0.03198	0.12856	0.17736	0.60949	0.61068
	Attention_RNN	0.03362	0.14699	0.18337	0.55249	0.55933
	CNN	0.03338	0.13343	0.18271	0.56521	0.58371
WDF_3	Convolutional_LSTM	0.03067	0.13453	0.18418	0.69534	0.68877
	S_LSTM	0.03392	0.14334	0.19154	0.67194	0.67274
	Attention_RNN	0.03855	0.15791	0.19636	0.64496	0.65595
	CNN	0.03719	0.13797	0.19285	0.64032	0.67551

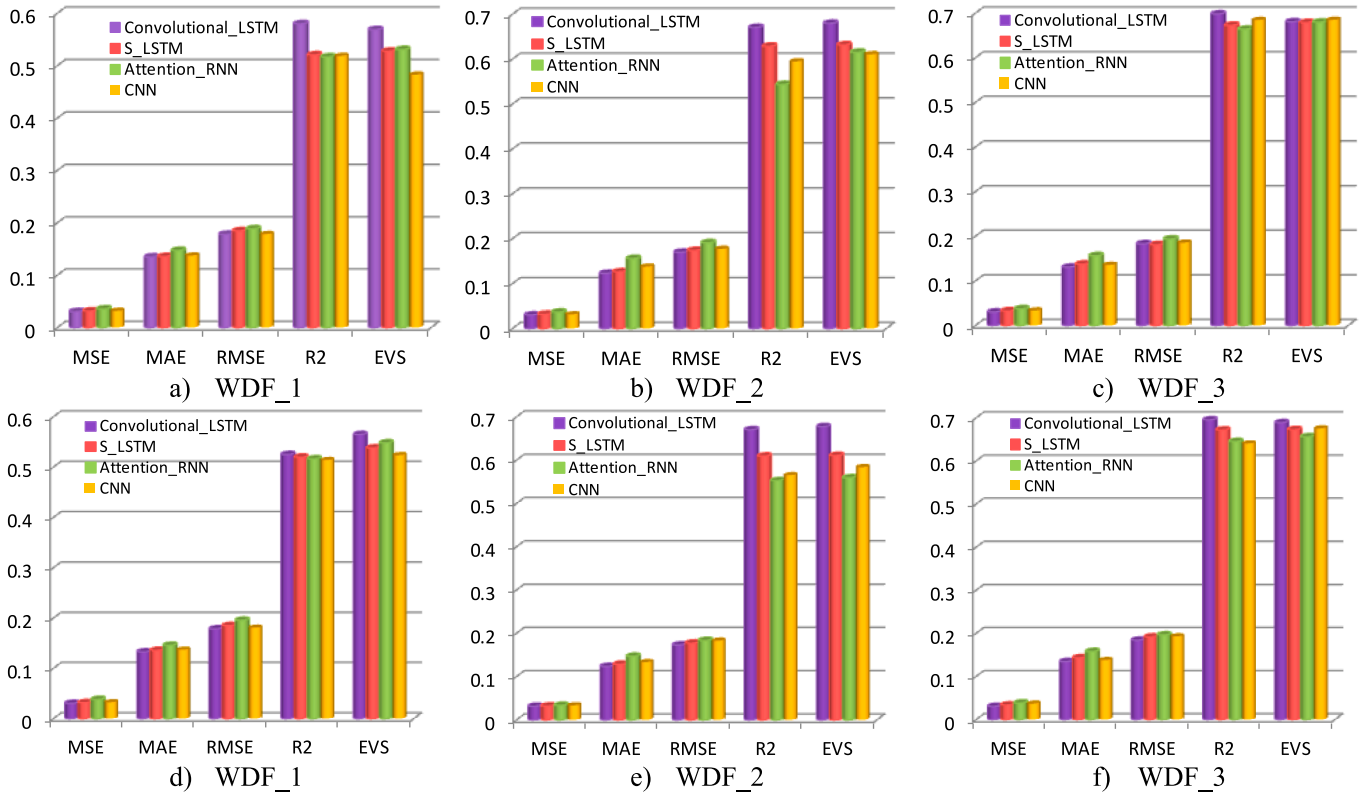


Fig. 9. Comparison of proposed 1D Convolutional LSTM with deep neural network techniques for data-I (a-c) and data-II (d-f)) in terms of five error measures.

the lowest values of MSE 0.03024, MAE 0.12258, RMSE 0.16975 for WDF_2 (wind farm) and highest values of R^2 0.6981, EVS 0.68045 for WDF_3. Additionally, it is noteworthy that the Convolutional LSTM model, when applied to data set-I, demonstrated superior performance across all error metrics when compared to the other models in the comparison.

Table 12
Comparison results of the proposed models with SOTA techniques w.r.t error measures.

	Models	MAE	RMSE
WDF_1	Convolutional_LSTM (data-I)	0.13463	0.17775
	Convolutional_LSTM (data-II)	0.13575	0.18504
	ARIMA (Saeed et al., 2017)	0.44702	0.54105
	SVR_rbf (Saeed et al., 2017)	0.23191	0.28382
	Grassi Model (Grassi and Vecchio, 2010)	0.17081	0.22155
WDF_2	Convolutional_LSTM (data-I)	0.12258	0.16975
	Convolutional_LSTM (data-II)	0.12348	0.17736
	ARIMA (Saeed et al., 2017)	0.44324	0.55437
	SVR_rbf (Saeed et al., 2017)	0.24632	0.29414
	Grassi Model (Grassi and Vecchio, 2010)	0.13524	0.17926
WDF_3	Convolutional_LSTM (data-I)	0.13067	0.18312
	Convolutional_LSTM (data-II)	0.13453	0.18418
	ARIMA (Saeed et al., 2017)	0.57114	0.68325
	SVR_rbf (Saeed et al., 2017)	0.30471	0.35202
	Grassi Model (Grassi and Vecchio, 2010)	0.14054	0.18885

Similarly, the comparative assessment between individual models and the proposed hybrid technique is presented in Table 11, demonstrating various error metrics. Specifically, for WDF_1, MSE is recorded at 0.03044. In the case of WDF_2, the MAE stands at 0.12348, while the RMSE is registered at 0.17736. Additionally, for WDF_3, both the R^2 and EVS exhibit commendable values, measuring 0.69534 and 0.68877, respectively. Upon thorough analysis, it is apparent that the Convolutional LSTM model outperforms the other models across all comparison metrics. This underscores the effectiveness of the proposed hybrid technique in enhancing predictive accuracy.

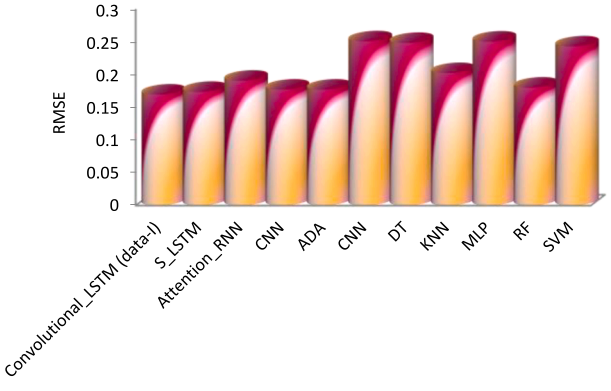
Moreover, Fig. 9 (a-g) provides a comprehensive comparison of the performance based on five error measures assessed for the four models: Convolutional_LSTM, S_LSTM, Attention_RNN, and CNN, for data-I, and data-II. The results demonstrate that the 1D Convolutional_LSTM models exhibit a notably superior performance when it comes to anticipating wind power. This reinforces the notion that the proposed model surpasses the existing ones in terms of predictive accuracy. Consequently, among all the models examined, data-I (*PkNN*-Conv-LSTM) stands out as the dataset yielding the highest predictive precision. This underscores the significance of selecting the appropriate dataset and model for accurate wind power forecasting.

The evaluation of wind power prediction accuracy involves a rigorous comparison with state-of-the-art models employed in prior studies (with missing data). Table 12 presents a detailed performance assessment of our proposed Convolutional_LSTM model on both data sets, I and II, across three wind farms. This evaluation is conducted in comparison to well-established models (with missing values) including ARIMA, SVR, (Saeed et al., 2017) and the Grassi Model (Grassi and Vecchio, 2010). A thorough examination of the comparison reveals that across a range of evaluation metrics applied to testing data, our proposed technique consistently outperforms the alternative models. Notably, the Convolutional_LSTM approach (without missing records) demonstrates superior performance across multiple metrics, displaying its competency in accurately predicting wind power generation.

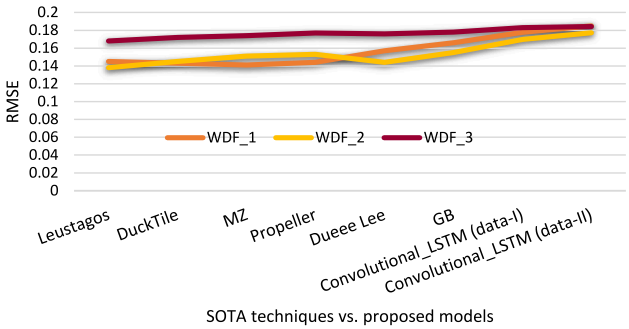
Moreover, the introduced methodology, denoted as *PkNN*-Conv-LSTM, is rigorously compared with state-of-the-art (SOTA) models in terms of RMSE using data-I. This evaluation aims to assess the predictive accuracy and efficacy of the proposed approach against established benchmarks. Bar plots and line graph as subfigures (a & b) of Fig. 10 demonstrates the comparable outcomes of the traditional approaches such as Tubulekas's model (GB) (Tubulekas, 2022) and Leustagos, DuckTile, MZ, Propeller, and Dueee Lee's techniques (Hong et al., 2012), and Wood's model (Wood, 2022) (with missing records) and suggested hybrid methodology integrated imputed techniques (without missing records) data analysis in terms of RMSE.

5. Discussion

Numerous factors such as humidity, temperature, and atmospheric pressure, have an impact on wind power, that causes uncertain and irregular behavior. Traditional machine learning algorithms have some limitations in predicting complex sequences based on a variety of influencing factors. In addition, electricity systems require highly accurate predictions of wind energy because these predictions can reduce operating costs for power systems. In order to produce forecasts with higher accuracy, it is important to use more advanced models. The proposed 1D convolutional LSTM model is computationally more efficient and significantly more accurate for one-day ahead predictions than the current forecasting methods S_LSTM, Attention_RNN, and CNN. The comparison bar charts as shown in Fig. 9, a prediction error made by the proposed model 1D Convolutional LSTM for data-I (*PkNN* data imputation) has less error rate. Results are enhanced by using k nearest neighbors for missing values in the wind farm dataset. The optimized hybrid method sometimes fails to deliver an accurate forecast on all test datasets, as can be seen in month-wise prediction tables, however, they do not result in maximum R^2 and EVS values. The wind power forecasting error in this work is reduced by 4.56 % by using the 1D



a) Bar plots contrasting the proposed methodology (*PkNN*-Conv-LSTM) to SOTA models [54] in terms of RMSE.



b) The line graph compares state-of-the-art techniques [52], [53] across *PkNN*-Conv-LSTM and *CMAR*-Conv-LSTM in terms of RMSE.

Fig. 10. Performance comparisons of proposed hybrid models (without missing records) for wind power prediction using the same GEFCom2012 dataset as previously published research (with missing values).

Convolutional LSTM model.

The primary contribution of this study lies in establishing that hybrid models, as a rule, outperform individual models in the realm of wind power forecasting, exhibiting heightened accuracy and utility. However, the integration of 1D Convolutional-LSTM networks has significantly advanced contemporary methodologies for wind power prediction, albeit with a limited number of existing approaches. Notably, a substantial portion of experiments documented in the literature does not encompass the analysis of missing datasets. This study addresses this gap and suggests its method's superior performance over various standard algorithms. Additionally, the feature selection in future work could incorporate certain improvements. Instead of using a user-defined range based on trial-and-error rules, certain statistical approaches followed by grid search can be used for dimension selection. Despite achieving high accuracy for predicting imputed wind power, it might not function well when the K-nearest neighbor value is high. Less k is more effective than a higher value. The behavior of algorithms of different training test sizes and computational approaches is also interesting to investigate, along with performance.

6. Conclusion

Wind power prediction may be used for renewable energy efficiency and has major implications for wind energy planning and power system stability. This paper conducts a comprehensive evaluation, validation, and comparative analysis of two distinct data imputation techniques, namely patterned K-Nearest Neighbor (*PkNN*) and Clue-Based Missing at Random (*CMAR*) imputation, integrated with time series-based regressor 1D Convolutional LSTM to deal with complex data by extracting the attributes of each wind farm. The experimental results indicate that the proposed hybrid technique outperformed individual methods such as S_LSTM, Attention_RNN, and CNN for all wind farms. The study further establishes the effectiveness of the 1D Convolutional LSTM-based model for data imputation (specifically, *PkNN*) in forecasting, as evidenced by achieving the lowest values of MSE and MAE. The use of the 1D Convolutional LSTM model has a considerable positive impact on the accuracy of wind power prediction. The proposed technique has a lot of potential users in forecasting renewable energy. However, the technique for imputation missing observations in K nearest neighbor with the integration of 1D convolutions is still being researched, therefore it will be a future field of study.

Funding

This work was funded by the National Natural Science Foundation of China NSFC62272419, U22A20102, Natural Science Foundation of Zhejiang Province ZJNSFLZ22F020010, and Zhejiang Normal University Research Fund ZC304022915, and research work for partially funded by Zhejiang Normal University research fund, YS304023947, and YS304023948.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Akcay, H., Filik, T., 2017. Short-term wind speed forecasting by spectral analysis from long-term observations with missing values. Appl. Energy 191, 653–662.
Al-Musaylh, M.S., et al., 2018. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. Adv. Eng. Inform. 35, 1–16.
Antoniou, C., Ben-Akiva, M., Koutsopoulos, H., 2008. Nonlinear Kalman filtering algorithms for on-line calibration of dynamic traffic assignment models. Intell. Transp. Syst. IEEE Trans. 8, 661–670.
Beam, A.L., Kohane, I.S., 2018. Big data and machine learning in health care. J. Am. Med. Assoc. 319 (13), 1317–1318.
Bigerna, S., et al., 2019. Green electricity investments: Environmental target and the optimal subsidy. Eur. J. Oper. Res. 279 (2), 635–644.
Chu, J., Guo, Z., Leng, L., 2018. Object detection based on multi-layer convolution feature fusion and online hard example mining. IEEE Access 6, 19959–19967.
Dairi, A., et al., 2020. Short-term forecasting of photovoltaic solar power production using variational auto-encoder driven deep learning approach. Appl. Sci. 10 (23), 8400.
Doretti, M., Geneletti, S., Stanghellini, E., 2018. Missing data: a unified taxonomy guided by conditional independence. Int. Stat. Rev. 86 (2), 189–204.
Emblemsvåg, J., 2022. Wind energy is not sustainable when balanced by fossil energy. Appl. Energy 305, 117748.
Emmanuel, T., et al., 2021. A survey on missing data in machine learning. J. Big Data 8 (1), 140.
Farah, S., et al., 2022. Short-term multi-hour ahead country-wide wind power prediction for Germany using gated recurrent unit deep learning. Renew. Sustain. Energy Rev. 167, 112700.
Fujiyoshi, H., Hirakawa, T., Yamashita, T., 2019. Deep learning-based image recognition for autonomous driving. IATSS Res. 43 (4), 244–252.
Grassi, G., Vecchio, P., 2010. Wind energy prediction using a two-hidden layer neural network. Commun. Nonlinear Sci. Numer. Simul. 15 (9), 2262–2266.
Hahn, M., 2020. Theoretical limitations of self-attention in neural sequence models. Trans. Assoc. Comput. Linguist. 8, 156–171.
Heinermann, J., Kramer, O., 2016. Machine learning ensembles for wind power prediction. Renew. Energy 89, 671–679.
Hong, T., P. Pinson, and S. Fan, *Global energy forecasting competition 2012*. 2014, Elsevier. p. 357-363.
Huang, L., et al., 2022. Multi-scale feature fusion convolutional neural network for indoor small target detection. Front. Neurorob. 16, 881021.
Jing, X., et al., 2022. A Multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest. Water Resour. Manag. 36 (4), 1159–1173.
Jung, S., et al., 2020. Bagging ensemble of multilayer perceptrons for missing electricity consumption data imputation. Sensors 20 (6), 1772.
Kusiak, A., Zheng, H., Song, Z., 2009. Models for monitoring wind farm power. Renew. Energy 34 (3), 583–590.
Li, Y., et al., 2021. A missing sensor measurement data reconstruction framework powered by multi-task Gaussian process regression for dam structural health monitoring systems. Measurement 186, 110085.
Liao, W., et al., 2021. Data-driven missing data imputation for wind farms using context encoder. J. Mod Power Syst. Clean Energy 10 (4), 964–976.
Liu, L., et al., 2023. Ultra-short-term wind power forecasting based on deep Bayesian model with uncertainty. Renew. Energy 205, 598–607.
Liu, X., Zhang, Z., 2021. A two-stage deep autoencoder-based missing data imputation method for wind farm SCADA data. IEEE Sens. J. 21 (9), 10933–10945.
Maarif, M.R., et al., 2023. Energy usage forecasting model based on Long Short-Term Memory (LSTM) and eXplainable Artificial Intelligence (XAI). Information 14 (5), 265.
Mittal, K., Aggarwal, G., Mahajan, P., 2019. Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. Int. J. Inf. Technol. 11, 535–540.
Norazian, M.N., et al., 2008. Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia 34 (3), 341–345.
Oktaviani, I.D., Putrada, A.G., 2022. KNN imputation to missing values of regression-based rain duration prediction on BMKG data. J. Infotel 14 (4), 249–254.
Otter, D.W., Medina, J.R., Kalita, J.K., 2020. A survey of the usages of deep learning for natural language processing. IEEE Trans. Neural Networks Learn. Syst. 32 (2), 604–624.
Peppanen, J., et al., 2016. Handling Bad or Missing Smart Meter Data through Advanced Data Imputation. IEEE.
Rinaldi, K.Z., et al., 2021. Wind and solar resource droughts in california highlight the benefits of long-term storage and integration with the western interconnect. Environ. Sci. Tech. 55 (9), 6214–6226.
Saeed, A., et al., 2017. Wind power prediction using deep neural network based meta regression and transfer learning. Appl. Soft Comput. 58.
Shabbir, Z., et al., 2019. Tetragonal Local Octa-Pattern (T-LOP) based image retrieval using genetically optimized support vector machines. Multimed. Tools Appl. 78 (16), 23617–23638.
Shahid, F., Zameer, A., Muneeb, M., 2021. A novel genetic LSTM model for wind power forecast. Energy 223, 120069.
Shukur, O., Lee, M.H., 2015. Imputation of Missing Values in Daily Wind Speed Data Using Hybrid AR-ANN Method. Mod. Appl. Sci. 9, 1.
Sudriani, Y., Ridwansyah, I., Rustini, H.A., 2019. Long short term memory (LSTM) recurrent neural network (RNN) for discharge level prediction and forecast in Cimandiri river, Indonesia. IOP Conference Series: Earth and Environmental Science. IOP Publishing.
Sun, P., et al., 2016. A generalized model for wind turbine anomaly identification based on SCADA data. Appl. Energy 168, 550–567.
Sun, C., Chen, Y., Cheng, C., 2021. Imputation of missing data from offshore wind farms using spatio-temporal correlation and feature correlation. Energy 229, 120777.
Tawn, R., Browell, J., Dinwoodie, I., 2020. Missing data in wind farm time series: Properties and effect on forecasts. Electr. Pow. Syst. Res. 189, 106640.
Tubulekas, A., *Exploring Machine Learning Techniques for Short-Term Wind Power Forecasting of Multiple Wind Parks*, in *IT*. 2022. p. 36.

Voyant, C., et al., 2017. Machine learning methods for solar radiation forecasting: A review. Renew. Energy 105, 569–582.
Wang, M.-C., Tsai, C.-F., Lin, W.-C., 2021. Towards missing electric power data imputation for energy management systems. Expert Syst. Appl. 174, 114743.
Wang, J., Zhou, Q., Zhang, X., 2018. Wind power forecasting based on time series ARMA model. IOP Conference Series: Earth and Environmental Science. IOP Publishing.
Waqas Khan, P., et al., 2020. Machine learning based hybrid system for imputation and efficient energy demand forecasting. Energies 13 (11), 2681.
Wood, D.A., 2022. Feature averaging of historical meteorological data with machine and deep learning assist wind farm power performance analysis and forecasts. Energy Syst. 1–27.
Wu, Q., et al., 2021. Ultra-short-term multi-step wind power forecasting based on CNN-LSTM. IET Renew. Power Gener. 15 (5), 1019–1029.
Yeh, S.-L., et al. *Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition*. in *Interspeech*. 2019.
Yu, C., et al., 2018. A novel framework for wind speed prediction based on recurrent neural networks and support vector machine. Energ. Conver. Manage. 178, 137–145.
Zameer, A.A., Junaid & Khan, Asifullah & Raja, Muhammad Asif Zahoor, *Intelligent and Robust Prediction of Short term Wind Power using Genetic Programming based ensemble of Neural Networks. IEEE*, 2017.
Zhang, J., et al., 2020. Non-iterative and fast deep learning: Multilayer extreme learning machines. J. Franklin Inst. 357.
Zhang, Y., et al., 2020. Mask-refined R-CNN: A network for refining object details in instance segmentation. Sensors 20 (4), 1010.
Zhang, J., et al., 2022. Physics-informed deep learning for musculoskeletal modeling: Predicting muscle forces and joint kinematics from surface EMG. IEEE Trans. Neural Syst. Rehabil. Eng. 31, 484–493.
Zhang, S., et al. Efficient kNN algorithm based on graph sparse reconstruction. in Advanced Data Mining and Applications: 10th International Conference, ADMA 2014, Guilin, China, December 19-21, 2014. Proceedings 10. 2014. Springer.
Zheng, Y., Lu, R., Shao, J., 2019. Achieving efficient and privacy-preserving k-NN query for outsourced ehealthcare data. J. Med. Syst. 43, 1–13.