Sustainable Machine Intelligence Journal

Journal Homepage: sciencesforce.com/smij

Sustain. Mach. Intell. J. Vol. 10 (2025) 1–22

Paper Type: Original Article

A Hybridized CNN-LSTM-MLP-KNN Model for Short-Term Solar Irradiance Forecasting

Doaa El-Shahat [1,*] and Ahmed Tolba [1]

[1] Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt.
Emails: doaazidan@zu.edu.eg; a.tolba24@fci.zu.edu.eg.

Received: 05 Jun 2024 **Revised**: 30 Nov 2024 **Accepted**: 30 Dec 2024 **Published**: 01 Jan 2025

Abstract

The accurate prediction of solar irradiance is essential for maximizing renewable energy production. Despite some researchers struggling to achieve sufficient prediction accuracy and model quality, we are working on developing a forecasting model for solar radiation using advanced artificial intelligence techniques. Solar energy offers a sustainable alternative to fossil fuels and has a wide range of applications, making the development of an effective forecasting model crucial. This paper proposes a hybrid model that combines deep learning and machine learning methods, denoted as CNN-LSTM-MLP-KNN. By combining Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP) and K-Nearest Neighbors (KNN), we aim to enhance the accuracy and effectiveness of time series forecasting models. The study focuses on extracting spatial and temporal patterns from sun irradiance data using CNN and LSTM and uses MLP to examine intricate connections. The KNN regressor algorithm is employed to make non-parametric forecasts based on the nearest neighbors, resulting in the final forecasting for solar irradiance. Our work relies on the historical Karachi dataset from 2017, 2018, and 2019, sourced from the NSRDB, which provides sun irradiance measurements at a 15-minute time interval using accurate meteorological instrumentation. This dataset covers a 3-year period and comprises 105120 samples with 24 features. Our proposed model provides more accurate predictions compared to the most recent published models, with an R2 Score of 0.9874, MSE of 0.0009, RMSE of 0.0311 and MAE of 0.0118. The source code is publicly accessible at https://github.com/Short-Term-Solar-Irradiance-Forecasting.

Keywords: Solar Irradiance; Deep Learning; Short-term Prediction; Time Series; Neural Networks; Hybrid Model.

1 | Introduction

In recent years, accurate and reliable Solar Irradiance Forecasting (SIR) has been critical for optimizing renewable energy generation [1]. There is a growing awareness of the necessity of finding alternative means to generate electrical energy without depleting the Earth's natural resources, as depleting those resources leads to environmental, social, and economic problems. Non-renewable resources, such as carbon-based fossil fuels [2], are formed by transforming organic matter under pressure and high heat to create fuels like gas or oil. However, using non-renewable resources poses significant environmental dangers due to their finite nature and the release of greenhouse gases during extraction and combustion. The overreliance on these resources

Corresponding Author: doaazidan@zu.edu.eg
https://doi.org/10.61356/SMIJ.2025.10448
Licensee **Sustainable Machine Intelligence Journal**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

contributes to climate change [3], air pollution [4], health damage [5] and environmental degradation [6], emphasizing the need for sustainable and renewable energy alternatives [7].

The creative solution is to employ renewable energy that is not reliant on fossil fuels. Renewable energy is the largest driver of CO2 abatement. One of the most significant renewable energy sources is solar energy, which uses sunlight to produce electricity. Furthermore, it serves as a sustainable, clean, and environmentally friendly option to restrict fossil fuel usage [8, 9]. Also, the 28th Conference of Parties (COP 28) for climate change comes to adopt the urgent need to mitigate fossil fuels. Moreover, it is the first time to observe an official health day in COP 28, which has not been seen in the past conferences [10].

For generating solar energy, it is necessary to use Photovoltaic (PV) cells [11], as they play a critical role in converting solar irradiance into electricity. Solar radiation refers to the emission of electromagnetic radiation from the sun that can be received on the surface area of solar panels during certain periods of time. Predicting the strength of solar radiation in the coming hours and days helps a lot in saving electrical energy for many solar energy-based systems, such as heating systems [12], cooling systems [13], and lightening systems [14]. Thus, SIF is very important task that determines the efficiency and performance of solar energy systems. Predicting the time when the sun does not appear helps to regularly generate electricity for these systems by using batteries to store solar energy for later utilization during the time when the sun does not appear. Forecasting plays a crucial role in refining the operation of PV systems, leading to economic benefits.

Due to its importance, predicting solar radiation has become a significant challenge for researchers to develop more accurate models. Predicting solar radiation has become a significant challenge for researchers to develop more accurate models. Recently, the spread of artificial intelligence, especially deep learning [15] and machine learning [16], has played a major role in improving the accuracy of solar radiation prediction. Although many linear and non-stationary solar radiation time series datasets are available that can be used to implement this system, some researches to date is still unsatisfactory in terms of prediction and model accuracy.

Many proposed and recently published algorithms for prediction were tested, but the results were not satisfactory either, so a new proposed model was made with high accuracy to obtain satisfactory results. The proposed model CNN-LSTM-MLP-KN is a new predictive model for 15-minute forecast of solar radiation using hybrid CNN-LSTM-MLP models integrated with convergence regression (K-neighbors regressor) and error correction. Combining hybrid CNN-LSTM-MLP models with convergence regression, K-neighbors regressor, and error correction enhances predictive accuracy by leveraging convolutional, recurrent, and dense neural networks for comprehensive feature extraction and temporal dependency modeling. Error correction mechanisms further refine predictions, ensuring robust and precise results. This approach is maximizing the model's capability to handle complex relationships in the data, resulting in superior performance across various applications especially applications based on time series data.

This study is based on the data recorded for 3 years from 2017 to 2019 using precise instruments, in Karachi, Pakistan. The proposed models and own proposed model will train, test, validate on the historical data Karachi and compare their performance by using different performance metrics, such as Coefficient of Determination (R2), Mean Square Error (MSE), Root Mean Square Error (RMSE), inference time. In the context of achieving sustainability, this paper aims to take advantage of advanced artificial intelligence, statistical models, deep learning models, hybrid models, and advanced data processing to provide more accurate predictions of solar radiation. Our main contributions are outlined as follows:

- A novel architecture of hybrid CNN-LSTM-MLP-KN for GHI predication is proposed.
- The proposed CNN-LSTM-MLP-KN is compared with some effective models published recently.
- Various statistical analyses of the forecasting accuracy are conducted across the photovoltaic grid depending on the spatial position.

The rest of this article is organized as follows. In Section 2, a literature review. In section 3, we present an architectural design of proposed model. Section 4 introduces the materials and methods. Section 5 provides the result and discussion. Section 6 presents the conclusions and future directions.

2 | Literature Review

Solar radiation prediction has always been the subject of various studies, and it has attracted the attention of many researchers around the world [17-19]. We will review and discuss some of the related work that was published in SIF. Various DL models [20] and Machine Learning (ML) models [21, 22] are employed for SIF. The authors [23] suggested a deep learning approach that combines an improved stacked BiLSTM/LSTM model with the Bayesian approach for hourly day-ahead PV forecasting of GHI and POA irradiance. Haider et al. [24] performed a comparative study of artificial neural network methods, such as ANN, CNN, and LSTM for GHI forecasting using data collected over four years and nine months in Islamabad, Pakistan. The research concluded that LSTM is the best, with R2 score of 0.984.

In [25], a heterogeneous ensemble dynamic selection model was proposed to forecast solar irradiance. Moreover, the model is capable of selecting the most appropriate forecasting algorithm from a group of seven well-known models, such as ARIMA, Random Forest (RF), Extreme Learning Machine (ELM), Support Vector Regression (SVR), Gradient Boosting (GB), multilayer perceptron neural network (MLP), and Deep Belief Network (DBN). The model has shown better results compared to the individual performance of the seven models. Lai, C.S., et al. [26] developed a hybrid deep learning model based on clustering by grouping GHI time series data into clusters. Then, the feature attention deep forecasting model assigns weights to the input features for better one-hour-ahead forecasting.

In this context, Khan, Hammad Ali, et al. [27] presented a comparison of various DL models, such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Temporal Convolution Network (TCN), and Gated Recurrent Unit (GRU) for very short-term SIF in the city of Karachi. LSTM is shown to be the best model among them. Azizi, Narjes, et al. [28] implemented multiple DL models, including MLP (Multiple Layer Perceptron), LSTM, GRU, Convolution Neural Network (CNN), and CNN-LSTM for long-term SIF and the results showed that CNN outperformed the other models. In [29], a comparative study is conducted using three types of SIF models, including image-based, time series-based, and hybrid models to forecast solar irradiance. Furthermore, the proposed hybrid MICNN-L model has proved to be more accurate for SIF. Another study [30] presented a DL model that predicts solar irradiance from one step (15 minutes) to six steps (1 hour and 30 minutes) ahead.

Lara-Benítez, Pedro, et al. [31] suggested a new data streaming approach for very short-term real-time SIF with changing weather and cloud coverage. Moreover, the authors [32] utilized the DL models to optimize renewable energy resources and battery systems by capturing the variations in solar irradiance, wind speed, and load demand during the year. For further improvement of the efficacy of the forecasting model [33], the hybrid CNN-LSTM-MLP model is incorporated with Variational Mode Decomposition (VMD) and error correction for one-hour ahead forecasting. Also, a study performed by Chandel, Shyam Singh, et al. [34] proved that the LSTM model is more precise rather than DL networks for different time horizons, whereas the GRU model has fewer parameters and may be better for small datasets.

Sivakumar, Mahima, et al. [35] concluded that the DL models (GRU, LSTM, bidirectional LSTM, CNN, Deep Neural Network (DNN)) are more accurate than the machine learning ones (Artificial Neural Network (ANN), Support Vector Machine(SVM)) for predicting the solar radiation across multiple cities in India. In addition, discrete wavelet transform and VMD techniques are employed for signal processing. Also, in [36], the authors proposed a three-stage hybrid model. The first stage typically employed the VMD, whereas the second stage applied the feature selection using a partial autocorrelation function. In the third stage, the data is passed to a Deep Belief Network-Online Sequential Extreme Learning Machine (DBN-OSELM) and a meta-heuristic algorithm is utilized to optimize the hyper-parameters of the model.

Neshat, Mehdi, et al. [37] integrated deep residual learning with GRU, LSTM, ResNet50 and an evolutionary strategy for one-hour ahead forecasting. The evolutionary strategy is used to tune ten hyper-parameters of the proposed hybrid model. Moreover, the authors in [38] adopted DNNs, such as AlexNet and ResNet-101 to identify the relevant features of total sky imager images, which were then fed into ensemble learning for training and forecasting. The statistical results by Ledmaoui, Younes, et al. [39] revealed that ANN is more effective compared to its peers of machine learning models (SVR, Random Forest (RF), Extreme Gradient Boosting (XGBOOST), Decision Tree (DT) and Generalized Additive Model (GAM)). Additionally, the authors [40] focused on very short-term SIF with time horizon from five to 15 minutes. The study introduced CNN model to sequences of infrared images from All-Sky Imager to forecast GHI based on different time horizons.

3 | Architectural Process Flow of the Proposed Model

Figure 1 illustrates the comprehensive procedure of sequential forecasting with error correction and provides a more elaborate explanation of the proposed model, starting with the time series data of solar irradiance to the final performance evaluation. At the first stage, the data is collected. Secondly, the data is prepared to be ready for processing. The data is normalized and divided into training, validation and testing sets. At the third stage, the model is trained on the training data and validation sets, while the test set is kept away for final evaluation. Finally, the goodness of results for the different models is evaluated using various performance metrics.

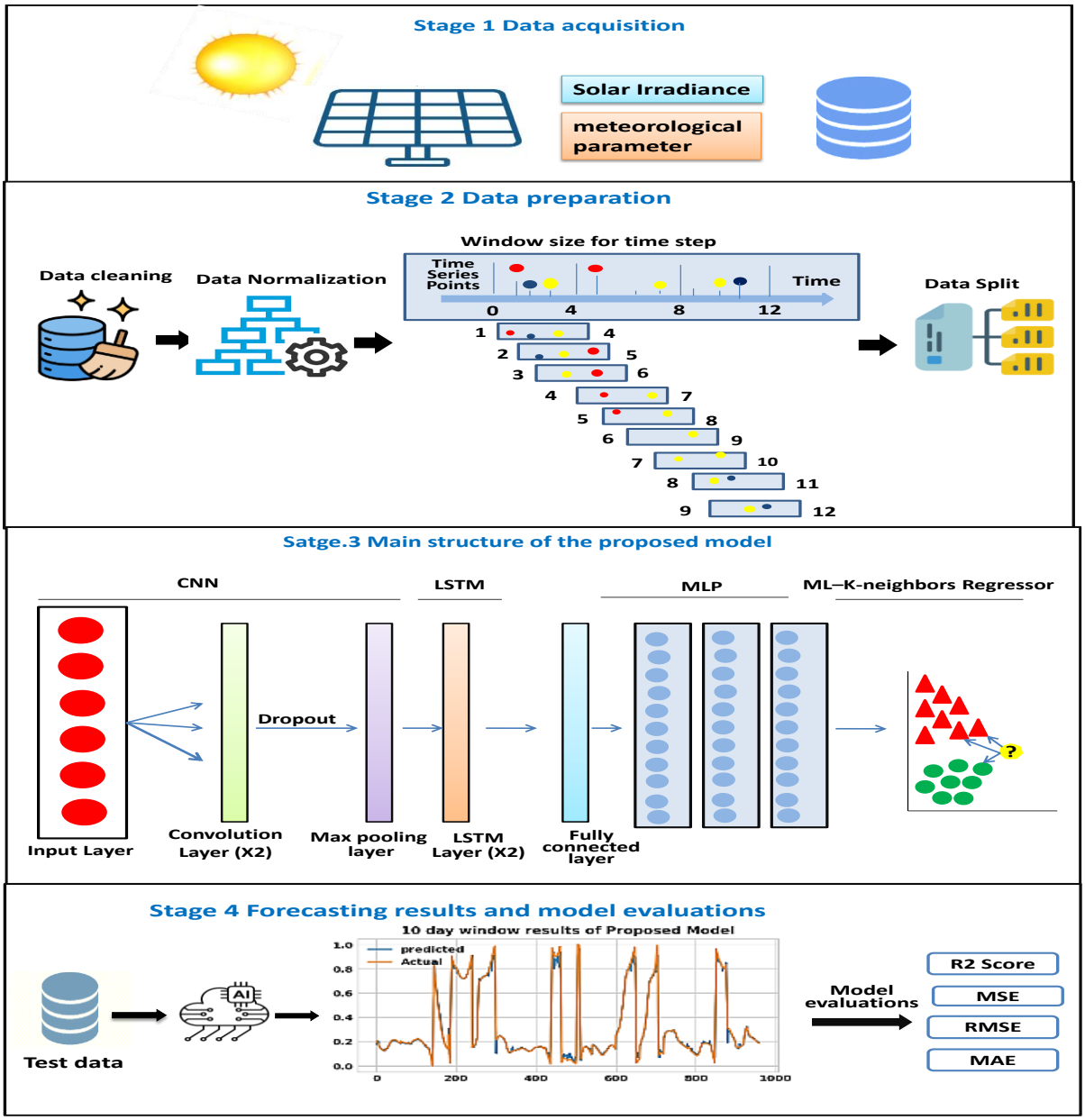


Figure 1. The comprehensive view of the process flow of the proposed model.

4 | Deep Neural Networks

The utilization of deep learning algorithms such as convolutional neural network, recurrent neural network, long short-term memory, multilayer perceptron, and hybrid models that combine deep learning and machine learning methods has demonstrated its efficacy in predicting solar irradiance due to its capacity to comprehend complex patterns and interconnections in time-series data. The objective in this particular situation is to predict the values of solar irradiance by utilizing Karachi dataset and pertinent meteorological variables.

4.1 | Convolutional Neural Network

A Convolutional Neural Network (CNN) is a specialized deep learning model designed specifically, for analyzing different datasets, such as images and time series data. Figure 2 describes the architecture of the CNN model. Convolutional layers are employed to automatically capture and extract features from the input data. The convolutional layers utilize filters to process the input, enabling the network to detect patterns and spatial relationships. The convolutional layers produce an output which is then sent to fully connected layers to carry out classification or regression tasks. It has shown considerable potential in managing time series data, such as forecasting wind speed [41], solar irradiance and stock prices [42]. In the context of 1D

convolution, the kernel serves as a filter that extracts features and uses max pooling to reduce the size of the input representation by selecting the highest value within a specified pool size window. The activation function used for one-dimensional convolutional feature extraction employs the Rectified Linear Unit (ReLU) activation function, where the activation function for the output layer is sigmoid for regression.

$$f(x) = max(0, x) \tag{1}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2}$$

σ is a sigmoid function and e is e is the mathematical constant of 2.718.

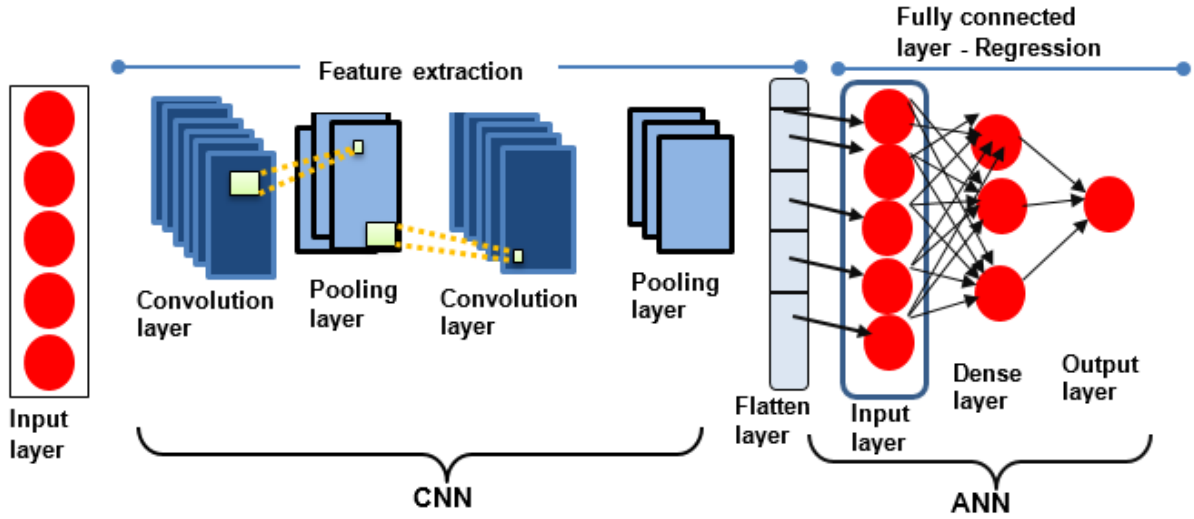


Figure 2. Architecture of CNN.

4.2 | Long Short-term Memory

Long Short-Term Memory (LSTM) is a specific form of Recurrent Neural Network (RNN) that has exceptional proficiency in handling sequential input [43]. The vanishing gradient problem is resolved by integrating memory cells with gating mechanisms. LSTM models have the ability to retain information throughout extensive sequences, which makes them well-suited for tasks involving natural language processing and time series analysis. They possess an inherent memory state that grants them the ability to deliberately discard or retain information, so facilitating the acquisition of long-term connections. LSTMs consist of input, forget and output gates that regulate the information flow inside the network in addition the activation function is tanh.

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{3}$$

The components of the system include the forget gate (ft), the memory state (Ct), the input gate (it), the output gate (ot), the hidden state (h) and the memory state (c), as shown in Figure 3.

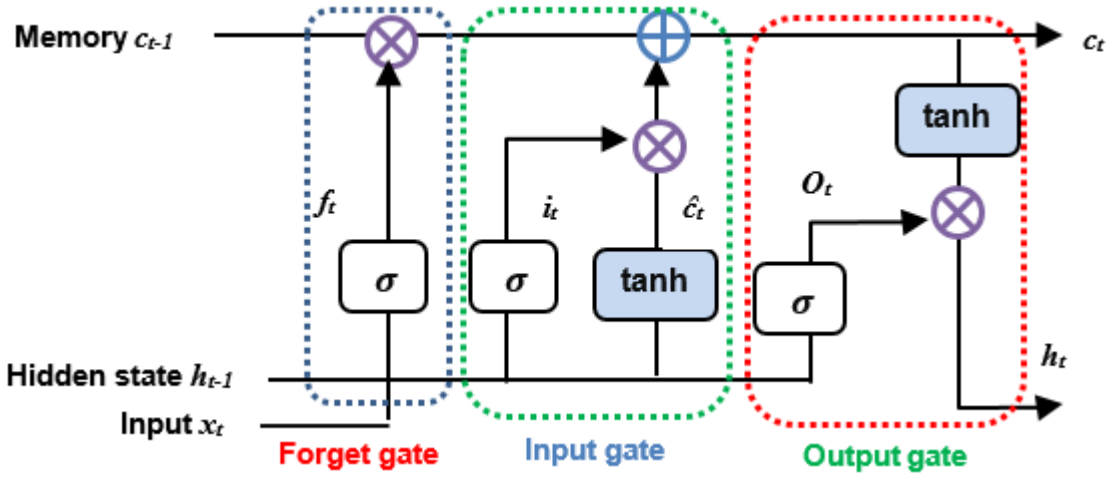


Figure 3. Architecture of LSTM.

4.3 | Multilayer Perceptron Network

Multilayer Perceptron (MLP) is a specific type of feedforward neural network and its architecture is illustrated in Figure 4. The architecture has an input layer, one or more hidden layers and an output layer [44]. Every layer consists of several interconnected artificial neurons or nodes. MLPs employ non-linear activation functions to incorporate non-linearity into the model, allowing for the acquisition of intricate patterns within the data. The training process involves utilizing backpropagation which is a technique in which the error is propagated in a reverse manner across the network to make adjustments to the weights and biases. MLPs are extensively employed for tasks such as classification, regression, and pattern recognition.

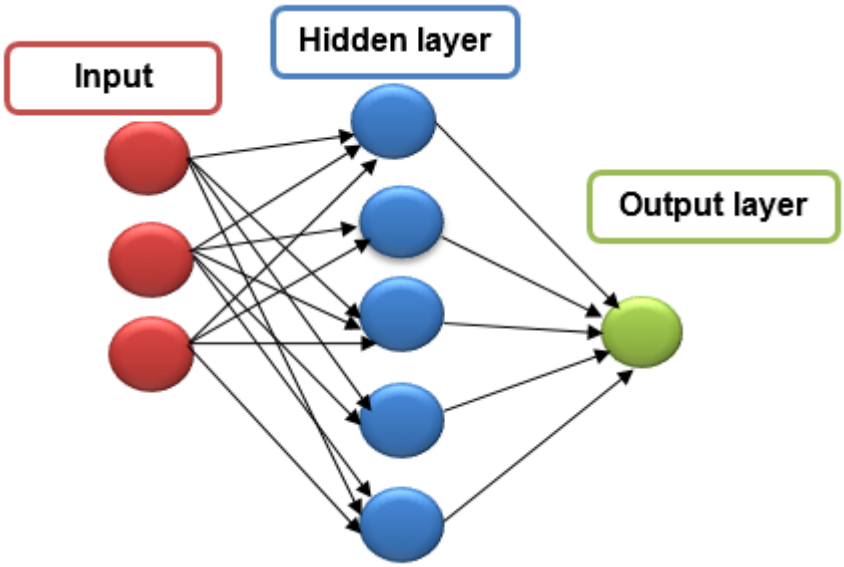


Figure 4. The structure of MLP.

4.4 | Multilayer Perceptron Network

Hybrid models are created by integrating various neural network architectures to capitalize on their own strengths and tackle unique issues. Hybrid models, which incorporate various models like CNNs, LSTMs, or their variations, have the ability to capture both local and global features, manage sequential or temporal dependencies, and handle diverse forms of input. These models are extensively utilized in many fields, such as computer vision, natural language processing, and time series analysis. Hybrid models utilize the combined

strengths of different architectures to improve efficiency and flexibility. This allows for more precise and thorough modeling of intricate data [45]. Figure 5 describes the structure of CNN-LSTM model.

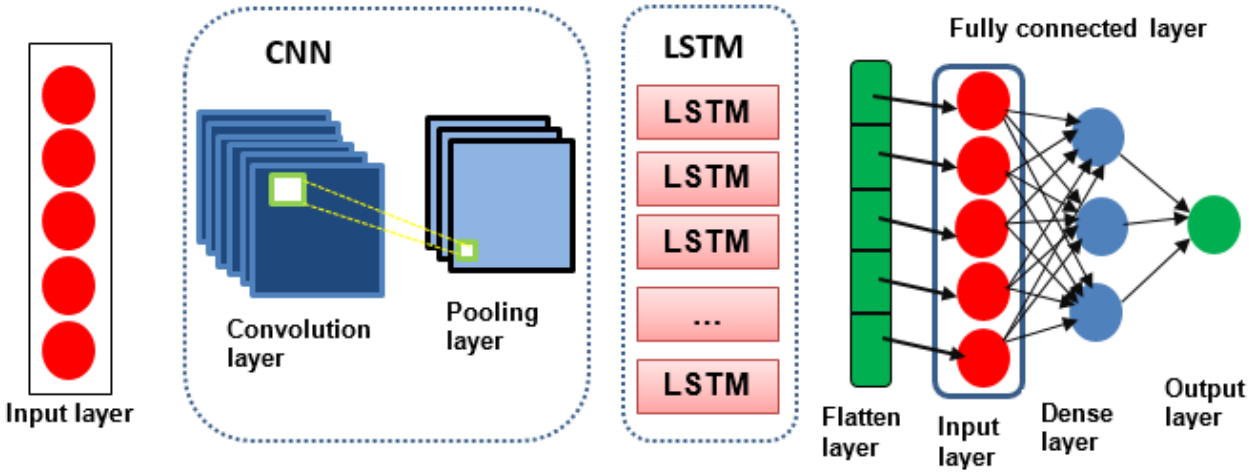


Figure 5. The structure of CNN-LSTM.

5 | The Proposed Hybrid CNN-LSTM-MLP-KNN Model

In this section, we described the proposed hybrid model, as can be shown in Figure 6. The structure of the proposed CNN-LSTM-MLP-KNN model is defined as follows:

- Convolutional layers
  - The first Conv1D layer consists of 128 filters with a kernel size of 3 units and ReLU activation function.
  - The second Conv1D layer consists of 128 filters with a kernel size of 3 and uses ReLU activation function.
  - Dropout layer with a dropout rate of 0.3.
  - The MaxPooling1D layer has a pool size of 1.
- LSTM layers
  - First LSTM layer with 64 units and returning sequences.
  - The second LSTM layer consists of 64 units.
- Flatten layer
  - It compresses the output prior to forwarding it to the dense layers.
- MLP layers
  - First dense layer consists of 64 units with ReLU activation function.
  - The second dense layer consists of 32 units with ReLU activation function.
  - The third dense layer consists of 16 units with a 'sigmoid' activation function.

The output from the last layer is utilized as input for a KNN regressor model with the specified parameters. The model contains a parameter K, which determines the number of neighbors to be taken into consideration. The K value is set to 20. The weights are set to 'uniform', indicating that all points are given identical weight in the regression. The model automatically chooses the most suitable algorithm based on the training data.

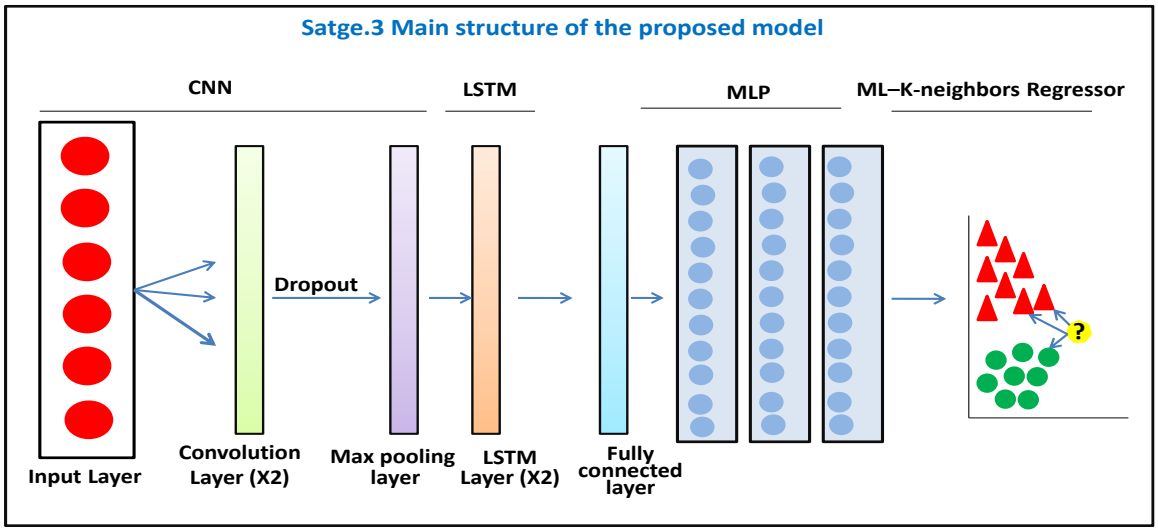


Figure 6. The structure of the proposed Model CNN-LSTM-MLP-KNN.

6 | Results and Discussion

6.1 | Experiment Setup

The various deep learning models are implemented within the same environment. Also, the Kaggle platform is employed for conducting and training all algorithms using the Nividia Tesla P100 GPU and 16 GB of RAM. Furthermore, the models are programmed in Python language of version 3.10.12. All deep learning algorithms were developed under identical conditions: Keras API Version 2.12.0, a window length of 288 and the AdamW optimizer with a learning rate of 0.01. The AdamW optimizer is a version of the Adam optimizer that integrates weight decay or L2 regularization directly into the update process. It is intended to enhance generalization performance and minimize overfitting in comparison to conventional weight decay techniques. The loss is calculated using Mean Squared Error when the model is compiled. The training and validation datasets are iterated through for 300 epochs.

6.2 | Models Evaluation

Evaluating the proposed models by using test data and assessing its performance using several evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R2 Score [46]. In addition to illustrative charts displaying the accuracy of the model in the training process, the time spent, and a box plot of the predicted values to know the level of data concentration and summarize data distributions, detect skewness, identify outliers and compare distributions.

- Mean squared error

  For calculating the MSE, take the real value, subtract the predicted value, and square that difference. Repeat that for all samples. Then, summation all of those squared values and divides by the number of samples. It is computed as follows:

  $$\mathrm{MSE} = \frac{\Sigma(y_i - p_i)^2}{n} \tag{4}$$

- Root mean square Error

  The Root Mean Squared Error (RMSE) is most performance indicators for a regression models. It measures the average difference between predicted values and real values. The lower the RMSE, the better the model and its predictions are. It can be calculated by:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}.(x_i - \hat{x_i})^2}{N}} \quad (5)$$

Where x_i denotes the real values and $\hat{x_i}$ determines the predicted values and N is the number of samples.

- Mean absolute error

  The Mean Absolute Error (MAE) is defined as the average variance between the real and predicted values. It can be mathematically formulated by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_1 - \hat{y_i}|^2 \quad (6)$$

Where y_i indicates the predicted value, p_i refers to the real values, n indicates the number of samples.

- R2 Score

  The regression coefficient determines the best possible score is 1.0, and less than that is less efficient and gets worse. R2 score can be calculated by:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\Sigma i(y_i - \hat{y_i})^2}{\Sigma i(y_i - \overline{y})^2} \quad (7)$$

6.3 | Dataset Description

The Karachi dataset from 2017, 2018 and 2019, obtained from [47], offers solar irradiance readings with a time resolution of 15 minutes for getting meteorological parameter by using precise meteorological instruments. The dataset spans a duration of 3 years and contains a total of 105,120 samples with 24 features [Year, Month, Day, Hour, Minute, Temperature, Clearsky DHI, Clearsky DNI, Clearsky GHI, Cloud Type, Dew Point, DHI, DNI, Fill Flag, GHI, Ozone, Relative Humidity, Solar Zenith Angle, Surface Albedo, Pressure, Perceptible Water, Wind Direction, Wind Speed and date [48]. Figure 7 depicts the distribution of data values over time.

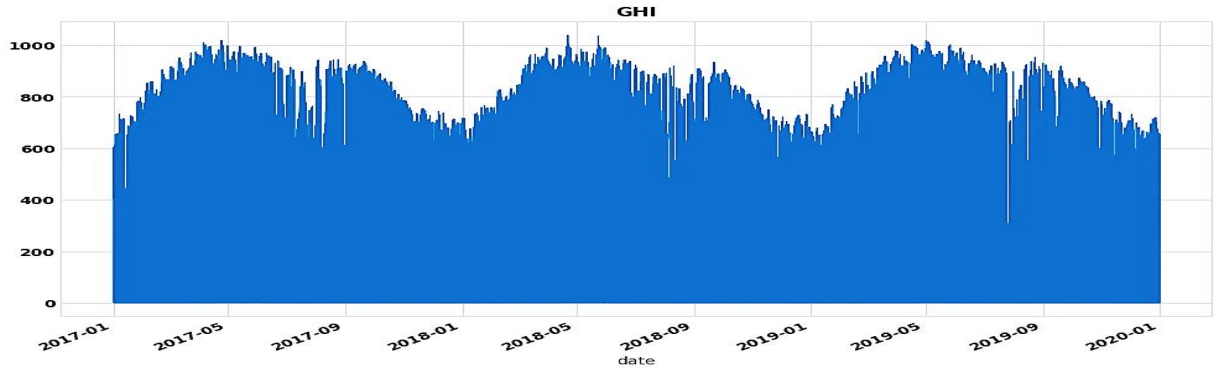


Figure 7. The distribution of data values over time.

6.4 | Dataset Distribution

Different statistical measures, such as mean, standard deviation, and quartiles (minimum, 25th percentile, median, 75th percentile, maximum) provides more insights into the central tendency, spread and distribution of the data. Additionally, it enables to identify the outliers and anomalies through the range of values and it's useful for data cleaning, quality assessment, exploratory data analysis to use it in predictions. The mean (μ) and Standard Deviation (*SD*) of these values are calculated by the following equations, respectively.

$$\mu = \frac{\sum x_i}{n} \quad (8)$$

$$SD = \sqrt{\frac{\sum(X-\mu)^2}{n}} \quad (9)$$

where μ refers to the mean and $x_i = i^{th}$ observation, $1 \leq i \leq n$. n is Number of observations. SD is the standard deviation. Box plots are employed to summarize data distributions, detect skewness, identify outliers, and compare distributions. They offer a graphical depiction of 5 quarters, as shown in Figure 8. The minimum (*Min*) value is the smallest value in the dataset not including any outliers. The first quartile (Q1), also known as the lower quartile, represents the number below which 25% of the data lies. The median (Q2) represents the value that divides the dataset into two equal halves. It divides the values into two equal parts, with half below and half above. The Third Quartile (Q3) represents the point at which 75% of the data falls below it. The maximum (*Max*) contains the highest value in the dataset, not including any outliers.

Table 1 records the statistical analysis for the distribution of data for Karachi dataset, in terms of their values count, mean, *Min*, *Max*, first quartile, third quartile and SD. The dataset have different distributions, where some variables have a wide range of values with significant variations, such as GHI, DNI, while other variables have smaller ranges and lower variations, such as Temperature, Wind Speed and Perceptible Water.

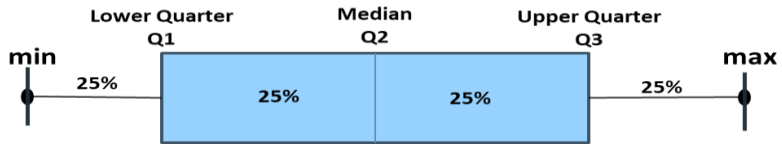


Figure 8. Box plot diagram.

Table 1. The statistical analysis for data distributions of Karachi dataset.

	Count	Mean	Min	First quartile	Median	Third quartile	Max	Std
Year	105120	2018	2017	2017	2018	2019	2019	0.8165
Month	105120	6.526027	1	4	7	10	12	3.447868
Day	105120	15.72055	1	8	16	23	31	8.796289
Hour	105120	11.5	0	5.75	11.5	17.25	23	6.922219
Minute	105120	22.5	0	11.25	22.5	33.75	45	16.77059
Temperature	105120	26.56804	11.1	23.4	27.2	30	40.1	4.880504
Clearsky DHI	105120	101.3397	0	0	0	195	603	128.7611
Clearsky DNI	105120	223.9575	0	0	0	487	957	280.521
Clearsky GHI	105120	251.9437	0	0	0	546	1039	325.4948
Cloud Type	105120	1.332544	0	0	0	2	12	2.373412
Dew Point	105120	18.75525	-10.6	13.4	21.6	25.1	28.5	7.557533
DHI	105120	100.9	0	0	0	191	603	132.7869
DNI	105120	186.825	0	0	0	383	957	265.8622
Fill Flag	105120	0.17519	0	0	0	0	5	0.764574
GHI	105120	225.1345	0	0	0	464	1039	303.9781
Ozone	105120	0.270101	0.217	0.261	0.273	0.281	0.341	0.016094
Relative Humidity	105120	66.10673	9.1	48.88	68.785	85.0025	100	21.82006
Solar Zenith Angle	105120	89.74373	1.6	52.93	89.45	126.76	178.59	43.10112
Surface Albedo	105120	0.177836	0.16	0.18	0.18	0.18	0.2	0.006577
Pressure	105120	1008.235	991	1002	1009	1014	1026	6.720981
Perceptible Water	105120	2.839464	0.3	1.6	2.4	3.9	8	1.657345
Wind Direction	105120	219.4983	0	219	245	263	360	83.69508
Wind Speed	105120	3.535639	0.2	2.3	3.3	4.6	10.2	1.64857
Date	105120	2/7/2018 52:30.0	1/1/2017 0:00	1/10/2017 17:56	2/7/2018 11:52	2/4/2019 5:48	31/12/20 19 23:45	NaN

6.5 | Data Preparation

Initially, the process of data cleaning is executed to eliminate any discrepancies, null values, and extreme values that have an impact on the effectiveness of the model. Additionally, doing data normalization is crucial to

guarantee that all features are standardized to a comparable scale, hence avoiding the dominance of any specific feature during model training. In addition feature selection to select features that have positive correlation and this enhance quality of data. Subsequently, the data is subjected to window sizing to enable the model to effectively capture temporal patterns in solar irradiance data. Ultimately, the dataset is divided into three distinct sets: training, testing, and validation. This division allows the model to be trained on one piece of the data, evaluated on another section, and validated on a separate portion to examine its capacity to generalize. The last ten days of data were exclusively allocated for testing, while the remaining dataset was divided into 80% for training and 20% for validation.

Exploratory data analysis uncovered a significant quantity of roughly 52,986 GHI values that are equal to zero. These values correspond to nighttime measurements and are clearly 0 because there is no solar radiation present. Hence, it is advisable to delete the values during nighttime. This may be achieved by opting to clear the data between 7:00 PM and 5:00 AM, during which the sunset occurs till sunrise at 5:00 AM. Utilize exclusively the data collected during daylight hours when solar radiation is present. Upon completion of these procedures, the values were lowered to 28,004. This number is considered quite large and considering the data spans a period of three years, encompassing a minimum of nine winter months. Each year consists of three months, namely December, January, and February. During this period, the duration of daylight is reduced and the skies are frequently overcast. According to the examination of exploratory data, the minimum temperature recorded is 11.1 degrees, suggesting that it corresponds to the winter season. This suggests occasional cold weather and insufficient visibility of the solar radiation

As mentioned there are various distributions with a wide range and small range in the dataset. It is crucial to perform data normalization by min–max normalization to scaled data in the range (0, 1) for several reasons: to standardize features, prevent biased results, enhance the stability and convergence of deep learning algorithms, mitigate the impact of outliers, and reduce computational complexity and storage needs.

$$x_{\text{scaled}} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{10}$$

x_{scaled} defines scaled data in the range (0, 1). X determines sample. x_{min} represents the minimum value. x_{max} indicates the maximum value.

Figure 9 displays the heatmap correlation of all features. The darker degree of red and blue colors indicates a strong correlation, while the brighter one indicates weak correlation. Pearson correlation coefficient finds the correlation between all the features. Our investigation revealed the existence of multiple attributes, including "year", "month", "day", "hour", and "minute". These attributes can be consolidated into a single feature that encompasses both date and time. Consequently, they can be extracted from the individual attributes and included into one unified feature, thereby eliminating redundancy. Some variables, such as "Cloud Type", "Ozone", "Solar Peak Angle", "Surface Albedo", "Clearsky DHI", "Clearsky DNI", "Clearsky GHI", "DHI", "DNI", and "Fill the tag", have a negative correlation. From this analysis, the variables with the strongest correlation to the target variable were chosen. These variables include "Temperature", "Dew Point", "Wind Speed", "Relative Humidity", "Pressure", "Precipitable Water", "Wind Direction", and "GHI". A heat map can be used to visually represent these variables.

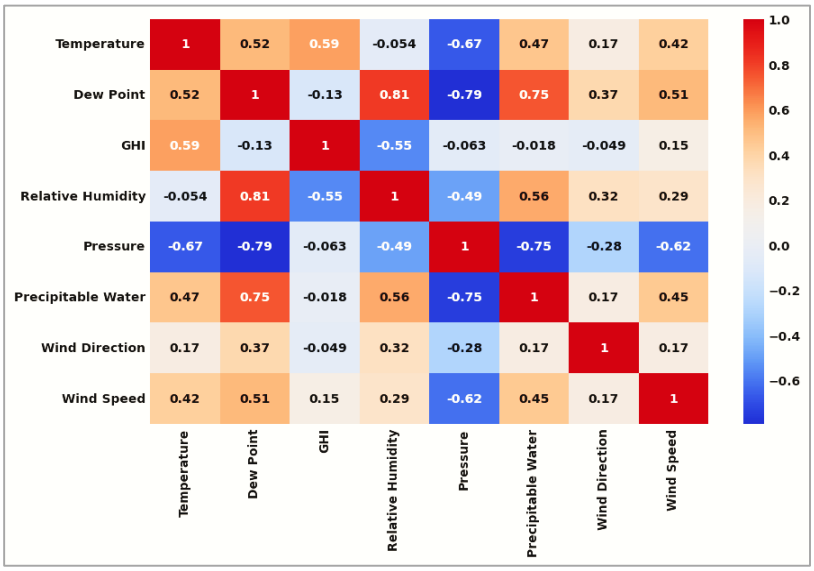


Figure 9. The heatmap correlation of the features.

In neural networks, a window size refers to the number of past time steps used as input to predict the next time step considered at each iteration (see Figure (10)). It is an important parameter that affects the model's ability to capture temporal dependencies and make accurate predictions when dealing with time-series data. By selecting a suitable window size, the model has the ability to capture both short-term and long-term patterns in the time-series data. A reduced window size prioritizes recent information and is appropriate for capturing temporary relationships. Conversely, a bigger window size enables the model to take into account a wider range of information and capture dependencies that span over a longer period of time. As we mentioned that the Karachi dataset provides solar irradiance readings at a time resolution of 15 minutes. For our analysis, we will focus on the past 3 days because we deal with short term task, resulting in a total of 288 readings. Each day consists of 24 hours, with each hour having 4 readings. Multiplying 4 readings per hour by 24 hours per day gives us 96 readings per day. Multiplying this by 3 days gives us a total of 288 readings.

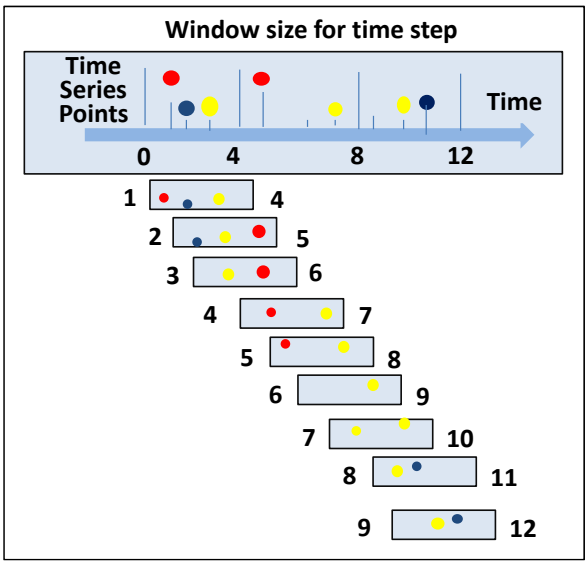


Figure 10. The window size of the neural networks.

6.6 | The Findings of the Proposed Model Against other Deep Learning Models

This subsection investigates the performance of the proposed model compared to nine models of the most well-known and recent deep learning models. Table 2 displays the results of the different deep learning model,

including CNN [28], LSTM [27], CNN-LSTM [30], MLP [31], CNN-LSTM-MLP [33], TCN [27], Attention mechanism [49], BI-LSTM [50] and CNN-Attention LSTM [51] compared to the proposed model (CNN-LSTM-MP-KNN) for short-term solar irradiance forecasting. It reveals that the proposed model attained the highest R2 value of 0.9874 and the lowest values for MSE, RMSE and MAE with values of 0.0311, 0.0009 and 0.0118, respectively. These values indicate the superiority of our proposed model over the other models. The LSTM model achieved a second-place ranking with an R2 value of 0.9695, followed by the CNN model. Although MLP takes the least training time with a value of 115.6595 seconds, its performance is poor with R2 score value = 0.7799. Unfortunately, the CNN-LSTM model has the worst performance with R2 value of 0.7520.

Figure 11 presents the ranks of each model for predicting the solar irradiance based on the four performance metrics: R2 score, MSE, RMSE and MAE to give us a general view for the models performance. We can see that the proposed CNN-LSTM-MLP-KNN model has the best ranks for all metrics. Figure 12 displays the models' prediction plots, demonstrating their performance and variations in the context of short-term solar irradiance forecasting. The figure shows that the predicted data of the proposed model better fits the actual data than the other models, whereas CNN-LSTM predictions doesn't fit well with the actual data.

Tables 3-5 bring graphs depicting the historical loss values and RMSE for all previous models. The data history across 300 epochs and show the loss function values from both the training and validation phases of the models. After examining the table, it is clear that the suggested model has outstanding stability and accuracy. The graphs depicting the loss values show steady and slight variations during both the training and validation stages. The stability demonstrates that the suggested model effectively learns from the input and generates dependable outcomes. The CNN-LSTM model also shows steady outcomes, although not as consistently as the suggested model. Although the CNN-LSTM model may experience slight changes in the loss values, it demonstrates strong performance. The charts clearly show that the proposed model outperforms other neural network models in terms of predictions.

Figure 13 and 14 illustrate the models' inference and training time. The charts indicate that the proposed model requires a significant amount of time for training, approximately 2097.1165 seconds, but significantly less time for prediction of MLP, around 115.6596 seconds. CNN-LSTM has the worst training value of 2468.7109. Although the proposed model comes in the seventh rank, it has the lowest inference time with value of 0.1703. Figure 15 shows the boxplot of loss values of the six models. As mentioned that the boxplots are used to summarize data distributions, detect skewness, identify outliers and compare distributions. They provide a visual representation of five metrics. We used it to summarize the distribution of model loss during training to identify skewness and detect outliers. MLP and CNN models exhibit outliers, but the proposed model shows a more even distribution of data. The loss values of the proposed model are distributed in a dense concentration.

Table 2. The numerical results obtained by the different deep learning models.

Model Name	# Parameters	Model training time	R^2 Score	MSE	RMSE	MAE
CNN-LSTM-MLP-KNN	141313	2097.1165	0.9874	0.0009	0.0311	0.0118
CNN [28]	716201	526.9438	0.9589	0.0031	0.0563	0.0265
CNN-LSTM [30]	2484609	542.6634	0.7520	0.1227	0.1385	0.1227
LSTM [27]	51521	1210.9236	0.9705	0.0022	0.0477	0.0172
MLP [31]	65217	115.6596	0.7799	0.0170	0.1304	0.0646
CNN-LSTM-MLP [33]	71713	1572.2155	0.8850	0.0088	0.0943	0.0890
CNN-Attention-LSTM [51]	2351713	1529.9048	0.9357	0.0561	0.2370	0.0561
Attention [49]	28929	865.3272	0.9817	0.0091	0.0957	0.0091
TCN [27]	2435073	2198.6368	0.9822	0.0145	0.1205	0.0145
BI- LSTM [50]	260097	2468.7109	0.9786	0.0246	0.1571	0.0246

Bold font indicates the best results

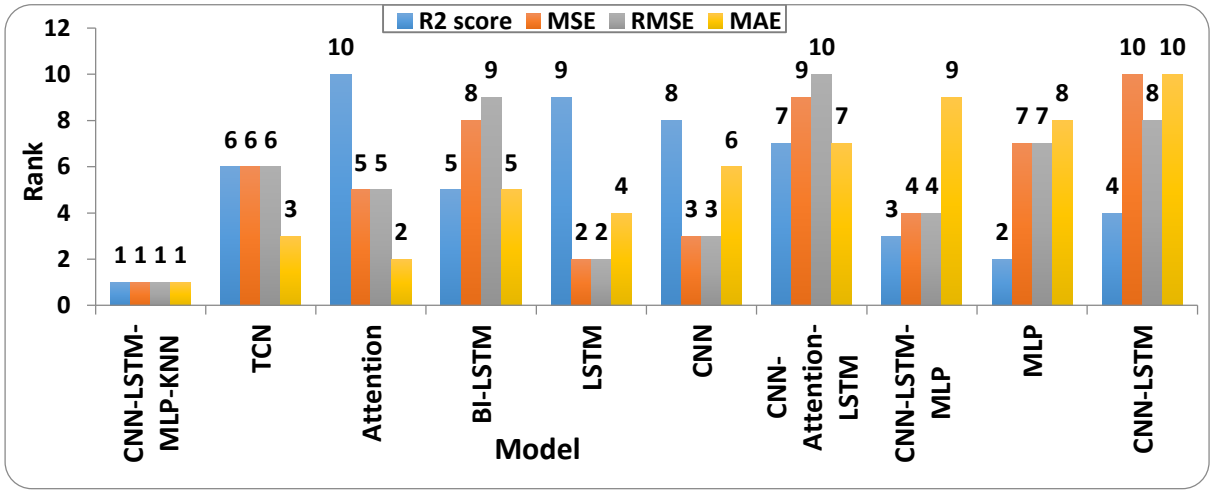
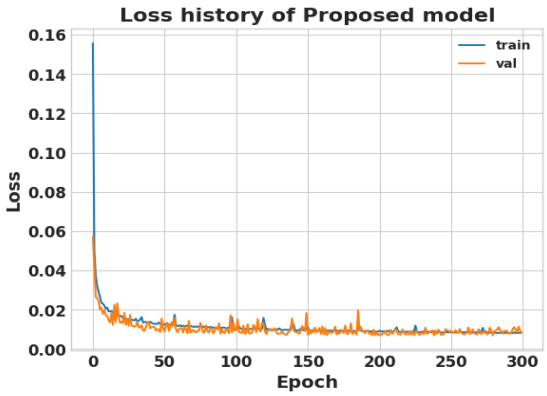
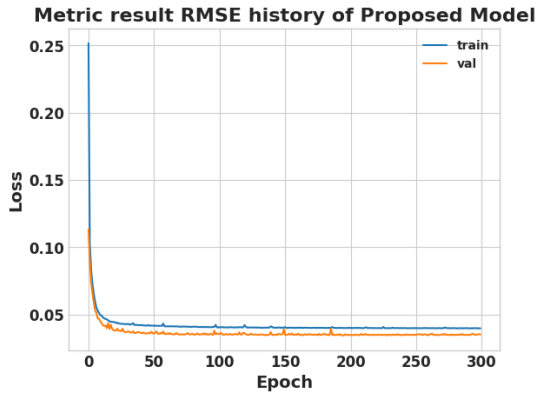
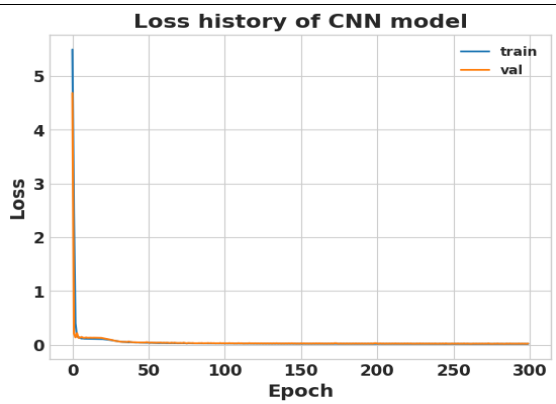
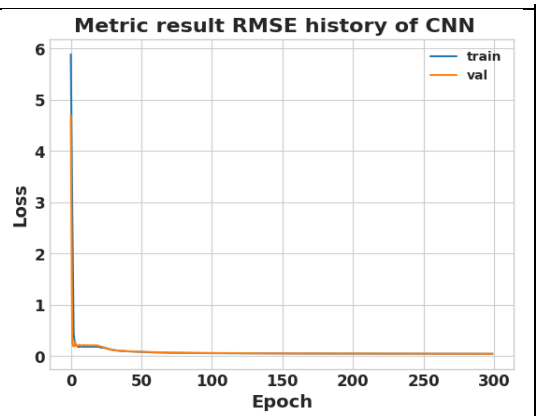
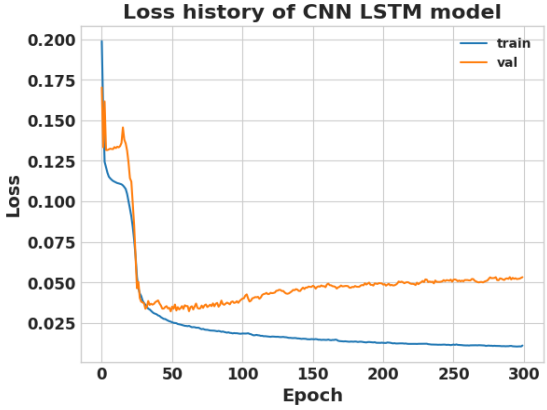
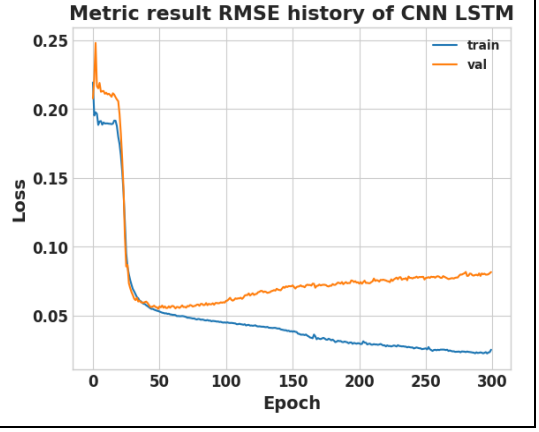


Figure 11. The ranks of the different models based on various performance metrics.

Table 3. Loss functions values and RMSE values during epochs for CNN, CNN-LSTM and the proposed CNN-LSTM-MLP-KNN models.

Model	Loss graphs	RMSE graphs
Proposed		
CNN		
CNN-LSTM		

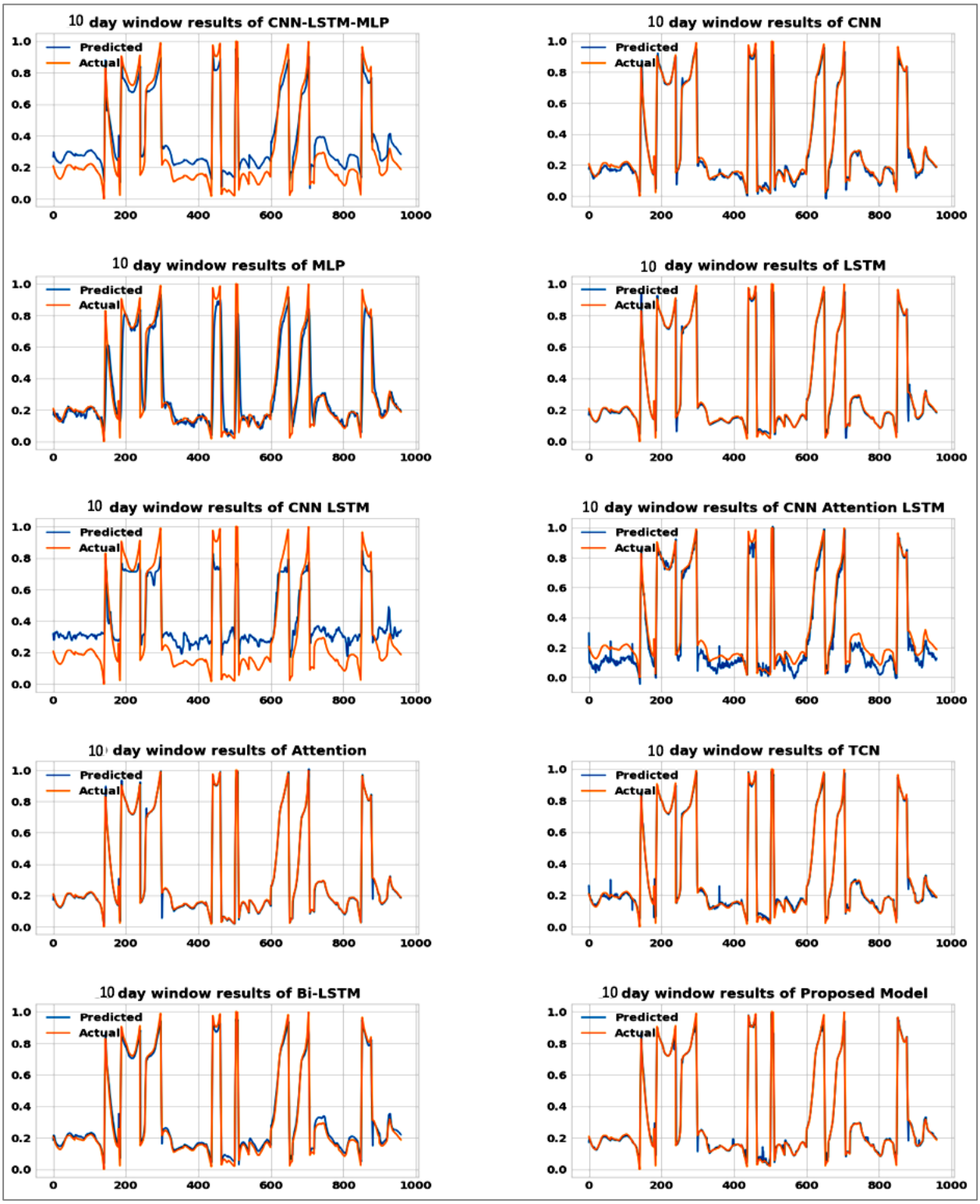


Figure 12. Predictions plots of the different deep learning models.

Table 4. Loss functions values and RMSE values during epochs for LSTM, MLP and the CNN-LSTM-MLP models.

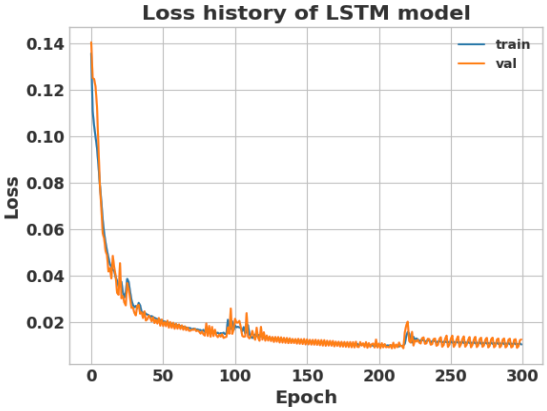
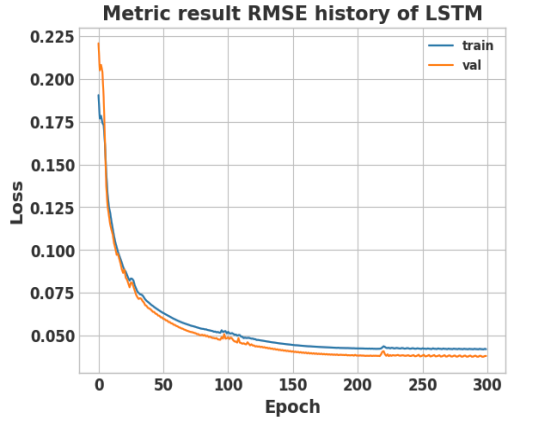
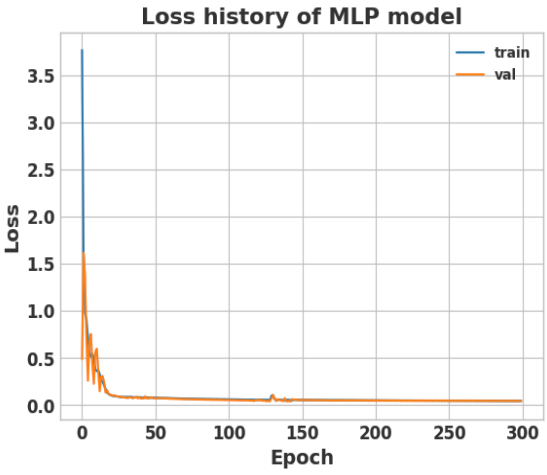
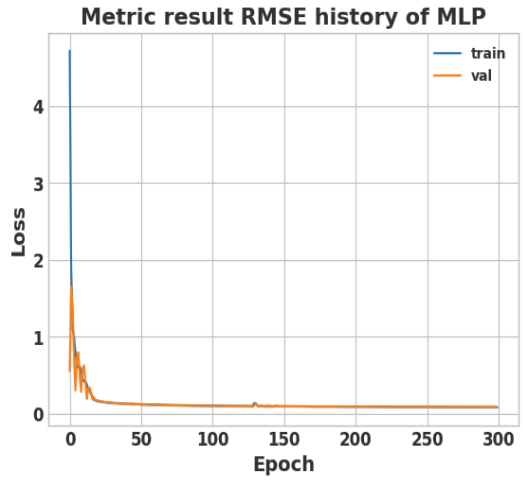
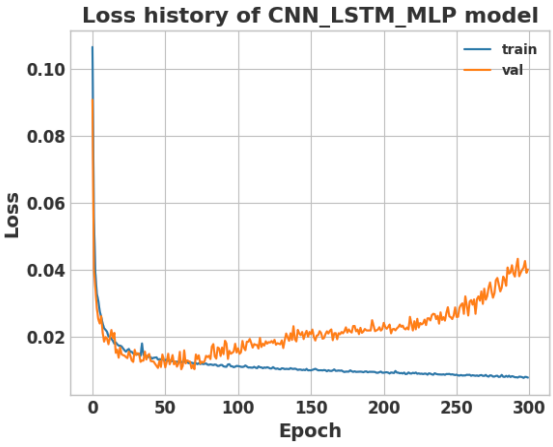
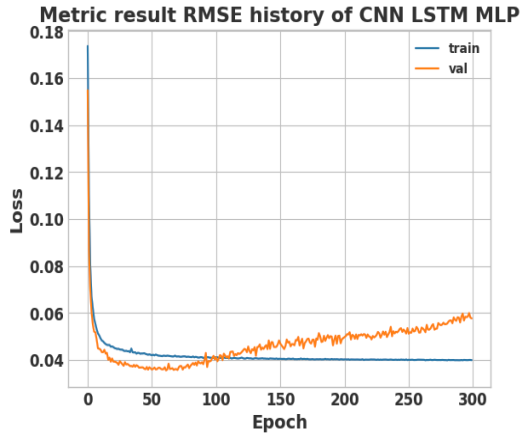
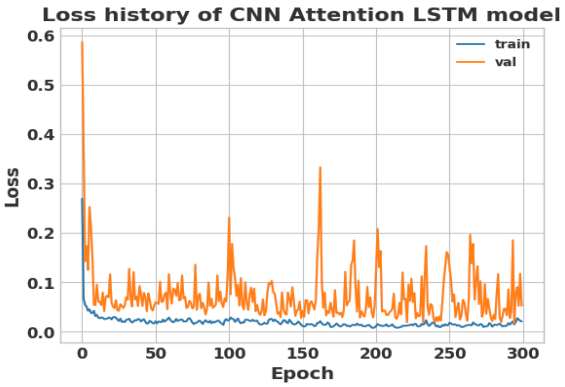
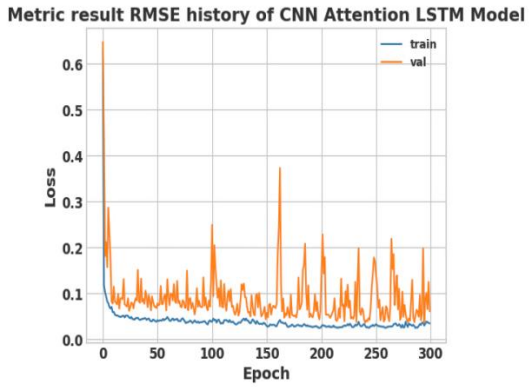
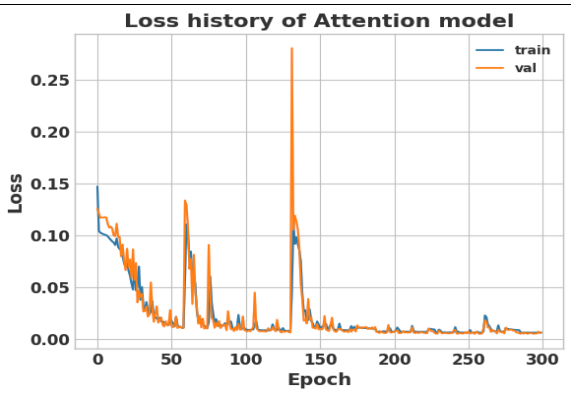
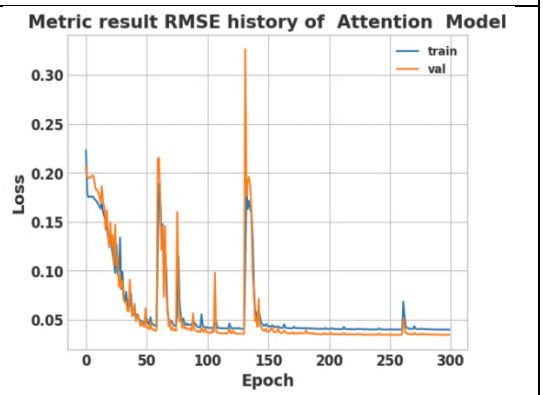
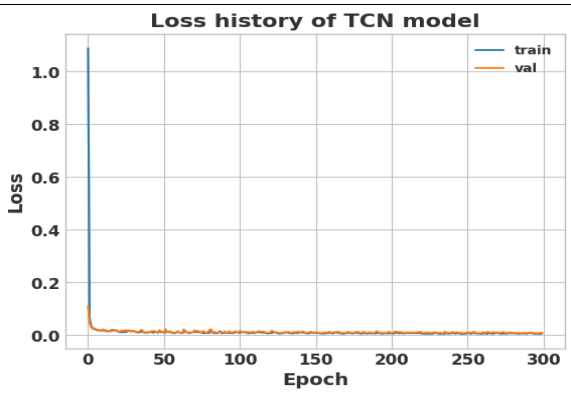
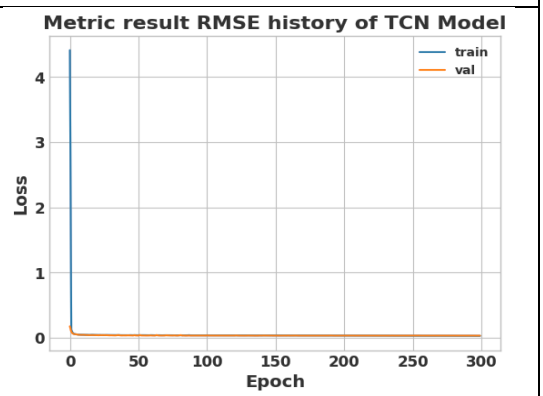
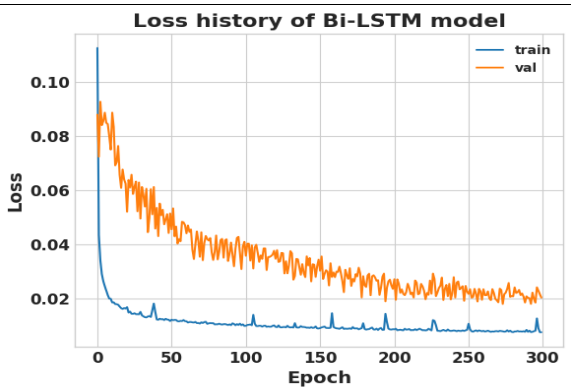
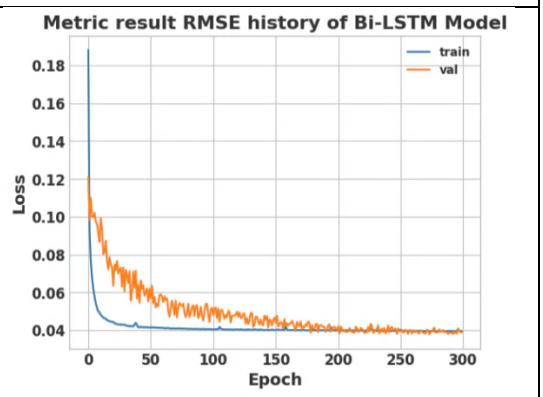
Model	Loss graphs	RMSE graphs
LSTM		
MLP		
CNN-LSTM-MLP		

Table 5. Loss functions values and RMSE values during epochs for CNN Attention LSTM, Attention, TCN and BI-LSTM models.

Model	Loss graphs	RMSE graphs
CNN Attention LSTM		
Attention		
TCN		
Bi-LSTM		

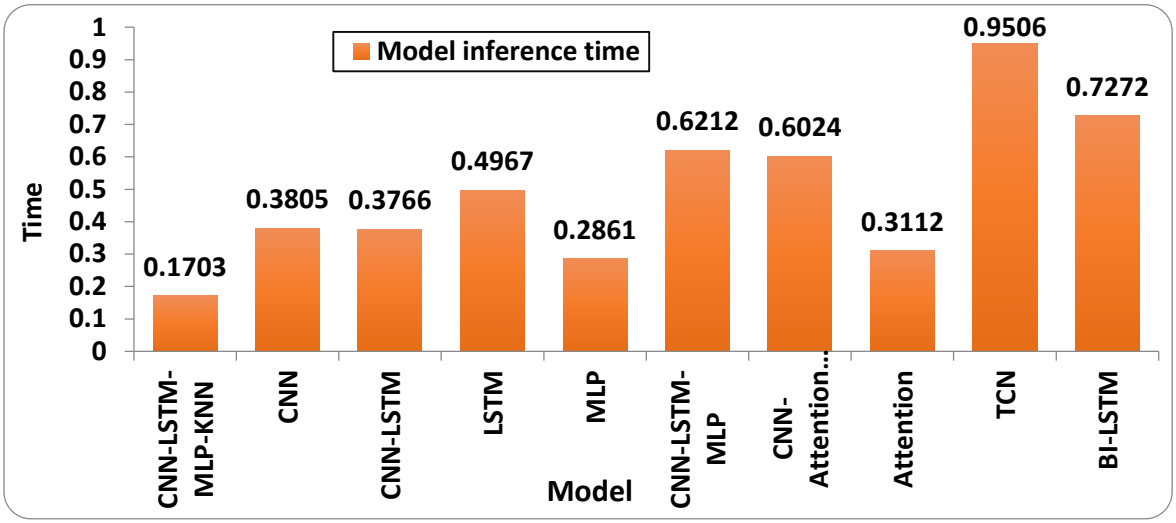


Figure 13. The Inference time of different models.

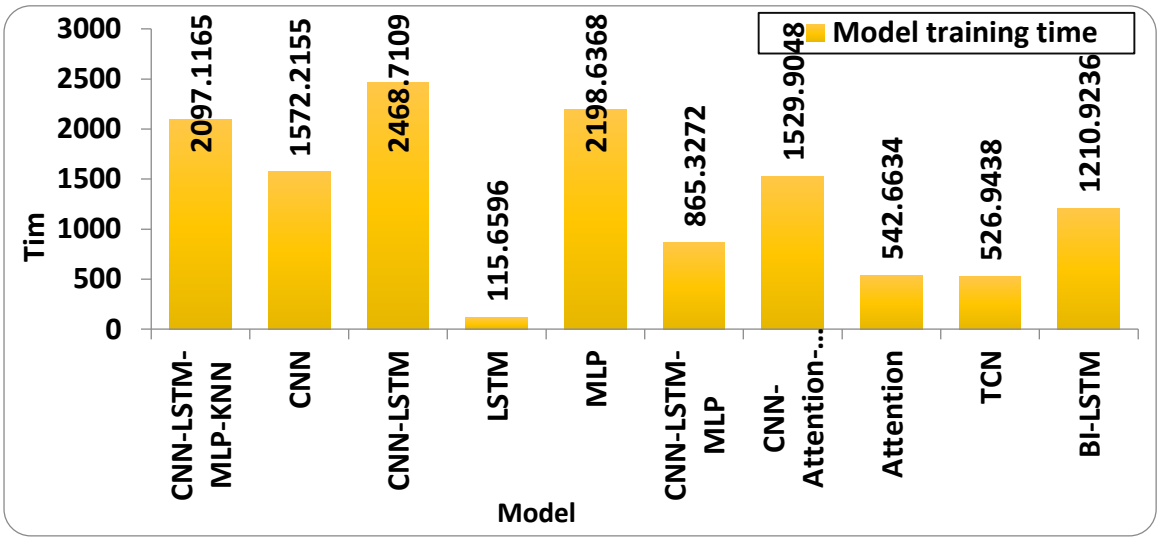


Figure 14. The training time of different models.

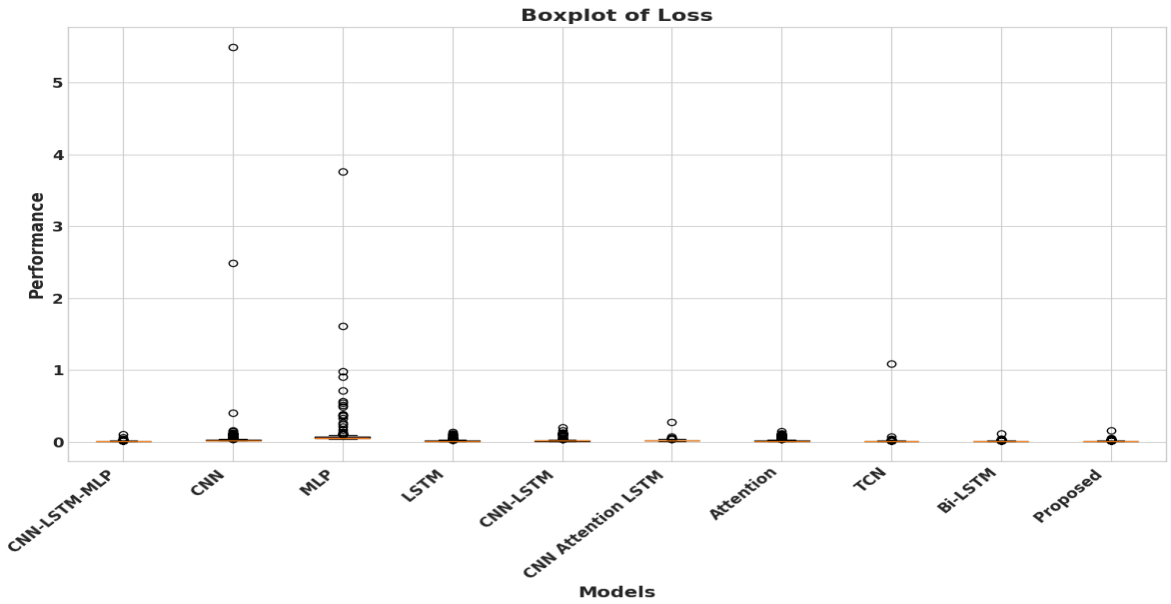


Figure 15. Boxplot of loss values for the different models.

7 | Conclusions and Future Works

This study aims to create a model for predicting solar radiation using advanced artificial intelligence methods. Solar energy is a sustainable alternative to fossil fuels with a broad range of applications. This study utilized a unique dataset named Karachi. The Karachi dataset obtained from the NSRDB, offers precise sun irradiance readings at a 15-minute frequency using reliable meteorological equipment. The dataset spans a 3-year period and consists of 105120 samples with 24 characteristics. The model being proposed is a hybrid that integrates deep learning and machine learning techniques. The project aims to identify spatial and temporal patterns from solar irradiance data by utilizing CNN and LSTM. Additionally, the proposed model employs MLP to analyze and comprehend complex relationships. The KNN regressor algorithm is utilized for non-parametric forecasting by using the nearest neighbors to provide the ultimate solar irradiance projection. Five recently published models were compared with the proposed model. The results showed that the proposed model is more efficient compared to the other five models.

In the future, we aim to apply our model to other forecasting issues, such as weather forecasting [52] and wind power forecasting [53]. Hyperparameter tuning is a crucial task, as it impacts the model efficiency. Hence, we hope to utilize a specific hyperparameters tuning technique to find the optimal parameters' values.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contributions

All authors contributed equally to this work.

Funding

This research was conducted without external funding support.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author declares that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Ajith, M., & Martínez-Ramón, M. (2023). Deep learning algorithms for very short term solar irradiance forecasting: A survey. Renewable and Sustainable Energy Reviews, 182, 113362.

[2] Al-Ja'Afreh, M. A. A., Amjad, B., Rowe, K., Mokryani, G., & Marquez, J. L. A. (2023). Optimal planning and forecasting of active distribution networks using a multi-stage deep learning based technique. Energy Reports, 10, 686-705.

[3] Algarni, S., Tirth, V., Alqahtani, T., Alshehery, S., & Kshirsagar, P. (2023). Contribution of renewable energy sources to the environmental impacts and economic benefits for sustainable development. Sustainable Energy Technologies and Assessments, 56, 103098.

[4] Alghool, D., Khir, R., & Haouari, M. (2024). Optimization and assessment of solar-assisted cooling systems: A multicriteria framework and comparative study. Energy Conversion and Management: X, 22, 100530.

[5] Alharbi, F. R., & Csala, D. (2021). Short-term solar irradiance forecasting model based on bidirectional long short-term memory deep learning. Paper presented at the 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE).

[6] Almarzooqi, A. M., Maalouf, M., El-Fouly, T. H., Katzourakis, V. E., El Moursi, M. S., & Chrysikopoulos, C. V. (2024). A hybrid machine-learning model for solar irradiance forecasting. Clean Energy, 8(1), 100-110.

[7] Azizi, N., Yaghoubirad, M., Farajollahi, M., & Ahmadi, A. (2023). Deep learning based long-term global solar irradiance and temperature forecasting using time series with multi-step multivariate output. Renewable Energy, 206, 135-147.

[8] Bhatt, A., Ongsakul, W., & Singh, J. G. (2022). Sliding window approach with first-order differencing for very short-term solar irradiance forecasting using deep learning models. Sustainable Energy Technologies and Assessments, 50, 101864.

[9] Brahma, B., Wadhvani, R., & Shukla, S. (2021). Attention mechanism for developing wind speed and solar irradiance forecasting models. Wind Engineering, 45(6), 1422-1432.

[10] Chandel, S. S., Gupta, A., Chandel, R., & Tajjour, S. (2023). Review of deep learning techniques for power generation prediction of industrial solar photovoltaic plants. Solar Compass, 8, 100061.

[11] de O. Santos Jr, D. S., de Mattos Neto, P. S., de Oliveira, J. F., Siqueira, H. V., Barchi, T. M., Lima, A. R., . . . Pereira, A. C. (2022). Solar irradiance forecasting using dynamic ensemble selection. Applied Sciences, 12(7), 3510.

[12] Farhidi, F. (2023). Impact of fossil fuel transition and population expansion on economic growth. Environment, Development and Sustainability, 25(3), 2571-2609.

[13] Haider, S. A., Sajid, M., Sajid, H., Uddin, E., & Ayaz, Y. (2022). Deep learning and statistical methods for short-and long-term solar irradiance forecasting for Islamabad. Renewable Energy, 198, 51-60.

[14] Hoseinzade, E., & Haratizadeh, S. (2019). CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Systems with Applications, 129, 273-285.

[15] Hou, X., Ju, C., & Wang, B. (2023). Prediction of solar irradiance using convolutional neural network and attention mechanism-based long short-term memory network based on similar day analysis and an attention mechanism. Heliyon, 9(11).

[16] Kaggle Retrieved 1 July 2024, from https://www.kaggle.com/datasets/ahmdtolba/solar-irradiance-dataset-karachi/data.

[17] Khan, H. A., Alam, M., Rizvi, H. A., & Munir, A. (2023). Solar Irradiance Forecasting Using Deep Learning Techniques. Engineering Proceedings, 46(1), 15.

[18] Khan, J., Lee, E., Balobaid, A. S., & Kim, K. (2023). A comprehensive review of conventional, machine leaning, and deep learning models for groundwater level (GWL) forecasting. Applied Sciences, 13(4), 2743.

[19] Krishnan, N., Kumar, K. R., & Inda, C. S. (2023). How solar radiation forecasting impacts the utilization of solar energy: A critical review. Journal of Cleaner Production, 388, 135860.

[20] Kumari, P., & Toshniwal, D. (2021). Deep learning models for solar irradiance forecasting: A comprehensive review. Journal of Cleaner Production, 318, 128566.

[21] Lai, C. S., Zhong, C., Pan, K., Ng, W. W., & Lai, L. L. (2021). A deep learning based hybrid method for hourly solar radiation forecasting. Expert Systems with Applications, 177, 114941.

[22] Lara-Benítez, P., Carranza-García, M., Luna-Romera, J. M., & Riquelme, J. C. (2023). Short-term solar irradiance forecasting in streaming with deep learning. Neurocomputing, 546, 126312.

[23] Ledmaoui, Y., El Maghraoui, A., El Aroussi, M., Saadane, R., Chebak, A., & Chehri, A. (2023). Forecasting solar energy production: A comparative study of machine learning algorithms. Energy Reports, 10, 1004-1012.

[24] Lelieveld, J., Haines, A., Burnett, R., Tonne, C., Klingmüller, K., Münzel, T., & Pozzer, A. (2023). Air pollution deaths attributable to fossil fuels: observational and modelling study. bmj, 383.

[25] Li, M., Wang, W., He, Y., & Wang, Q. (2024). Deep learning model for short-term photovoltaic power forecasting based on variational mode decomposition and similar day clustering. Computers and Electrical Engineering, 115, 109116.

[26] Li, Y., Qing, C., Guo, S., Deng, X., Song, J., & Xu, D. (2023). When my friends and relatives go solar, should I go solar too?——Evidence from rural Sichuan province, China. Renewable Energy, 203, 753-762.

[27] Liu, J., Huang, X., Li, Q., Chen, Z., Liu, G., & Tai, Y. (2023). Hourly stepwise forecasting for solar irradiance using integrated hybrid models CNN-LSTM-MLP combined with error correction and VMD. Energy Conversion and Management, 280, 116804.

[28] Maibach, E., Kotcher, J., & Patel, L. (2024). We can use our superpower to help end fossil fuel pollution and rise to the challenge of climate change. Journal of Communication in Healthcare, 1-3.

[29] Manandhar, P., Temimi, M., & Aung, Z. (2023). Short-term solar radiation forecast using total sky imager via transfer learning. Energy Reports, 9, 819-828.

[30] Michael, N. E., Hasan, S., Al-Durra, A., & Mishra, M. (2022). Short-term solar irradiance forecasting based on a novel Bayesian optimized deep Long Short-Term Memory neural network. Applied Energy, 324, 119727.

[31] Murugan, D. K., Said, Z., Panchal, H., Gupta, N. K., Subramani, S., Kumar, A., & Sadasivuni, K. K. (2023). Machine learning approaches for real-time forecasting of solar still distillate output. Environmental Challenges, 13, 100779.

[32] Nayak, P. K., Mahesh, S., Snaith, H. J., & Cahen, D. (2019). Photovoltaic solar cell technologies: analysing the state of the art. Nature Reviews Materials, 4(4), 269-285.
[33] Neshat, M., Nezhad, M. M., Mirjalili, S., Garcia, D. A., Dahlquist, E., & Gandomi, A. H. (2023). Short-term solar radiation forecasting using hybrid deep residual learning and gated LSTM recurrent network with differential covariance matrix adaptation evolution strategy. Energy, 278, 127701.
[34] NSRDB. Retrieved 1 July 2024, from https://nsrdb.nrel.gov/
[35] Ogliari, E., Sakwa, M., & Cusa, P. (2024). Enhanced Convolutional Neural Network for solar radiation nowcasting: All-Sky camera infrared images embedded with exogeneous parameters. Renewable Energy, 221, 119735.
[36] Oh, J., So, D., Jo, J., Kang, N., Hwang, E., & Moon, J. (2024). Two-Stage Neural Network Optimization for Robust Solar Photovoltaic Forecasting. Electronics, 13(9), 1659.
[37] Peng, T., Li, Y., Song, Z., Fu, Y., Nazir, M. S., & Zhang, C. (2023). Hybrid intelligent deep learning model for solar radiation forecasting using optimal variational mode decomposition and evolutionary deep belief network-Online sequential extreme learning machine. Journal of Building Engineering, 76, 107227.
[38] Rani, T., Wang, F., Rauf, F., Ain, Q. u., & Ali, H. (2023). Linking personal remittance and fossil fuels energy consumption to environmental degradation: evidence from all SAARC countries. Environment, Development and Sustainability, 25(8), 8447-8468.
[39] Ritchie, H., Rosado, P., & Roser, M. (2024). Fossil fuels. Our world in data.
[40] Salman, A. G., Kanigoro, B., & Heryadi, Y. (2015). Weather forecasting using deep learning techniques. Paper presented at the 2015 international conference on advanced computer science and information systems (ICACSIS).
[41] Saravanakumar, S., JA, B. R., Kumar, D., Janani, T., & Saravana, I. (2024). Solar Energy Powered Advanced Smart Streetlight For Remote Areas. International Research Journal on Advanced Engineering Hub (IRJAEH), 2(04), 1003-1009.
[42] Sarbu, I., & Sarbu, I. (2021). Solar heating and cooling systems: Springer.
[43] Sivakumar, M., George, S. T., Subathra, M., Leebanon, R., & Kumar, N. M. (2023). Nine novel ensemble models for solar radiation forecasting in Indian cities based on VMD and DWT integration with the machine and deep learning algorithms. Computers and Electrical Engineering, 108, 108691.
[44] Smeeth, L., & Haines, A. (2023). COP 28: Ambitious climate action is needed to protect health (Vol. 383): British Medical Journal Publishing Group.
[45] Taud, H., & Mas, J.-F. (2018). Multilayer perceptron (MLP). Geomatic approaches for modeling land change scenarios, 451-455.
[46] Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. Machine learning for brain disorders, 601-630.
[47] Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. Renewable Energy, 105, 569-582.
[48] Wang, H.-z., Li, G.-q., Wang, G.-b., Peng, J.-c., Jiang, H., & Liu, Y.-t. (2017). Deep learning based ensemble approach for probabilistic wind power forecasting. Applied Energy, 188, 56-70.
[49] Wang, J., Yu, L.-C., Lai, K. R., & Zhang, X. (2016). Dimensional sentiment analysis using a regional CNN-LSTM model. Paper presented at the Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers).
[50] Woodruff, T. J. (2024). Health effects of fossil fuel–derived endocrine disruptors. New England Journal of Medicine, 390(10), 922-933.
[51] Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.
[52] Zhang, Z., & Yin, J. (2024). Spatial-temporal offshore wind speed characteristics prediction based on an improved purely 2D CNN approach in a large-scale perspective using reanalysis dataset. Energy Conversion and Management, 299, 117880.
[53] Zhou, H., Zhang, Y., Yang, L., Liu, Q., Yan, K., & Du, Y. (2019). Short-term photovoltaic power forecasting based on long short term memory neural network and attention mechanism. IEEE Access, 7, 78063-78074.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.