Applied Soft Computing Journal 155 (2024) 111441

Contents lists available at ScienceDirect

Applied Soft Computing

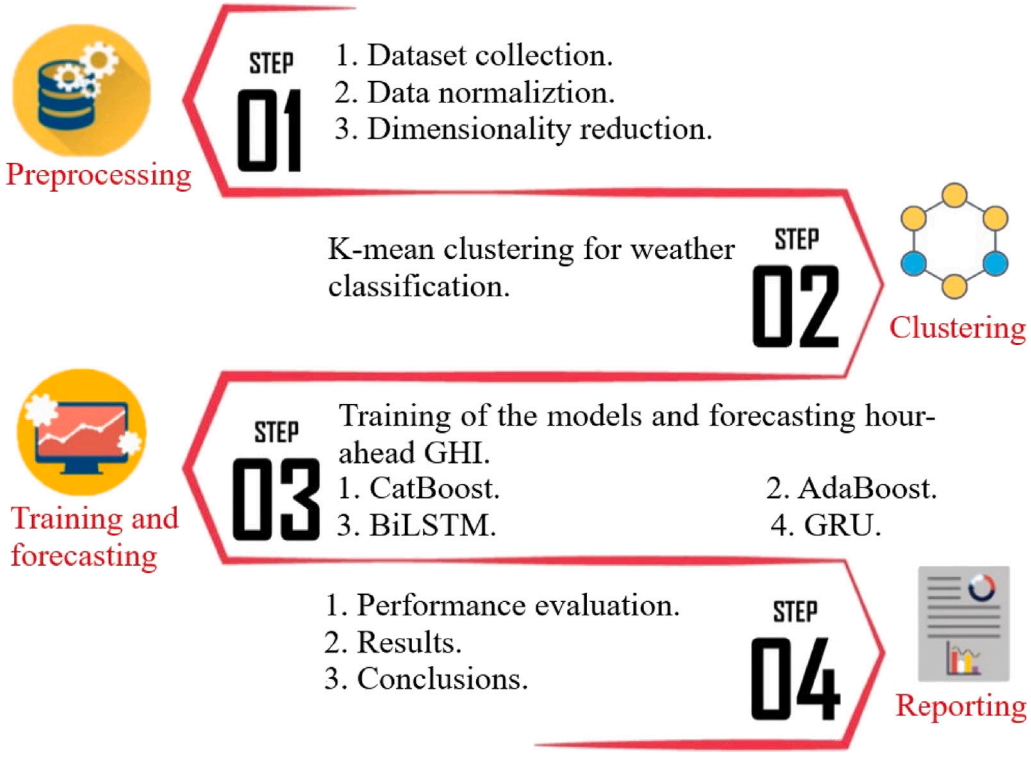
journal homepage: www.elsevier.com/locate/asoc

Short-term global horizontal irradiance forecasting using weather classified categorical boosting

Ubaid Ahmed [a], Ahsan Raza Khan [b,*], Anzar Mahmood [a], Iqra Rafiq [a], Rami Ghannam [b], Ahmed Zoha [b]

[a] Electrical Engineering, Mirpur University of Science and Technology (MUST), Mirpur, 10250, Azad Jammu & Kashmir, Pakistan
[b] James Watt School of Engineering, University of Glasgow, G12 8QQ, United Kingdom

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Global Horizontal Irradiance (GHI) forecasting
Weather classification
Categorical boosting (CatBoost model)

ABSTRACT

Accurate short-term solar irradiance (SI) forecasting is crucial for renewable energy integration to ensure unit commitment and economic load dispatch. However, hourly SI prediction is very challenging due to atmospheric conditions and weather fluctuations. This study proposes a hybrid approach using weather classification and boosting algorithms for short-term global horizontal irradiance (GHI) forecasting. In data pre-processing steps,

* Corresponding author.
E-mail address: a.khan.9@research.gla.ac.uk (A.R. Khan).

https://doi.org/10.1016/j.asoc.2024.111441
Received 29 May 2023; Received in revised form 20 February 2024; Accepted 26 February 2024
Available online 11 March 2024

Bidirectional long short-term memory (BiLSTM)
Random forest (RF)

we employ random forest for feature selection and K-means clustering for weather classification. The weather-based clustered data is used for the model training using categorical boosting (CatBoost). The proposed weather-classified categorical boosting (WC-CB) scheme is compared with benchmarks in literature like adaptive boosting (AdaBoost), bi-directional long short-term memory (BiLSTM) and gated recurrent unit (GRU) using datasets from two distinct geographical locations obtained from the National Solar Radiation Database (NSRDB). The results show that the proposed WC-CB hybrid approach has lower forecast errors compared to conventional CatBoost modelling. The error reduction of 16% and 39% in root mean square error and 6% and 9% in mean absolute error is recorded for the two datasets, respectively. These findings demonstrate the importance of weather classification in improving forecasting accuracy with potential implications for broader renewable energy applications.

1. Introduction

The global energy landscape is changing significantly due to rising energy demand and environmental concerns. According to the International Energy Agency (IEA) report, energy consumption is expected to rise between 15000–21000 Million Tons of Oil Equivalent (MTOE) by 2040 due to a projected 25% population growth [1]. Fossil fuels continue to be the primary energy source, which is approximately 63.1% of the world's energy mix [2]. However, owing to the widespread usage of fossil fuels, these resources are depleting rapidly [3]. Another drawback of the extensive use of fossil fuels is environmental repercussions, particularly global warming, underscore the urgency to shift towards sustainable energy sources [4].

Renewable energy resources (RERs), particularly solar energy, emerge as a promising alternative to fossil fuels due to their abundant, clean, and green attributes. However, the widespread adoption of solar energy is hindered by its inherent intermittent nature, making the planning, management, and maintenance of photovoltaic (PVs) systems challenging. The performance of PV systems is notably susceptible to variable atmospheric conditions, emphasizing the need for accurate and robust solar energy forecasting for the reliable operation of power systems integrated with PVs [5]. PV system forecasting has different applications with different forecasting horizons. Various forecasting applications are shown in Fig. 1, along with their time scales and horizons.

Usually, short-term forecasting solves unit commitment and economic load dispatch problems [6,7]. Unit commitment is one of the basic issues in an electrical power system. It focuses on the key choices to schedule the power production units so that the energy demand is fulfilled at the lowest possible cost [8]. The study of the economic load dispatch problem enables the power system to be operated cost-effectively and efficiently, ensuring an uninterrupted power supply [9].

An essential component of a PV power system is solar irradiance (SI), which is the amount of electromagnetic radiation received from the sun by a particular area of the land, which is expressed in watts per meter square. In the literature, numerous techniques have been proposed for SI forecasting. Numerical weather prediction (NWP) is a physical model based on complex mathematical equations to predict SI. However, the performance of the NWP model becomes vulnerable to severe fluctuations in weather conditions [10]. In recent years, machine learning (ML) models have been developed that forecast the SI with better accuracy. With their ability to learn complex patterns and relationships from historical data, ML models have shown promise in enhancing SI forecasting accuracy. These models, ranging from support vector machines (SVM) to neural networks (NN), have been pivotal in addressing the non-linearities and uncertainties inherent in SI data. Leveraging the data-driven ML technique, the research community has shifted more towards hybrid models, aiming to capitalize on the strengths and mitigate the weaknesses of individual approaches.

For instance, the long short-term memory (LSTM) model with K-mean clustering has been proposed in [11] for forecasting hour-ahead and one-day-ahead SI. The dataset is classified into sunny days and completely and partially cloudy days using the K-mean clustering algorithm. Findings indicate that LSTM performs better than recurrent neural networks (RNN) for hour-ahead SI forecasting. In the case of day-ahead forecasting, RNN draws less error than the LSTM model. In [12], a hybrid model of LSTM with sky image data is used for very short-term forecasting of SI with a time interval of five and ten minutes. For a short-term solar power forecast, a hybrid Mycielski–Markov approach is proposed in [13]. With the combination of stochastic and deterministic techniques, the coefficient of determination (R^2) of the forecasting model is found to be 0.8749.

Accurate solar energy forecasting is of utmost importance in the context of the global push towards sustainable energy sources. The objective is not only to harness solar energy but also to do it in a predictable, reliable, and efficient way. Although existing forecasting methods have made significant progress, they often struggle with rapid weather changes, computational efficiency, and overfitting on complex datasets. Therefore, the balance between the performance and computational efficiency of the forecasting approach is very crucial, particularly given the critical role of solar energy in modern power systems. Therefore, keeping these challenges in mind, this study introduces the weather-classified categorical boosting (WC-CB) algorithm, a novel hybrid approach that seamlessly integrates the strengths of multiple techniques. Unlike traditional LSTM and RNN hybrid models, WC-CB is designed to handle rapid weather fluctuations efficiently and effectively captures both sequential and non-sequential patterns in historical data, ensuring robustness against overfitting and enhanced computational efficiency. This holistic approach streamlines the forecasting process by focusing on salient features using random forest (RF) and weather-specific data clusters using K-mean clustering. The weather classification aids in reducing the inherent variability and unpredictability associated with solar energy. Finally, categorical boosting (CatBoost) is used for model training using clustered data. Our hybrid model combines feature selection, data clustering, and gradient boosting to provide an accurate and robust forecast for SI forecasting. The main contributions of this paper are listed as follows:

1. This study introduces the hybrid approach WC-CB for short-term GHI forecasting. The proposed scheme combines RF for feature selection, K-mean clustering for weather categorization, and the CatBoost algorithm for model training.
2. The robustness and efficiency of the WC-CB methodology are evaluated utilizing two distinct datasets sourced from the National Solar Radiation Database (NSRDB). The performance is benchmarked against adaptive boosting (AdaBoost), bidirectional (Bi-LSTM), and gated recurrent unit (GRU). Additionally, the research compares the performance of WC-CB and a traditional CatBoost model without clustering.
3. A comprehensive analysis of clustering techniques, comparing one-dimensional and two-dimensional K-means clustering for classifying GHI data, is conducted. The findings showed that single-parameter clustering outperforms two-parameter clustering in terms of predictive performance, making it more appropriate for forecasting purposes.

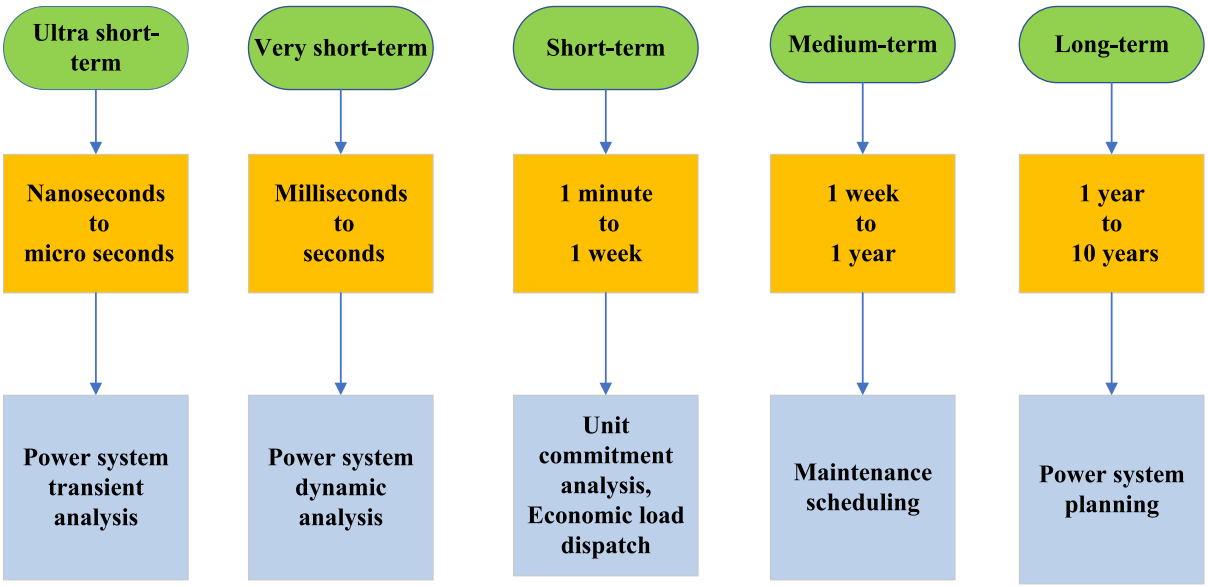


Fig. 1. Forecasting horizons with their time scale and applications.

The remaining portion of the paper is organized as follows. Section 2 presents a literature review, the proposed hybrid WC-CB methodology is in Section 3. Section 4 covers the data processing and model configurations, followed by results and discussion in Section 5. The conclusions are presented in Section 6.

2. Literature review

Different models have been proposed in the literature for forecasting SI, which can be broadly classified into two types: physical and data-driven [14]. NWP is a type of physical model that uses differential equations to describe the changes in the atmosphere for prediction. The forecasting performance of the NWP model depends on the amount of available data and the complexity of weather conditions. If the forecasting period is longer, the performance of the NWP model worsens. This is because the model might need to interpolate more data and variability of weather conditions [15]. On the other hand, data-driven models harness historical data, sometimes enriched with external weather parameters like temperature, relative humidity, wind speed, and pressure, to make predictions. This literature review will explore statistical models, ML-based techniques, boosting algorithms, and hybrid approaches in SI forecasting, aiming to provide a comprehensive overview of the diverse methodologies, emphasizing their merits, constraints, and performance outcomes.

2.1. Statistical models for SI forecasting

Statistical models, particularly autoregressive (AR) and Markov chain (MC), have been prominently utilized for SI forecasting. These models, grounded in statistical theories, offer insights into the temporal patterns of SI data. For instance, the study in [16] proposed autoregressive moving-average (ARMA) for hour-ahead SI forecasting and compared it with the smart persistence (SP) model. The results demonstrate the effectiveness of ARMA over the SP model. Hour-ahead SI forecasting using autoregressive integrated moving average (ARIMA) is proposed in [17]. The datasets used in this study were collected from the weather stations in Miami and Orlando. Different combinations of input variables are provided to the ARIMA model. In the first combination, GHI data is the only input parameter used for hour-ahead forecasting. Secondly, direct normal irradiance (DNI) and direct horizontal irradiance (DHI) are forecasted individually, and their results are combined to predict GHI. Finally, the cloud cover effect is also considered an input combination that outperforms the other two approaches.

In [18], the ARIMA model predicts hour-ahead GHI. The performance of the model is evaluated through the root mean square error (RMSE) and R^2 score. The RMSE and R^2 values were found to be 72.88 Wm^{-2} and 88.63%, respectively. In [19], a second-order MC model is presented that forecasts day-ahead SI for the cities of Bhadla, Jodhpur, and Rajasthan. The method uses a similarity algorithm to segment the data into similar groups. The proposed model formulates an ordered transition matrix of size 81×9, calculated using the data's mean and standard deviation. This matrix is then used to forecast the SI. In [20], a comparison of the MC model with ARIMA, artificial neural networks (ANN), and support vector machines (SVM) for day-ahead SI forecasting is presented. The performances of the models are evaluated using RMSE, mean absolute error (MAE), and mean absolute percentage error (MAPE). Results demonstrate that the second-order MC model outperforms other techniques. Short-term SI forecasting using ARIMA and ANN models is presented in [21]. In the proposed work, datasets from five different geographical areas are used for model training, and normalized root mean square error NRMSE, normalized mean absolute deviation (NMAD), and normalized mean bias error (NMBE) are used as performance indicators. Results demonstrate that the ARIMA model performs better than the ANN model. Findings also indicate that the performances of ARIMA and ANN models are better for continental sites than island sites. In [22], an experimental model is carried out for improving the power generation capacity of the PV panels. The thermoelectric system is integrated with the PV panel and uses the temperature difference between the panel and the environment to enhance the power generation capacity. Results demonstrate that the combined system generates 4.2% more power than the conventional PV system. In [23], the authors highlight the importance of multi-cycle production development planning strategies for increasing the share of RERs in sustainable power systems. Four different multi-cycle production development planning strategies, hierarchical production planning, multi-criteria decision support, generation expansion planning and multi-energy complementary systems, are discussed for maximizing the share of RERs. The Weather Research and Forecasting (WRF) model is presented for day-ahead SI forecasting for the dataset of Singapore in [24]. The comparison of the WRF model with persistence, exponential smoothing (ES) and seasonal autoregressive

integrated moving average (SARIMA) models is also studied. The WRF approach shows better performance than other models. Moreover, the findings also depict that if the WRF and ES forecasting outputs are combined, the error can be reduced to 49%.

2.2. Machine learning models for SI forecasting

The artificial intelligence (AI)-based models, particularly ML, forecast SI with better accuracy because of their tendency to learn the non-linear relationship between inputs and output of the prediction task. RNN, feed-forward neural networks (FFNN), SVM, LSTM and adaptative fuzzy neural networks (AFNN) are ML models that are mostly used in forecasting SI. A comparative study of SVM with facebook propet (FBP) is performed in [25], in which the datasets contain parameters of three locations: Boston, Denver and Seattle. Results indicate that for one and two-hours-ahead forecasting, SVM performs better than the FBP model. However, for 3 hours-ahead forecasting, the performance of SVM degrades. In [26], FFNN is used for day-ahead SI forecasting. The dataset of 19 years is collected from the meteorological station of Ajaccio, France. The performance of FFNN is compared with ARIMA, MC Bayesian Interference and k-nearest neighbor (KNN). FFNN beats other models with NRMSE found to be 21%. For hour-ahead GHI forecasting, a comparative study of FFNN, LSTM and SVR is presented in [27]. The dataset is obtained from a radiometric station at the University of Pretoria. Results indicate that FFNN draws less error than other models. Moreover, the results of ML models are combined using Quantile Regression Average (QRA). Findings demonstrate that QRA shows better performance than ML models with best-recorded RMSE and MAE, which are 34.87 and 20.039 Wm^{-2}, respectively. The ANN-based model is presented in [28] to forecast day-ahead GHI and DNI, for which the dataset is collected from the US National Weather Service (NWS) database. The gamma test (GT) and genetic algorithm (GA) are used for the selection of relevant input parameters in the proposed model. The ANN with GT and GA as feature extraction techniques is compared with conventional ANN. Results show an improvement of 10%–15% in the RMSE of ANN, which is trained by the relevant parameters as compared to the conventional ANN model. Short-term SI forecasting using deep recurrent neural network (DRNN) is performed in [29], and K-mean clustering is used to separate night and day hours data. The night hours data, where SI becomes zero, are removed from the dataset. DRNN is compared with SVM and FFNN using RMSE, MSE and main bias error (MBE). DRNN shows better performance with MBE of 0.003 Wm^{-2}.

Deep learning (DL), the branch of AI, has attracted the attention of many researchers in recent years. Deep learning networks (DLNs) have the feature of multiple hidden layers that enable them to learn the data pattern accurately. DLNs find applications in classification, computer vision, forecasting and natural language processing [30]. Different DLNs are used for forecasting SI in literature. For the short-term forecasting of Florida's PV power plant, the LSTM model is used in [31]. In the proposed study, the pearson correlation coefficient (PCC) is used for dimensionality reduction and sky-type classification of SI data is performed by K-mean clustering. Findings indicate that LSTM performs better than generalized recurrent neural network (GRNN) and extreme learning machine (ELM). A comparative analysis of LSTM with SVM on the dataset of Johannesburg is performed in [10], in which NRMSE is used for the performance evaluation of the models. Results demonstrate the superiority of LSTM over SVM. A comparison of the LSTM model with FFNN, SVM and persistence model for day-ahead SI forecasting is presented in [14]. K-mean clustering is used to divide the data into sunny and cloudy days, and PCC is used for extracting suitable features. The datasets contain meteorological parameters of three locations. LSTM performs better than other models for each location. In [11], hour-ahead and day-ahead SI forecasting is performed using LSTM. The dataset is classified into sunny days and completely and partially cloudy days using the K-mean clustering algorithm. The LSTM model is compared with ARIMA, SVM, RNN, convolutional neural networks (CNN) and back-propagation neural networks (BPNN) using RMSE, R^2 and MAE. Findings indicate that for hour-ahead SI forecasting, LSTM draws less error than other models. In the case of day-ahead forecasting, RNN performs better than the LSTM model. A comparative study of LSTM, linear least square regression and multilayered FFNN with back-propagation for hour-ahead SI forecasting is presented in [32]. The 11 years of historical SI data of Santiago, Cape Verde, is used in the study. Results indicate that the RMSE for the LSTM model is improved by 42.9% compared to FFNN with back-propagation.

2.3. Boosting algorithms in SI forecasting

In the literature, some boosting algorithms are also proposed for forecasting applications. Boosting algorithms aim to enhance the forecasting power by training the weak models, each of which addresses the shortcomings of the precursor. It is contrary to the ML algorithms, which have a single model to concentrate on for accurate forecasting [33]. In SI forecasting, boosting algorithms are very popular, and various studies are presented to show the effectiveness of these algorithms. For instance, the AdaBoost regressor model is presented in [34] for day-ahead SI forecasting. The dataset consisting of four months is collected from the HI-SEAS meteorological station. AdaBoost regressor performance is compared with RF regressor and linear regressor model using RMSE, MAE and MSE. The RMSE is found to be 135.77, 164.76 and 195.4 Wm^{-2} for AdaBoost, RF and linear regressor respectively. A hybrid model of a gradient boosting (GB) algorithm with NWP for short-term SI forecasting is performed on the dataset of San Diego city in [35]. The effectiveness of the hybrid approach is evaluated by RMSE, MAE, MSE and MAPE. The forecasting interval is of 30 minutes, and RMSE for three different seasons, winter, summer and spring, is found to be 6.6, 6.2 and 6.3 Wm^{-2} respectively. A comparative analysis of RF, ANN and SP models for GHI, DHI and beam normal irradiance (BNI) forecasting is presented in [36]. Historical SI data from Odeillo, France, is used in the proposed study. Models' performances are evaluated through RMSE, MAE, NRMSE and NMAE. Results indicate that the RF model outperforms other techniques. The RMSE of 88.62, 189.50 and 48.53 are recorded for hour-ahead forecasting of GHI, BNI and DHI, respectively. In [37], a hybrid model of extreme gradient boosting forest (XGBF) with deep neural network (XGBF-DNN) is proposed for hour-ahead GHI forecasting. The meteorological parameters of three different locations of India, New Delhi, Jaipur, and Gangtok, are included in the datasets. The XGBF-DNN model performance is compared with SP, support vector regressor (SVR), extreme gradient boost (XGBoost), RF and DNN. Results indicate the effectiveness of the hybrid XGBF-DNN technique over other models. A comparative study of SVM and with XGBoost for hour-ahead GHI forecasting is performed in [38]. Findings demonstrate that the forecasting accuracy XGBoost model is better than the SVM model. Moreover, the computational speed of the XGBoost model is found to be 3.07 s as compared to 31.61 s of SVM.

2.4. Hybrid models for SI forecasting

Hybrid models are a combination of different methodologies that aim to utilize the advantages of each technique while minimizing their limitations. In the domain of SI forecasting, hybrid models have gained popularity due to their promising potential for providing better accuracy and resilience. For instance, a hybrid model of MC with the 'persistence approach' and 'neighbor inference approach' is compared with the 'persistence' model in [39] on the datasets containing input parameters of four locations: Athens, Bucharest, Berlin and Helsinki. The hybrid model performs better than the persistence model on each dataset. In the case of Berlin city, an average improvement of 2.5 Wm^{-2} in the MAE for each month is achieved over the persistence model. A hybrid LSTM model with meta-heuristics bio-inspired algorithm cuckoo

search (CS) for day-ahead SI forecasting is presented in [40]. The PCC is used for feature extraction, and hyper-parameters of the LSTM model are optimized using CS and improved cuckoo search (ICS) algorithms. For comparison between ICS-LSTM and CS-LSTM models, RMSE and MAE are used as performance indicators. Findings depict that ICS-LSTM produces better results than the CS-LSTM model. For 5- and 10-minute intervals forecast, the hybrid model of LSTM with sky image data is performed in [12]. Findings show that for a 5-minute interval forecast, LSTM with sky image data produces better results. While the performance of the model without a sky image is better for a 10-minute interval forecast. A hybrid approach consisting seasonal clustering forecasting technique (SCFT) with an LSTM model for hour-ahead SI forecasting is performed in [41]. The datasets of 19 years of six different locations are collected from NSRDB. The dataset is first clustered into 4 types using a seasonality clustering algorithm. In the next stage, further classification of data into sunny, cloudy and rainy hours is performed by the K-mean clustering algorithm. 3D-LSTM model is then trained on these clusters of data. The best recorded RMSE is 13.48 Wm^{-2}, which is for the dataset of Tripoli, Libya. A hybrid model of deep LSTM with the aggregation function based on the choquet integral is presented in [42]. The dataset contains parameters of six different locations in Finland. The proposed model's performance for hour-ahead SI forecasting is evaluated using RMSE, which is found to be 26.71, 30.33, 29.88, 20.77, 30.32 and 19.72 Wm^{-2} for six different sites. A hybrid technique of the CNN-LSTM model for hour-ahead GHI forecasting for the dataset of Texas, USA, is studied in [43]. CNN network is used to extract the spatial features then the LSTM model is applied with spatiotemporal correlation for hour-ahead prediction. The proposed technique is compared with CNN and LSTM models. Results demonstrate the superiority of the hybrid CNN-LSTM approach over the CNN and LSTM models.

Various techniques for SI forecasting have been explored in the literature, including statistical models, ML algorithms and hybrid approaches. While these methods have merits, some key limitations create scope for improving accuracy and robustness. For instance, the primary advantage of statistical techniques lies in their ability to capture linear relationships and patterns in time series data, making them suitable for short-term forecasts. However, their performance can be limited when dealing with non-linearities or when external factors like weather conditions influence the SI. On the other hand, ML models such as SVM, ANN and LSTM offer enhanced performance compared to statistical techniques. The DL models often require huge amounts of data for model training, and their black-box nature makes them less interpretable. Despite these challenges, their ability to adapt and learn from non-linear data makes them a promising avenue for SI predictions. However, complex DL architectures like LSTM and RNN are prone to overfitting and have high computational costs.

Ensemble learning techniques, such as boosting algorithms, have become popular in SI forecasting due to their ability to adapt and handle complex patterns in data. Though boosting algorithms have advantages such as flexibility and reduced overfitting, they can be computationally demanding and sensitive to hyper-parameter tuning. They may have interpretability issues, which require careful consideration before deployment. To overcome these challenges, hybrid approaches have been proposed to combine the strengths of different techniques while mitigating their weaknesses. However, the success of these models depends on a well-designed methodology. In particular, weather conditions significantly impact the performance of SI forecasting models. Thus, data must be categorized based on weather conditions. Additionally, as the complexity of data increases, there is a higher risk of overfitting and computational challenges. Therefore, feature engineering is crucial in selecting the most relevant features for the model, reducing computational demands, and minimizing overfitting. As a solution, this study proposes a WC-CB algorithm that combines RF and K-means clustering for feature selection and weather categorization. Finally, CatBoost is used for prediction, ensuring a balance between performance and computational efficiency.

3. Proposed methodology

The complete methodology of the proposed WC-CB approach consists of many steps, as shown in Fig. 2. The first step is the data collection and data pre-processing, followed by K-mean clustering. In the third step, model training is done using the clustered data and performance evaluation is done in the fourth step. The details of each step of our proposed scheme are discussed as follows:

1. The SI data of two distant geographical locations is collected from NSRDB [44]. This data is enriched with a plethora of weather parameters, including DHI, DNI, temperature, pressure, wind speed, wind direction, and cloud type. In the first step, data preprocessing is done, which includes data normalization and dimensionality reduction.
2. K-mean clustering algorithm is executed based on clear sky DHI for the weather classification of data into cloudy and sunny hours. Moreover, to access the effect of clustering based on multiple parameters, the classification of data with two parameters based on clear sky DHI and cloud type is performed.
3. In the third step, the CatBoost model is trained, and results are compared with models like AdaBoost, BiLSTM and GRU using weather-classified data for a fair comparison.
4. Finally, the models' performances are evaluated using different error measurement techniques and results are documented.

The details of each step are further elaborated in the subsequent section.

3.1. Data preprocessing

Data preprocessing is one of the crucial steps, ensuring high-quality structured data as input to improve the predictive performance. In this study, the preprocessing phase includes data normalization and feature engineering. Data normalization is an important process when dealing with raw data that comes from diverse sources or sensors. This type of data can have discrepancies in scale and range, which can negatively impact the performance of ML models. Normalization solves this issue by transforming all numeric columns to a standard scale, thus ensuring that no particular feature has an unfair influence on the model due to its range. In this work, we use the Min–Max scaler, which rescales every feature to the interval [0, 1]. This means that the smallest value in the dataset becomes 0, the largest becomes 1, and all other values are adjusted proportionally [45]. Using this scaling method retains the original distribution of the data, ensuring that relationships between values remain intact.

When dealing with complex datasets, the amount of data can be enormous. However, not all features contribute equally to the predictive power of a model. Some may be redundant or even cause noise, which can lead to sub-optimal performance. Therefore, it is important to perform feature selection. In this study, RF is used due to its effectiveness in feature selection. The RF is a supervised learning algorithm that constructs multiple decision trees during training and provides the mean prediction for regression. One of its major strengths is the ability to calculate a feature importance score that indicates the contribution of each feature to the prediction [46]. The RF's built-in estimator function gauges the significance of each feature concerning the target variable [47].

3.2. K-mean clustering algorithm

Clustering algorithms are broadly classified into two types: partitional and hierarchical. A non-overlapping group of data is formed in partitional clustering, while in hierarchical clustering, a set of hierarchical clusters is created by a distance matrix [48]. K-mean clustering is a partitional clustering method first introduced by Stuart Lloyd [49].

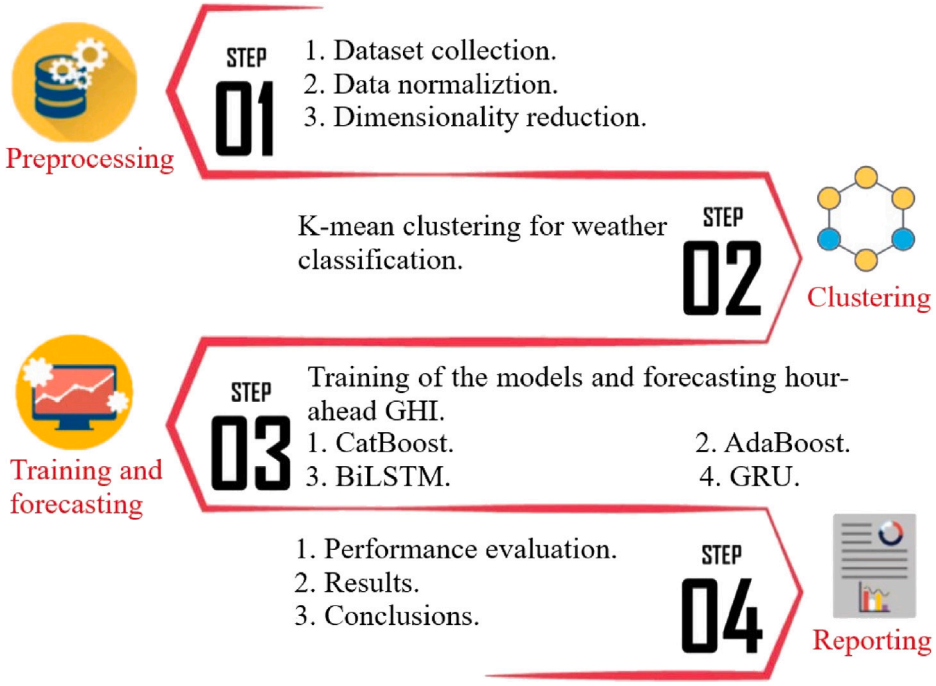


Fig. 2. The flow diagram of proposed WC-CB.

The number of centroids in which data is divided is predetermined. The data points are clustered into different groups by evaluating the mean distance between them. The mechanism is repeated over time to give accurate classification [14]. By iteration, the algorithm tries to minimize the square error function to link data points and centroids.

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n} \left\| x_{ij} - m_i \right\|^2, \tag{1}$$

where $\left\| x_{ij} - m_i \right\|^2$ is the distance between data points x_{ij} and cluster point m_i. There are many ways to determine the number of clusters into which dataset is divided. In the present study, we use the elbow method to determine the optimal number of clusters into which the dataset is divided [50]. Weather data is complex, with multiple variables, including temperature, humidity, wind speed, and others, that interact in intricate ways. Clustering, particularly K-means, provides a means of identifying unique patterns within this complexity, which is essential for SI forecasting. In essence, by segmenting the data into meaningful clusters, models can be tailored to specific weather patterns, leading to improved performance and more nuanced insights.

3.3. Categorical boosting (CatBoost)

CatBoost is a GB on decision trees algorithm developed by Yandex researchers and engineers. CatBoost algorithm finds its application in recommendation systems, personal assistance, self-driving cars and forecasting tasks [51]. CatBoost has two distinct features over other GB algorithms, and that is ordered boosting and efficient dealing with categorical features [52]. In this algorithm, a complex ensemble learning technique is used in which decision trees are created sequentially. During training, the decision trees are created in such a way that each subsequent tree learns from its forerunner to minimize the loss function [53]. Unlike other GB algorithms, CatBoost work on oblivious trees. In oblivious trees, only one feature is selected on a specific level of a tree. The decision rules about splitting criteria are made according to that specific feature. In other boosting algorithms, the weak learners are enhanced during each iteration which causes the over-fitting of a final learner. Because of the oblivious trees chance of over-fitting is low, and the execution speed of the CatBoost model also improves [54]. In the ordered boosting of the CatBoost N different supporting model say $(N_1, N_2, \ldots, N_n)$ are created. Each new model is trained on the residual training set unseen by the previous model. At the first stage $a+1$, independent random permutation $\{\sigma_1, \sigma_2, \ldots, \sigma_s\}$ for the training set are generated. The leaf values from the obtained tree are selected using permutation σ_0. Let us consider we have N_{rj} model and $N_{rj}(i)$ is the prediction of the model at the ith example with permutation σ_r. During model execution, the permutation from the set $\{\sigma_1, \sigma_2, \ldots, \sigma_s\}$ is sampled for the construction of the tree T_t. The gradient for the corresponding prediction $N_{rj}(i)$ is calculated as

$$grad_{rj} = (\partial L(b_i, a))/\partial a, a = N_{r,j}(i) \tag{2}$$

Where b represents the target variable [55]. The CatBoost algorithm is mostly used for classification tasks. In the proposed study, we implement CatBoost model for time series forecasting in hybridization with the K-mean clustering algorithm.

3.4. Adaptive boosting (AdaBoost)

Freund and Schapire first introduced AdaBoost algorithm, which is widely used in different sectors with different applications [56]. AdaBoost is a learning algorithm in which more attention is paid to weak classifiers of the base learner. The AdaBoost algorithm also works sequentially. At first, a base learner is divided into a weak and strong classifiers. At each iteration, weak classifiers are enhanced by adding sample weight to improve the model's performance. The next base learner is trained by these added samples [57]. AdaBoost overcomes the two main problems with the other boosting algorithms: the adjustment of a weak classifier with a training dataset and the combination of a trained weak classifiers to create a strong classifier. The AdaBoost algorithm was first introduced for classification purposes. However, in recent years it also finds its application in different regression problems e.g., for day-ahead SI forecasting [34].

3.5. Bidirectional long short-term memory network (BiLSTM)

The RNN model has two shortcomings. First is the carrying and retrieval of information over a long period. Second, the vanishing

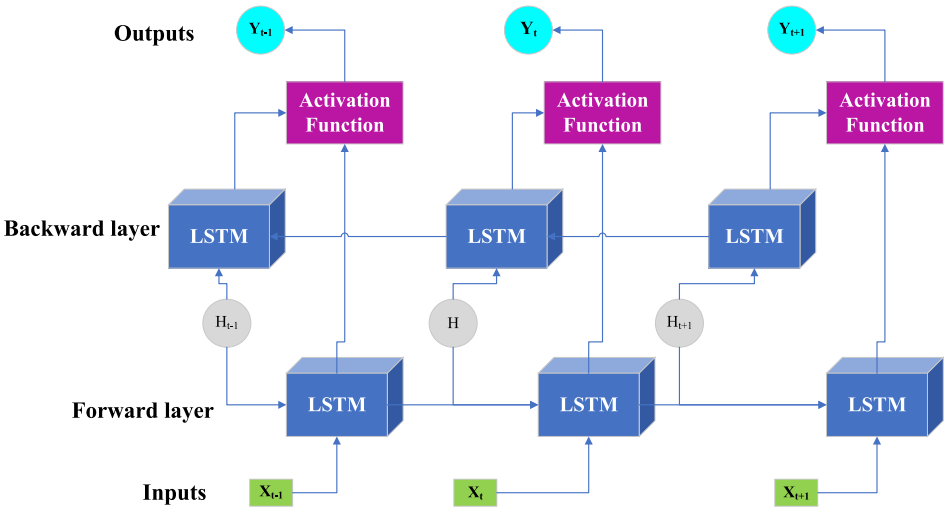


Fig. 3. Bidirectional LSTM structure.

gradient is because of backpropagation. To overcome the associated problems of RNN, the LSTM model is developed as one of the variants [58]. LSTM has a gated structure consisting of input, output and forget gates. Sigmoid activation and 'tanh' activation functions, in conjunction with the gating logic, at layers architecture, carry the information over a long period and overcome the vanishing gradient problem [31]. In this paper, for comparative analysis, we use the BiLSTM model for the prediction of hour-ahead GHI. Both forward and backward time series are used by the BiLSTM model to extract more information about past and future timestamps. This bidirectionality feature provides more information to the network [59]. This is contrary to the conventional LSTM model in which the flow of input is either in a forward or backward direction. This bidirectionality feature improves the models' forecasting accuracy. The basic structure of BiLSTM is shown in Fig. 3. The input layer data X_{t-1} together with the outputs of forward and backward layers gives the output Y_{t-1} at the output layer. The BiLSTM works under the following equations [60].

$$f(\tau) = \sigma[W_f x(\tau) + U_f h(\tau - 1) + b_f] \tag{3}$$

$$i(\tau) = \sigma\left[W_i x(\tau) + U_i h(\tau - 1) + b_i\right] \tag{4}$$

$$c_o(\tau) = \varphi[W_c x(\tau) + U_c h(\tau - 1) + b_c] \tag{5}$$

$$o(\tau) = \sigma\left[W_o x(\tau) + U_o h(\tau - 1) + b_o\right] \tag{6}$$

$$c(\tau) = f(\tau) \odot c(\tau - 1) + i(\tau) \odot c_o(\tau) \tag{7}$$

$$h(\tau) = o(\tau) \odot \varphi[c(\tau)] \tag{8}$$

Where (W_f, W_i, W_o, W_c, U_f, U_i, U_o, U_c) and (b_f, b_i, b_o, b_c) are the weights and biases, respectively, which are independent of time. Moreover, $f(\tau)$, $i(\tau)$ and $o(\tau)$ denote forget, input and output gates, respectively. Whereas $c(\tau)$ and $h(\tau)$ are cell and hidden states, respectively. The symbols: φ, σ and $\odot$ represent 'tanh', sigmoid and element-wise multiplication functions, correspondingly.

3.6. Gated recurrent unit (GRU)

GRU is one of the types of neural networks (NNs) first introduced by Kyunghyun Cho in 2015 [61]. The goal of GRU network is to solve the vanishing gradient problem of the conventional RNN model. Like LSTM, GRU also has gated architecture. It consists of two gates: an update gate and a reset gate. The retention of information is determined by the update gate. The reset gate governs which information is not worthy, and it helps the model to get rid of it. This gated architecture helps the model to overcome the vanishing gradient problem [62]. GRU network is different from the LSTM network in one aspect it does not have memory cells [63]. The following equations summarize the working of GRU [64].

$$h_t = (1 - z_t)\, h_{t-1} + z_t\, h'_t \tag{9}$$

$$z_t = \sigma(W_z\, x_t + U_z\, (h_{t-1})) \tag{10}$$

$$h'_t = \tanh(W_h\, x_t + U(r_t \odot h_{t-1})) \tag{11}$$

$$r_t = \sigma(W_r\, x_t + U_r\, h_{t-1}) \tag{12}$$

Where h_t and $h_t\prime$ denote output and candidate output, respectively. The update and reset gates are denoted by z_t and r_t. W_z, W_r, W_h, U_z and U_r represent metrics in GRU. The element-wise multiplication is represented by $\odot$.

3.7. Performance evaluation

The performances of predictive models can be evaluated by using different error measurement techniques. In the proposed study, four error measurement techniques: RMSE (Wm^{-2}), NRMSE (%), MAE (Wm^{-2}) and MSE (Wm^{-2}) are used for performance evaluation. These error measurement techniques are defined by the following equations [65].

$$RMSE = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(X_I - Y_I)^2} \tag{13}$$

$$NRMSE = \frac{RMSE}{\max(Y_I) - \min(Y_I)} * 100 \tag{14}$$

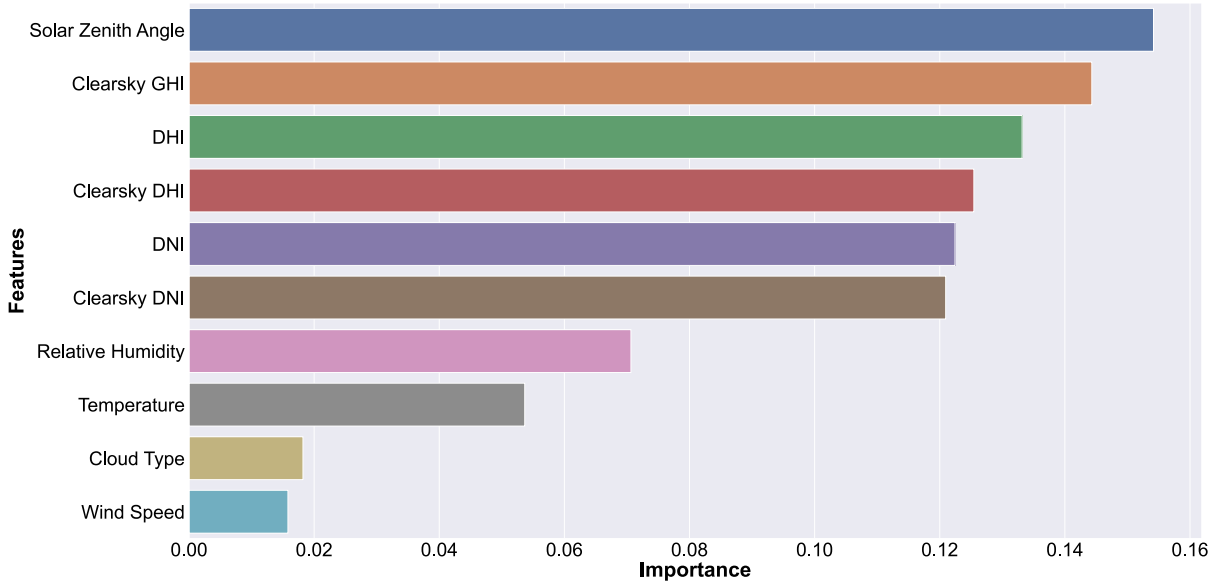
$$MAE = \frac{1}{N}\sum_{I=1}^{N}|X_I - Y_I| \tag{15}$$

$$MSE = \frac{1}{N}\sum_{I=1}^{N}(X_I - Y_I)^2 \tag{16}$$

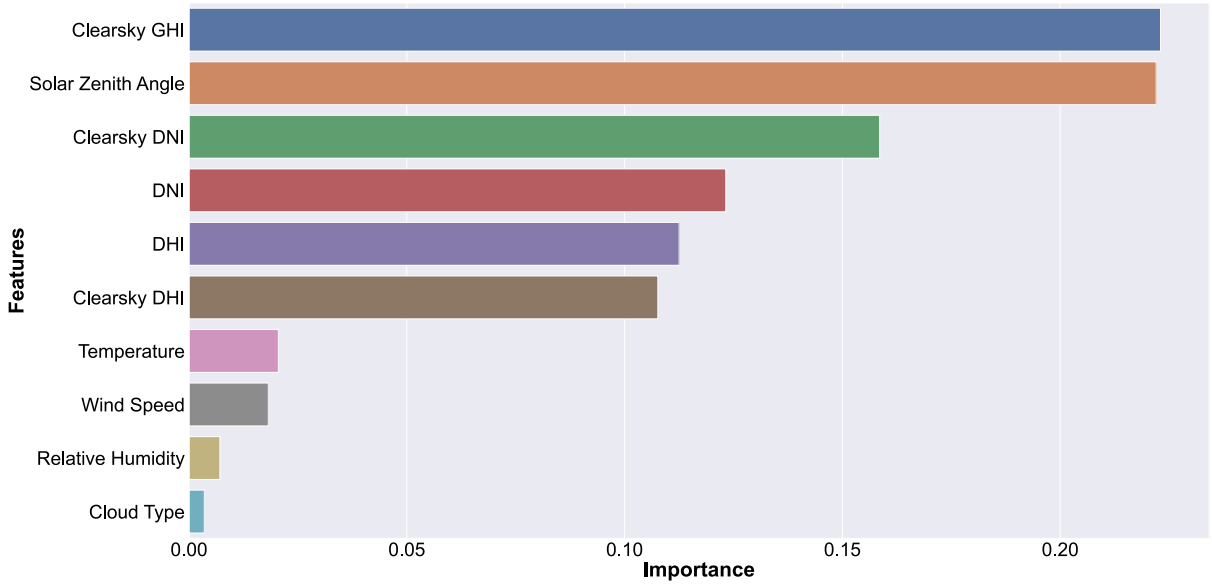
Where X_I, Y_I are measured and predicted values, respectively, and N represents the total number of values.

4. Data processing and model configuration

This section presents the dataset description, the process of feature engineering and hyper-parameter tuning for model training. The details are further discussed in the subsequent section.



(a)



(b)

Fig. 4. Feature Importance identification with strong influence on target variable using RF for the dataset of two cities: (a) Golden and (b) Johannesburg.

Table 1
Dataset description.

Location	Latitude	Longitude	Data size (Years)
Golden, Colorado, USA	39.75700°	−105.22058°	2 (2019–2020)
Johannesburg, SA	−26.195246°	28.034088°	1 (2019)

4.1. Dataset description and division

The datasets used in this study were gathered from two different locations, Golden, Colorado, USA, and Johannesburg, South Africa. The selection of distinct locations was intentional as we aimed to test the robustness and adaptability of our proposed model to diverse climatic conditions and SI patterns. This diversity ensures the model's applicability across varied terrains and weather conditions, avoiding over-reliance on specific geographic or climatic contexts. These two datasets are sourced from the NSRDB and are detailed in Table 1. The data from Golden spans two years, from January 1, 2019, to December 31, 2020, while the Johannesburg dataset covers the entirety of 2019.

These datasets encompass a rich array of meteorological features, including DHI (Wm^{-2}), DNI (Wm^{-2}), clear sky DHI (Wm^{-2}), clear sky DNI (Wm^{-2}), clear sky GHI (Wm^{-2}), temperature (°C), relative humidity (%), pressure (mbr), wind speed (ms^{-1}), wind direction (°), surface albedo, dew point (°C), solar zenith angle (°), perceptible water, and cloud type. Such comprehensive data offers a holistic view of the weather conditions, which is crucial for accurate GHI predictions. In the proposed study, we adopt an 80–20 split, where 80% of the data is used for the training and 20% for the testing.

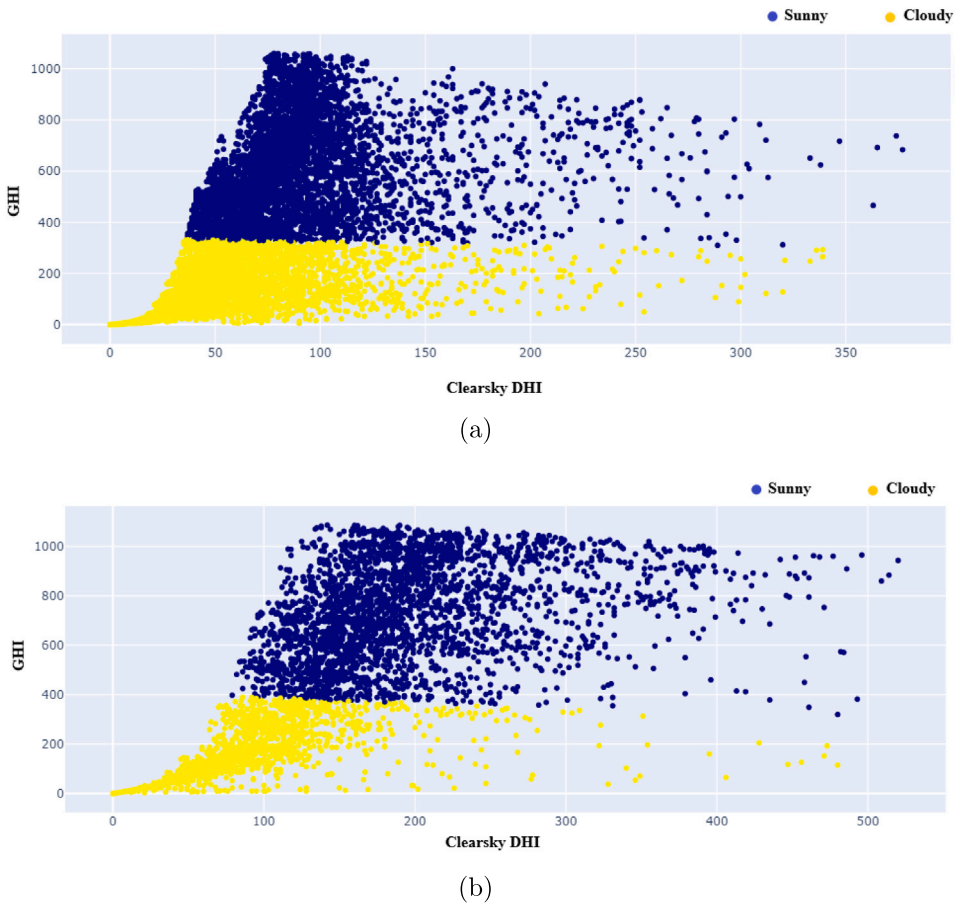


Fig. 5. Weather classification using K-mean clustering algorithm (a) Golden (b) Johannesburg.

Table 2
Search space for RF hyper-parameters.

Hyper-parameters	Search space
N-estimators	{20, 60, 100, 120, 150, 200}
Min-samples-split	{2, 5}
Min-samples-leaf	{1, 2}
Max-samples	{0.5, 0.6, 0.7, 0.75}
Max-features	{0.2, 0.6, 1.0}
Max-depth	{2, 8, None}
Bootstrap	{True, False}

Table 3
Search space for the CatBoost model's hyper-parameters.

Hyperparameters	Search space
Iteration	{200, 300, 400}
Learning rate	{0.1, 0.01, 0.03}
Depth	{2, 4, 6,8}
L2 leaf regularization	{0.2, 0.5, 1, 3}

4.2. Feature selection

The datasets of both locations contain 15 features. In the first stage, RF is used for feature selection, and with extensive analysis, the top 10 features given in Fig. 4 are chosen for model training. Specifically, RF identified the top 10 features that have the maximum influence on hourly GHI forecasting. The results show that parameters like DHI, DNI, clear sky DHI, clear sky DNI, temperature, humidity, solar zenith angle, wind speed, and cloud type emerged as the most important features influencing the performance of the model. The search space for RF hyper-parameters is given in Table 2.

Once feature selection is done, the K-mean clustering is used to categorize the data in cloudy and sunny hours. The results in Fig. 5 show the weather classification of the SI data, which has been categorized into sunny and cloudy hours based on the DHI single parameter clustering. This weather classification is significant because it separates sunny conditions with high clear sky DHI from cloudy conditions with low DHI reducing the variability and uncertainty associated with SI forecasting. The separation of data into distinct clusters allows the model to capture distinct relationships and trends within each weather category, improving the effectiveness of the training process.

4.3. Hyper-parameter tuning

After selecting the best subset of features with weather-categorized data, the next step is training the model. This requires extensive hyper-parameter tuning to ensure a robust ML model. Appropriate tuning of the parameters not only influences the model's predicting accuracy but the computational speed and memory requirements are also affected. Different algorithms are introduced in the literature for tuning parameters. In this proposed study, parameters are tuned for cloudy and sunny hours cluster data individually. For the proposed CatBoost model, the number of iterations, learning rate, depth and L2 leaf regularization are the hyper-parameters that are tuned by grid search. The search space used for the parameters' tuning for the CatBoost model is presented in Table 3.

Three different optimization solvers: adaptive movement estimation (ADAM), stochastic gradient descent (SGD) and root mean squared propagation (RMSprop) are tested for the BiLSTM model. ADAM was found to be the best solver. Moreover, the hyper-parameters optimized for BiLSTM are also used for GRU. The search space used for

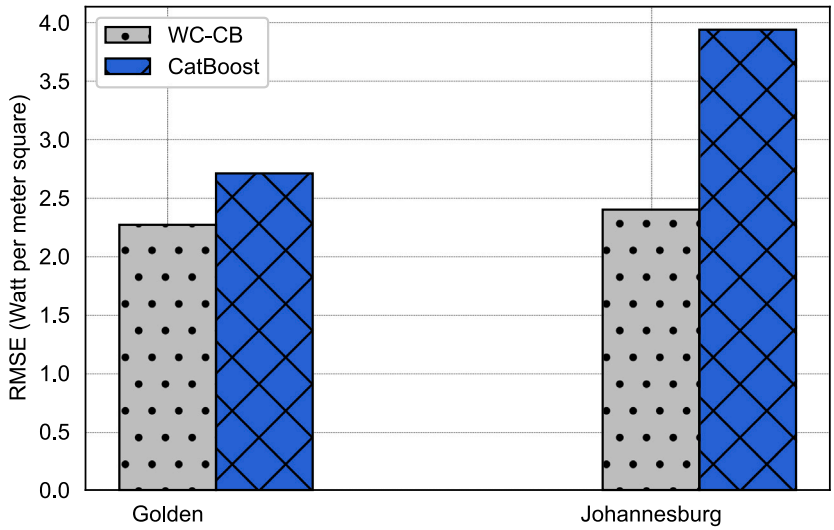


Fig. 6. Bar chart representation of RMSE for hybrid and conventional CatBoost models.

Table 4
Search space for the BiLSTM model's hyper-parameters.

Hyperparameters	Search space
Input shape	3D
No. of hidden layer	{2, 3, 4, 5}
No. of units in each hidden layer	{32, 64, 96, 128}
Learning rate	{0.1, 0.05, 0.001, 0.0001}
Optimizers	{ADAM, SGD, RMSprop}
Batch size	{16, 32}
No. of epochs	{50, 100, 150, 200}

Table 5
Performance evaluation of hybrid and conventional CatBoost models.

Location	Model	RMSE (Wm^{-2})	MSE (Wm^{-2})	MAE (Wm^{-2})	NRMSE (%)
Golden	**WC-CB**	**2.27**	**7.19**	**1.32**	**0.48**
	Conventional	2.71	7.37	1.46	0.38
Johannesburg	**WC-CB**	**2.4**	**6.96**	**1.5**	**0.49**
	Conventional	3.94	15.57	1.61	0.46

hyper-parameters tuning of BiLSTM model is given in Table 4. The AdaBoost parameters: the maximum number of estimators and learning rate are also optimized for all types of weather-classified data. The learning rate of 0.1 and 100 estimators are found to be the best for all weather-classified data by the grid search algorithm.

5. Results and discussion

This section explains the results obtained from our WC-CB model, emphasizing their implications and broader significance. The results of the proposed WC-CB are compared with the established models, BiLSTM, GRU, and AdaBoost, specifically for hour-ahead GHI forecasting. Our analysis leverages two distinct real-world datasets to ensure a robust evaluation of the proposed model. Furthermore, a thorough comparative analysis is presented, including graphical representations that compare the forecasted and measured GHI. Moreover, detailed insights into the model's performance under various clustering parameters are also discussed. In addition, a comparison of the proposed WC-CB model with traditional techniques is given in the literature to place our findings in the broader research context.

5.1. Hybrid vs. conventional CatBoosting

Initially, the proposed WC-CB approach is compared with the conventional CatBoost model, which is trained on non-weather classified data. In Table 5, the results of the comparative analysis of hybrid and conventional CatBoost models are presented. The improvement of 16.23% and 9.5% on the dataset of Golden City, while 39.08% and 6.8% for Johannesburg is achieved in RMSE and MAE, respectively, compared to the conventional model. Tailoring the modelling to distinct sunny and cloudy conditions through weather-based clustering can significantly reduce forecast errors. This is because creating data subsets corresponding to different irradiance patterns enables more specialized learning of the unique relationships and trends. The WC-CB approach, therefore, provides improved modelling and higher accuracy than global modelling on the full dataset without weather context. In Fig. 6, the RMSE of WC-CB and conventional CatBoost models is presented as a bar chart. Like the hybrid approach, the large dataset size also improves the performance of the traditional CatBoost model, as all four performance indicators give better results on the dataset of Golden than Johannesburg city.

5.2. Single parameter clustering

The data is categorized based solely on the clear sky DHI parameter in single-parameter clustering. Using the K-means clustering algorithm, the classification outcomes for two distinct locations, Golden, Colorado, and Johannesburg, South Africa, are presented in Fig. 5. The data is divided into two clusters systematically, representing sunny and cloudy conditions. This strategic classification facilitates the identification of distinct irradiance patterns corresponding to varying weather conditions, helping the ML model to learn specific relationships inherent to each weather type. This nuanced approach significantly improves the forecasting accuracy of the models. We employ four metrics to evaluate the model's performance: RMSE, MSE, MAE, and NRMSE. The corresponding results are presented in Table 6.

Upon analysis, it is evident that the proposed WC-CB model performs much better when compared to other models, namely AdaBoost, BiLSTM, and GRU, across both datasets. Specifically, for the Golden city dataset, the WC-CB model achieved average RMSE, MSE, MAE, and NRMSE values as 2.27, 7.19, 1.32, and 0.48, respectively. In contrast, the recorded average values for the Johannesburg dataset were 2.40, 6.96, 1.5, and 0.49, respectively. We made an interesting observation during our study regarding the impact of dataset size on the accuracy of predictions. Like traditional ML models, the prediction accuracy of the WC-CB model improves when trained on larger datasets. This is

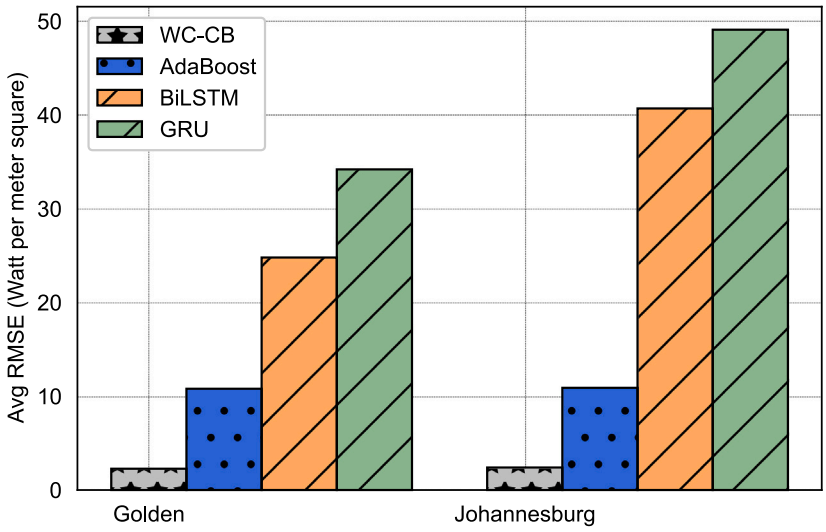


Fig. 7. The average RMSE of all four models in the form of a bar chart. It is clearly evident from the results that the proposed WC-CB performs better compared to other techniques.

Table 6
Proposed hybrid and other models obtained results.

Location	Model	RMSE (Wm^{-2})			MSE (Wm^{-2})			MAE (Wm^{-2})			NRMSE (%)		
		Cloudy	Sunny	Avg.	Cloudy	Sunny	Avg.	Cloudy	Sunny	Avg.	Cloudy	Sunny	Avg.
Golden	**WC-CB**	**0.82**	**3.71**	**2.27**	**0.67**	**13.72**	**7.19**	**0.32**	**2.36**	**1.32**	**0.24**	**0.73**	**0.48**
	AdaBoost	4.08	17.47	10.8	16.65	305.46	161.1	2.93	15.58	9.25	1.33	3.9	2.62
	BiLSTM	19.4	40.29	24.8	376.42	1623.4	999.9	8.09	27.86	17.9	6.44	8.09	7.27
	GRU	24.33	44.12	34.2	592.05	1946.7	1269.4	13.26	31.37	22.3	9.32	8.13	8.73
Johannesburg	**WC-CB**	**1.32**	**3.49**	**2.4**	**1.73**	**12.19**	**6.96**	**0.52**	**2.41**	**1.5**	**0.34**	**0.64**	**0.49**
	AdaBoost	3.45	18.47	10.9	11.89	341.18	176.5	2.17	16.25	9.21	0.95	3.39	2.17
	BiLSTM	24.19	57.25	40.7	585.49	3277.7	1931.6	11.41	39.76	25.6	7.35	10.64	8.99
	GRU	30.63	67.62	49.1	940.25	4573.4	2756.8	12.35	51.27	31.8	10.5	13.89	12.2

Table 7
Inference time of models.

Location	Model	Inference time (Seconds)
Golden	**WC-CB**	**0.064**
	AdaBoost	0.111
	BiLSTM	2.62
	GRU	1.6

evident from the superior forecasting results achieved with the Golden City dataset compared to the Johannesburg dataset.

The results in Fig. 7, depict the bar chart of average RMSE. The results show that the WC-CB model consistently outperforms the other models in terms of RMSE for both locations. The GRU model lagged in forecasting accuracy compared to other techniques. Furthermore, the WC-CB model demonstrated better forecasting results during cloudy than sunny hours. Another crucial aspect of model performance is inference time, tabulated in Table 7. The WC-CB model boasts the best inference time, closely followed by AdaBoost. In contrast, the BiLSTM model exhibited the longest inference time among the models studied.

To show the effectiveness of the WC-CB approach, a fitted line plot of measured GHI against the output of the hybrid model is depicted in Fig. 8. A fitted line plot is a scatter plot displaying points against the regressor line. The model's predictive accuracy improves as the points get closer to the regressor line and vice versa. Thus, Fig. 9 demonstrates that the proposed WC-CB model's forecasting accuracy is high for both types of clustered data as the predicted result of the proposed approach is close to the fitted line.

5.3. Two parameter clustering

This section delves into the outcomes when employing a two-dimensional clustering approach, utilizing two variables for GHI data classification. The Golden City dataset is used to evaluate how two parameters affect model classification performance. Fig. 10 depicts the outcome of classification using two parameters. The results of all four models on classified data are presented in Table 8. The result demonstrates the superiority of the proposed approach over other models. The average RMSE 2.75, 11.3, 34.5 and 42.3 Wm^{-2} is recorded for WC-CB, AdaBoost, BiLSTM and GRU, respectively. With two parameters-based clustering techniques, the errors are further reduced for cloudy cluster data, while the performances of models deteriorate for the cluster of sunny hours. This is because more data points are now located in a cloudy cluster, and the size of the sunny cluster dataset has decreased. However, the overall performances of the models are not improved from the former classification approach. Therefore, the classification of data with one parameter is suited for prediction tasks because of better predicting accuracy than a classification approach based on two parameters.

The results in Table 9 thoroughly assess the performance of the ML models, WC-CB, AdaBoost, BiLSTM, and GRU, across two locations, Golden and Johannesburg, under different weather conditions, such as cloudy and sunny. The performance metric used in the evaluation is RMSE, a widely accepted measure for evaluating the performance of a predictive model by measuring the deviation from the actual values. To ensure the reliability of the assessment, the study uses bootstrap resampling, a well-known method for estimating the distribution of a statistic (in this case, RMSE) by random sampling with replacement from the test dataset. This method helps calculate the confidence

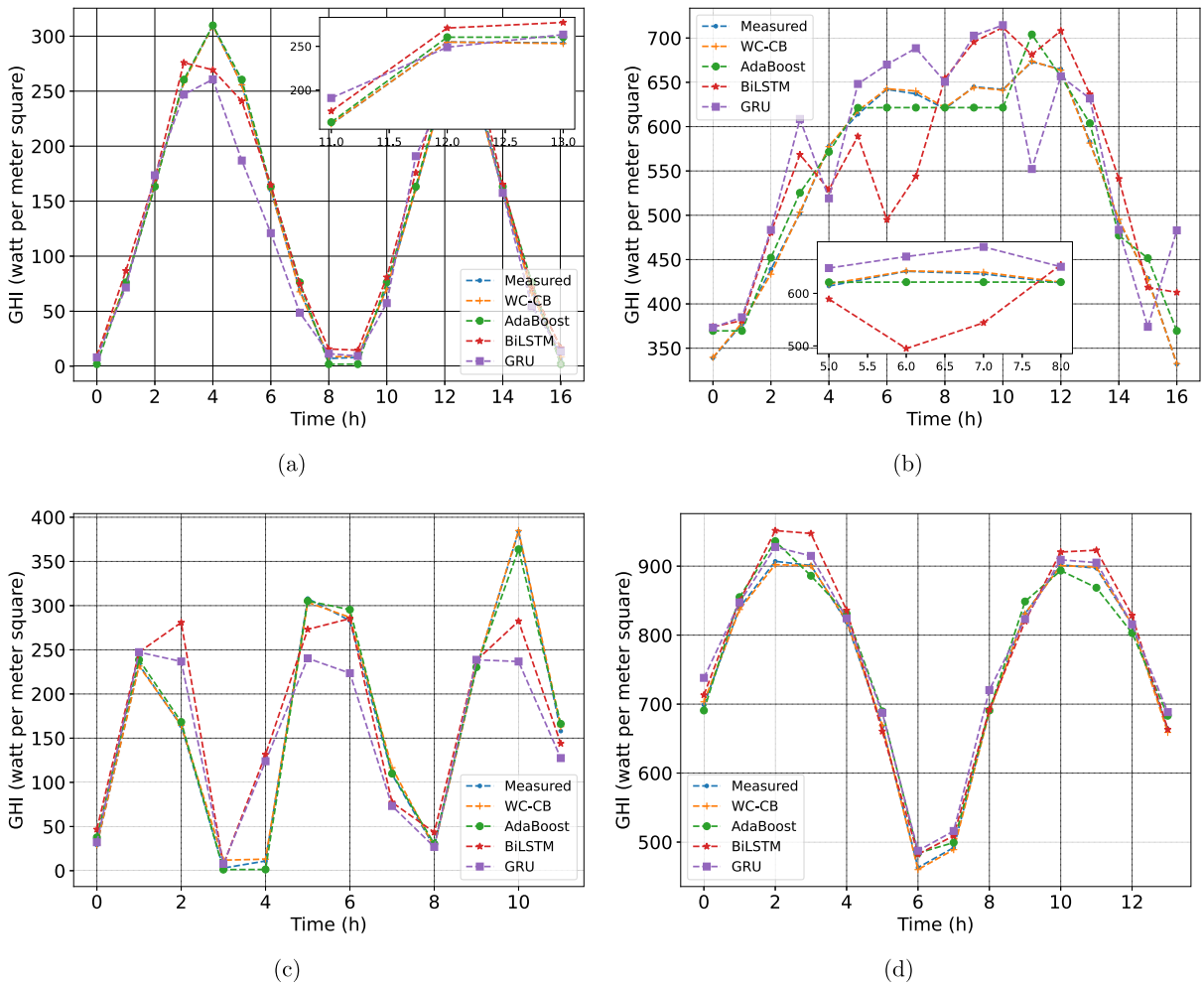


Fig. 8. Graphical representation of measured and predictive GHI. (a) Golden cloudy hours. (b) Golden sunny hours. (c) Johannesburg cloudy hours. (d) Johannesburg sunny hours.

Table 8
Results of the proposed and other models with two parameters clustering.

Location	Model	RMSE (Wm^{-2})			MSE (Wm^{-2})			MAE (Wm^{-2})			NRMSE (%)		
		Cloudy	Sunny	Avg.	Cloudy	Sunny	Avg.	Cloudy	Sunny	Avg.	Cloudy	Sunny	Avg.
Golden	**WC-CB**	**0.69**	**4.78**	**2.75**	**0.49**	**22.93**	**11.7**	**0.31**	**2.51**	**1.41**	**0.21**	**0.93**	**0.57**
	AdaBoost	4.06	18.59	11.3	16.52	345.61	181.1	3	16.64	9.82	1.32	4.19	2.75
	BiLSTM	18.63	50.32	34.5	347.18	2531.9	1439.5	7.16	36.15	21.7	6.1	10.75	8.43
	GRU	38.08	46.5	42.3	1449.9	2162.3	1806.1	21.84	32.37	27.1	5.25	9.82	7.53

intervals for each model's RMSE, providing insights into the precision and reliability of predictions.

We present a 95% confidence interval for each model, and the range indicates where the actual RMSE value is likely to be found with 95% probability. We derive this confidence interval from the distribution of RMSE values obtained from numerous bootstrap samples from our test set. The narrow confidence interval (0.69, 0.99) of the WC-CB model for the Golden City location indicates a relatively high level of precision in cloudy conditions. The AdaBoost, BiLSTM, and GRU models follow a similar trend, with generally wider intervals under sunny conditions than cloudy ones. This suggests that the model's performance varies with weather conditions. Furthermore, our results suggest that the WC-CB model consistently outperforms the other models in terms of RMSE across different weather conditions and locations. We use bootstrap resampling for confidence interval estimation, which provides a comprehensive understanding of the models' performance variability. This reinforces the reliability of our comparative analysis.

5.4. Literature comparison

In this section, we compare the performance of the proposed WC-CB model with the techniques reported in the literature for hour-ahead GHI forecasting. In [11,41], the datasets of New York and Tripoli cities are used, collected from NSRDB. We collected these two cities' datasets and applied them to the proposed model. In Table 10, a comparison of the proposed technique with the models presented in [11,41] is presented.

The comparative analysis demonstrates the superiority of the proposed WC-CB model over other approaches presented in the literature for predicting hour-ahead GHI. The effectiveness of the proposed model is due to the boosting phenomena of CatBoost and the clustering strategy, which reduce uncertainties. The datasets used in this study also contain exogenous variables. The model's forecasting accuracy also depends on the accurate measurement of exogenous variables.

In light of our comprehensive analysis, the results underscore the efficacy and robustness of the proposed WC-CB model, particularly

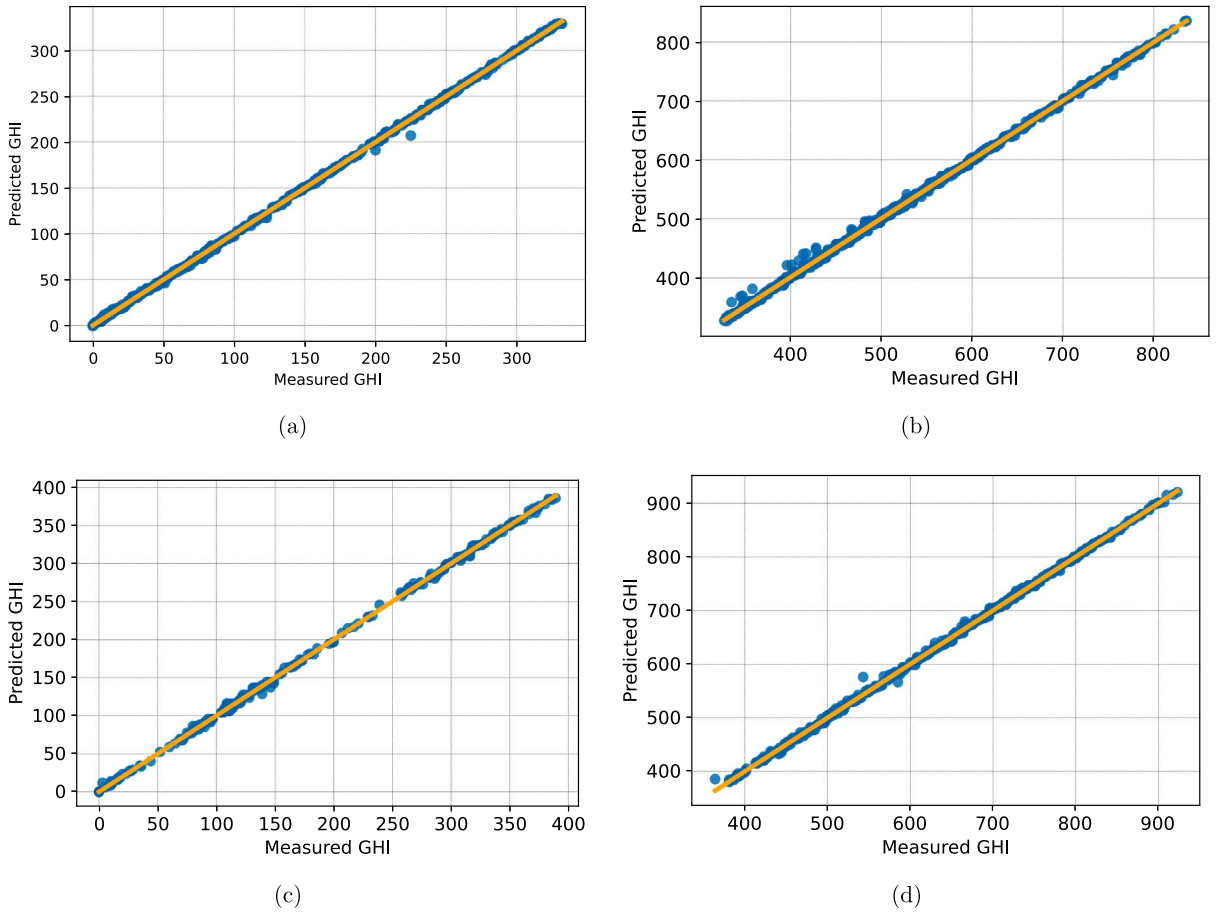


Fig. 9. Graphical representation of measured and predictive GHI. (a) Golden cloudy hours. (b) Golden sunny hours. (c) Johannesburg, cloudy hours. (d) Johannesburg sunny hours.

Table 9
Comparative Bootstrap confidence intervals for RMSE for comparison under different weather conditions in Golden and Johannesburg cities. The narrower confidence interval indicates higher precision..

Location	Cluster	Model	Confidence interval	RMSE (W/m^2)
Golden	Cloudy	WC-CB	[0.69, 0.97]	0.82 ± 0.15
		AdaBoost	[3.96, 4.37]	4.08 ± 0.29
		BiLSTM	[17.52, 21.15]	19.40 ± 1.75
		GRU	[22.53, 26.29]	24.33 ± 1.96
	Sunny	WC-CB	[3.23, 4.19]	3.71 ± 0.48
		AdaBoost	[16.98, 18.01]	17.47 ± 0.54
		BiLSTM	[37.32, 43.13]	40.29 ± 2.84
		GRU	[41.11, 46.88]	44.12 ± 3.01
Johannesburg	Cloudy	WC-CB	[1.16, 1.47]	1.32 ± 0.15
		AdaBoost	[3.10, 3.78]	3.45 ± 0.33
		BiLSTM	[20.92, 27.47]	24.19 ± 3.28
		GRU	[27.44, 33.83]	30.63 ± 3.2
	Sunny	WC-CB	[2.94, 4.18]	3.49 ± 0.69
		AdaBoost	[17.83, 19.08]	18.47 ± 0.61
		BiLSTM	[50.65, 63.58]	57.25 ± 6.33
		GRU	[61.11, 74.46]	67.62 ± 6.84

for hour-ahead GHI forecasting. The model's performance is noteworthy, especially compared to other established models such as BiLSTM, GRU, and AdaBoost. The single-parameter clustering, which categorizes data based on the clear sky DHI parameter, has enhanced the model's forecasting performance. This strategic classification facilitates the identification of distinct irradiance patterns corresponding to varying weather conditions, enabling the ML model to capture specific relationships inherent to each weather type. Furthermore, the reduced inference time of the proposed WC-CB, especially compared to other models like BiLSTM, underscores its efficiency and potential for real-time applications. It is important to highlight the limitations despite the promising results of the WC-CB model presented in the study. The approach of clustering with a single parameter is effective but may not capture all the complexities of different weather patterns. Hence, using more sophisticated clustering techniques could lead to improved accuracy, which needs a through investigation. There are also concerns about the model's generalization, particularly concerning long-term weather patterns or datasets with varying temporal resolutions. Additionally, the hybrid nature of the WC-CB model adds complexity, which could pose challenges when interpreting and troubleshooting the results.

6. Conclusions and future work

In this paper, we proposed a hybrid WC-CB model for hourly SI forecasting using historical data. Our approach has a unique strength as it strategically integrates multiple methodologies. For instance, RF captures feature importance, effectively addressing the inherent variability in solar data to avoid overfitting. Additionally, K-means clustering is used for data segmentation into distant weather conditions (sunny/cloudy), which ensures that the model is trained on more homogeneous data subsets. This reduces the potential for large errors due to sudden weather shifts. Weather-specific data partitioning is a key reason for the model's low errors. Finally, CatBoost, with power handling of categorical data and iterative GB mechanism, refines predictions by learning from previous errors, ensuring accuracy and computational efficiency. This ensemble design outperforms BiLSTM, GRU and AdaBoost thanks to its adaptation to specific domain intricacies and prior feature selection and clustering steps, resulting in superior forecasting performance and faster inference.

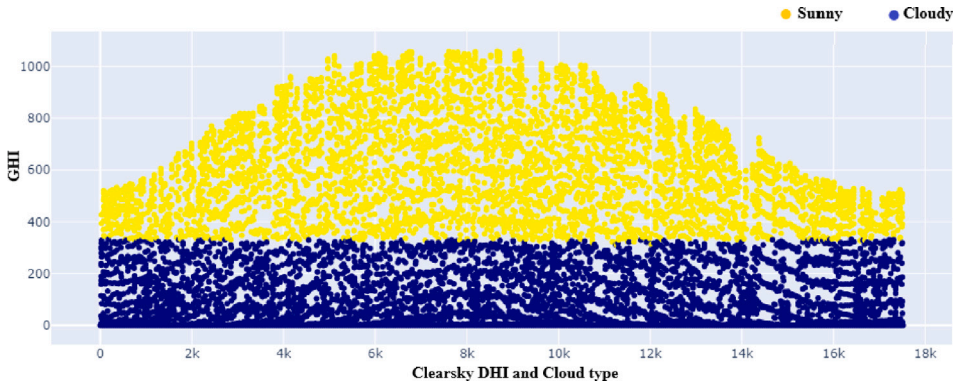


Fig. 10. Clustering using two parameters.

Table 10
Comparison of the proposed model with other approaches.

Ref	Journal publisher	Location	Mode	RMSE (Wm^{-2})	MAE (Wm^{-2})
[11]	IEEE Access	New York, USA	K-mean-LSTM	41.37	30.19
[41]	IEEE Transaction on Industrial informatics	Tripoli, Libya, North Africa	SCFT-LSTM	13.48	10.05
Proposed model		New York, USA	**WC-CB**	**2.16**	**1.37**
Proposed model		Tripoli, Libya, North Africa	**WC-CB**	**2**	**1.36**

The extensive analysis on two distinct real-world datasets from geographically diverse locations demonstrates the robustness of WC-CB. The approach not only outperforms the conventional CatBoost model trained without clustering but also underscores the scientific value added by this research. Additionally, comprehensive benchmarking also proves WC-CB's superiority over state-of-the-art techniques like BiLSTM, GRU and AdaBoost in metrics of RMSE, MAE, and inference time. In the context of performance comparison, the WC-CB model achieves the overall improvement of 16.23% and 9.5% on the dataset of Golden city while 39.08% and 6.8% for Johannesburg's dataset in terms of RMSE and MAE, respectively, as compared to the conventional model. The improved SI forecasting can facilitate better renewable energy integration, load scheduling, and grid management. Applications such as unit commitment and economic dispatch can leverage improved predictions to create more efficient and sustainable energy systems.

On the scientific front, this research has contributed to the unique integration, not commonly seen in existing literature, combining RF, K-means, and CatBoost, tailored specifically for SI forecasting. The workflow, which combines selective feature extraction, weather categorization, and tuned boosting, provides a holistic and impactful solution, setting a benchmark for future research in this domain. However, it is essential to acknowledge potential limitations and areas for future research. Evaluating performance under diverse weather scenarios and investigating optimal classification techniques can help achieve robustness across broader deployments. Incorporating the forecasting model into practical power system operations will validate its real-world efficacy. Furthermore, hyper-parameter tuning and algorithmic refinements tailored to solar data nuances can potentially further improve accuracy, offering a promising direction for subsequent studies. While WC-CB has showcased significant promise, its adaptability to regions with extreme weather fluctuations or its scalability to larger datasets remains to be explored.

CRediT authorship contribution statement

Ubaid Ahmed: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Ahsan Raza Khan:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Anzar Mahmood:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Iqra Rafiq:** Writing – review & editing, Investigation, Formal analysis. **Rami Ghannam:** Writing – review & editing, Conceptualization. **Ahmed Zoha:** Writing – review & editing, Supervision, Project administration, Conceptualization.

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.

Data availability

This study used the publicly available dataset.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to improve the language, grammar, and connectivity of the manuscript. After using ChatGPT, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

References

[1] J. Sheffield, World population growth and the role of annual energy use per capita, Technol. Forecast. Soc. Change 59 (1) (1998) 55–87.
[2] Key world energy statistics 2021, IEA, Paris. URL https://www.iea.org/reports/key-world-energy-statistics-2021. (Accessed 8 March 2022).
[3] S. Miao, G. Ning, Y. Gu, J. Yan, B. Ma, Markov chain model for solar farm generation and its application to generation performance evaluation, J. Clean. Prod. 186 (2018) 905–917.
[4] I.N. Jiya, R. Gouws, Overview of power electronic switches: A summary of the past, state-of-the-art and illumination of the future, Micromachines 11 (12) (2020) 1116.
[5] D. Gielen, R. Gorini, N. Wagner, R. Leme, L. Gutierrez, G. Prakash, E. Asmelash, L. Janeiro, G. Gallina, G. Vale, et al., Global Energy Transformation: A Roadmap to 2050, Institution of Gas Engineers and Managers (IGEM), 2019.
[6] A.R. Khan, A. Mahmood, A. Safdar, Z.A. Khan, N.A. Khan, Load forecasting, dynamic pricing and DSM in smart grid: A review, Renew. Sustain. Energy Rev. 54 (2016) 1311–1322.

[7] H. Malik, N. Fatema, A. Iqbal, Intelligent Data-Analytics for Condition Monitoring: Smart Grid Applications, Academic Press, 2021.

[8] M.F. Anjos, A.J. Conejo, et al., Unit commitment in electric energy systems, Found. Trends® Electr. Energy Syst. 1 (4) (2017) 220–310.

[9] N. Singh, Y. Kumar, Multiobjective economic load dispatch problem solved by new PSO, Adv. Electr. Eng. 2015 (2015).

[10] C.N. Obiora, A. Ali, A.N. Hasan, Forecasting hourly solar irradiance using long short-term memory (LSTM) network, in: 2020 11th International Renewable Energy Congress, IREC, Hammamet, Tunisia, IEEE, 2020, pp. 1–6.

[11] Y. Yu, J. Cao, J. Zhu, An LSTM short-term solar irradiance forecasting under complicated weather conditions, IEEE Access 7 (2019) 145651–145666.

[12] Y. Lin, D. Duan, X. Hong, X. Cheng, L. Yang, S. Cui, Very-short-term solar forecasting with long short-term memory (LSTM) network, in: 2020 Asia Energy and Electrical Engineering Symposium, AEEES, Chengdu, China, IEEE, 2020, pp. 963–967.

[13] F. Serttas, F.O. Hocaoglu, E. Akarslan, Short term solar power generation forecasting: A novel approach, in: 2018 International Conference on Photovoltaic Science and Technologies, PVCon, Ankara, Turkey, IEEE, 2018, pp. 1–4.

[14] R. Zafar, B.H. Vu, M. Husein, I.-Y. Chung, Day-ahead solar irradiance forecasting using hybrid recurrent neural network with weather classification for power system scheduling, Appl. Sci. 11 (15) (2021) 6738.

[15] H. Tyler, Why is the weather so hard to predict?, Let's Talk Science. URL https://letstalkscience.ca/educational-resources/stem-in-context/why-weather-so-hard-predict. (Accessed 6 November 2022).

[16] B. Singh, D. Pozo, A guide to solar power forecasting using ARMA models, in: 2019 IEEE PES Innovative Smart Grid Technologies Europe, ISGT-Europe, Bucharest, Romania, IEEE, 2019, pp. 1–4.

[17] D. Yang, P. Jirutitijaroen, W.M. Walsh, Hourly solar irradiance time series forecasting using cloud cover index, Sol. Energy 86 (12) (2012) 3531–3543.

[18] S. Hussain, A. Al Alili, Day ahead hourly forecast of solar irradiance for abu dhabi, UAE, in: 2016 IEEE Smart Energy Grid Engineering, SEGE, Oshawa, ON, Canada, IEEE, 2016, pp. 68–71.

[19] S. Garg, A. Agrawal, S. Goyal, K. Verma, Day ahead solar irradiance forecasting using Markov chain model, in: 2020 IEEE 17th India Council International Conference, INDICON, New Delhi, India, IEEE, 2020, pp. 1–5.

[20] S. Garg, A. Agrawal, S. Goyal, K. Verma, Day ahead solar irradiance forecasting using different statistical techniques, in: 2020 IEEE International Conference on Power Electronics, Drives and Energy Systems, PEDES, Jaipur, India, IEEE, 2020, pp. 1–4.

[21] J. Boland, M. David, P. Lauret, Short term solar radiation forecasting: Island versus continental sites, Energy 113 (2016) 186–192.

[22] C. Harsito, T. Triyono, E. Rovianto, Analysis of heat potential in solar panels for thermoelectric generators using ANSYS software, Civ. Eng. J. 8 (7) (2022) 1328–1338.

[23] W. Musa, V. Ponkratov, A. Karaev, N. Kuznetsov, L. Vatutina, M. Volkova, O. Shalina, A. Masterov, Multi-cycle production development planning for sustainable power systems to maximize the use of renewable energy sources, Civ. Eng. J. 8 (11) (2022) 2628–2639.

[24] A.W. Aryaputera, D. Yang, W.M. Walsh, Day-ahead solar irradiance forecasting in a tropical environment, J. Sol. Energy Eng. 137 (5) (2015).

[25] P. Bendiek, A. Taha, Q.H. Abbasi, B. Barakat, Solar irradiance forecasting using a data-driven algorithm and contextual optimisation, Appl. Sci. 12 (1) (2021) 134.

[26] C. Paoli, C. Voyant, M. Muselli, M.-L. Nivet, Forecasting of preprocessed daily solar radiation time series using neural networks, Sol. Energy 84 (12) (2010) 2146–2160.

[27] T. Mutavhatsindi, C. Sigauke, R. Mbuvha, Forecasting hourly global horizontal solar irradiance in South Africa using machine learning models, IEEE Access 8 (2020) 198872–198885.

[28] R. Marquez, C.F. Coimbra, Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database, Sol. Energy 85 (5) (2011) 746–756.

[29] A. Alzahrani, P. Shamsi, M. Ferdowsi, C. Dagli, Solar irradiance forecasting using deep recurrent neural networks, in: 2017 IEEE 6th International Conference on Renewable Energy Research and Applications, ICRERA, San Diego, California, USA, Ieee, 2017, pp. 988–994.

[30] N. Sharma, R. Sharma, N. Jindal, Machine learning and deep learning applications-a vision, Glob. Transit. Proc. 2 (1) (2021) 24–28.

[31] M.S. Hossain, H. Mahmood, Short-term photovoltaic power forecasting using an LSTM neural network and synthetic weather forecast, IEEE Access 8 (2020) 172524–172533.

[32] X. Qing, Y. Niu, Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM, Energy 148 (2018) 461–468.

[33] Z. Zixuan, Boosting algorithms explained, Medium. URL https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30. (Accessed 8 November 2022).

[34] M. Kamble, S. Ghosh, P. Patel, Solar irradiance prediction using meteorological data by ensemble models, in: 2nd International Conference on Data, Engineering and Applications, IDEA, Bhopal, India, IEEE, 2020, pp. 1–6.

[35] S. Tiwari, R. Sabzehgar, M. Rasouli, Short term solar irradiance forecast using numerical weather prediction (NWP) with gradient boost regression, in: 2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems, PEDG, Charlotte, NC, USA, IEEE, 2018, pp. 1–8.

[36] L. Benali, G. Notton, A. Fouilloy, C. Voyant, R. Dizene, Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components, Renew. Energy 132 (2019) 871–884.

[37] P. Kumari, D. Toshniwal, Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance, J. Clean. Prod. 279 (2021) 123285.

[38] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, Y. Xiang, Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China, Energy Convers. Manage. 164 (2018) 102–111.

[39] X. Hou, K. Papachristopoulou, Y.-M. Saint-Drenan, S. Kazadzis, Solar radiation nowcasting using a Markov chain multi-model approach, Energies 15 (9) (2022) 2996.

[40] V.A. Tikkiwal, S.V. Singh, H.O. Gupta, Day-ahead forecasting of solar irradiance using hybrid improved cuckoo search-lstm approach, in: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN, Greater Noida, India, IEEE, 2020, pp. 84–88.

[41] N. Omar, H. Aly, T. Little, Seasonal clustering forecasting technique for intelligent hourly solar irradiance systems, IEEE Trans. Ind. Inform. (2022).

[42] M. Abdel-Nasser, K. Mahmoud, M. Lehtonen, Reliable solar irradiance forecasting approach based on choquet integral and deep LSTMs, IEEE Trans. Ind. Inform. 17 (3) (2020) 1873–1881.

[43] H. Zang, L. Liu, L. Sun, L. Cheng, Z. Wei, G. Sun, Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations, Renew. Energy 160 (2020) 26–41.

[44] NSRDB: National solar radiation database, NREL Transforming Energy. URL https://nsrdb.nrel.gov/data-viewer. (Accessed 22 October 2022).

[45] Normalize data component, Azure, Microsoft. URL https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/normalize-datar. (Accessed 10 November 2022).

[46] M.B. Kursa, W.R. Rudnicki, Feature selection with the boruta package, J. Stat. Softw. 36 (2010) 1–13.

[47] M.K. Boutahir, Y. Farhaoui, M. Azrour, I. Zeroual, A. El Allaoui, Effect of feature selection on the prediction of direct normal irradiance, Big Data Min. Anal. 5 (4) (2022) 309–317.

[48] A. Ahmad, L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data, Data Knowl. Eng. 63 (2) (2007) 503–527.

[49] P. Vora, B. Oza, et al., A survey on k-mean clustering and particle swarm optimization, Int. J. Sci. Mod. Eng. 1 (3) (2013) 24–26.

[50] M.A. Syakur, B.K. Khotimah, E.M.S. Rochman, B.D. Satoto, Integration k-means clustering method and elbow method for identification of the best customer profile cluster, in: IOP Conference Series: Materials Science and Engineering, Yekaterinburg, Russia, vol. 336, IOP Publishing, 2018, 012017.

[51] CatBoost, Yandex. URL https://catboost.ai/r. (Accessed 10 November 2022).

[52] S. Hussain, M.W. Mustafa, T.A. Jumani, S.K. Baloch, H. Alotaibi, I. Khan, A. Khan, A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection, Energy Rep. 7 (2021) 4425–4436.

[53] G. Huang, L. Wu, X. Ma, W. Zhang, J. Fan, X. Yu, W. Zeng, H. Zhou, Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions, J. Hydrol. 574 (2019) 1029–1041.

[54] T. Simon, CatBoost regression in 6 minutes, Medium. URL https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329. (Accessed 10 November 2022).

[55] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: Unbiased boosting with categorical features, Adv. Neural Inf. Process. Syst. 31 (2018).

[56] R.E. Schapire, Explaining AdaBoost, in: Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, Springer, 2013, pp. 37–52.

[57] F. Wang, Z. Li, F. He, R. Wang, W. Yu, F. Nie, Feature learning viewpoint of AdaBoost and a new algorithm, IEEE Access 7 (2019) 149890–149899.

[58] H. Widiputra, GA-optimized multivariate CNN-LSTM model for predicting multi-channel mobility in the COVID-19 pandemic, Emerg. Sci. J. 5 (5) (2021) 619–635.

[59] H. Cheng, Z. Xie, L. Wu, Z. Yu, R. Li, Data prediction model in wireless sensor networks based on bidirectional LSTM, EURASIP J. Wireless Commun. Networking 2019 (1) (2019) 1–12.

[60] L. Rahman, N. Mohammed, A.K. Al Azad, A new LSTM model by introducing biological cell state, in: 2016 3rd International Conference on Electrical Engineering and Information Communication Technology, ICEEICT, Dhaka, Bangladesh, IEEE, 2016, pp. 1–6.

[61] Y. Gao, D. Glowacka, Deep gate recurrent neural network, in: Asian Conference on Machine Learning, Bangkok, Thailand, PMLR, 2016, pp. 350–365.
[62] K. Simeon, Understanding GRU networks, Medium. URL https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be. (Accessed 11 November 2022).
[63] K. Yao, T. Cohn, K. Vylomova, K. Duh, C. Dyer, Depth-gated recurrent neural networks, 9, 2015, p. 98, arXiv preprint arXiv:1508.03790.
[64] R.A. Rajagukguk, R.A. Ramadhan, H.-J. Lee, A review on deep learning models for forecasting time series data of solar irradiance and photovoltaic power, Energies 13 (24) (2020) 6623.
[65] Z. Wang, T. Zhang, Y. Shao, B. Ding, LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement, Appl. Acoust. 172 (2021) 107647.