Research article

Attention-Based Models for Multivariate Time Series Forecasting: Multi-step Solar Irradiation Prediction

Sadman Sakib, Mahin K. Mahadi, Samiur R. Abir, Al-Muzadded Moon, Ahmad Shafiullah [*], Sanjida Ali, Fahim Faisal, Mirza M. Nishat

Department of Electrical and Electronic Engineering, Islamic University of Technology, Gazipur, 1704, Bangladesh

ARTICLE INFO

ABSTRACT

Bangladesh's subtropical climate with an abundance of sunlight throughout the greater portion of the year results in increased effectiveness of solar panels. Solar irradiance forecasting is an essential aspect of grid-connected photovoltaic systems to efficiently manage solar power's variation and uncertainty and to assist in balancing power supply and demand. This is why it is essential to forecast solar irradiation accurately. Many meteorological factors influence solar irradiation, which has a high degree of fluctuation and uncertainty. Predicting solar irradiance multiple steps ahead makes it difficult for forecasting models to capture long-term sequential relationships. Attention-based models are widely used in the field of Natural Language Processing for their ability to learn long-term dependencies within sequential data. In this paper, our aim is to present an attention-based model framework for multivariate time series forecasting. Using data from two different locations in Bangladesh with a resolution of 30 min, the Attention-based encoder-decoder, Transformer, and Temporal Fusion Transformer (TFT) models are trained and tested to predict over 24 steps ahead and compared with other forecasting models. According to our findings, adding the attention mechanism significantly increased prediction accuracy and TFT has shown to be more precise than the rest of the algorithms in terms of accuracy and robustness. The obtained mean square error (MSE), the mean absolute error (MAE), and the coefficient of determination (R^2) values for TFT are 0.151, 0.212, and 0.815, respectively. In comparison to the benchmark and sequential models (including the Naive, MLP, and Encoder-Decoder models), TFT has a reduction in the MSE and MAE of 8.4–47.9% and 6.1–22.3%, respectively, while R^2 is raised by 2.13–26.16%. The ability to incorporate long-distance dependency increases the predictive power of attention models.

1. Introduction

The combustion of fossil fuels for conventional electrical systems releases greenhouse gases that significantly contribute to global warming. Extensive efforts have been made to understand and promote renewable energy to reduce reliance on nonrenewable sources [1,2]. The photovoltaic system has emerged as a viable alternative to conventional electricity, offering green energy and a reduced carbon footprint [3]. As awareness grows regarding the financial and ecological benefits of transitioning to renewable energy sources, there has been a notable increase in the adoption of photovoltaic systems in households and small businesses [4]. Integrated

---

* Corresponding author.

E-mail address: ahmadshafiullah@iut-dhaka.edu (A. Shafiullah).

photovoltaic systems mainly consist of distributed systems, such as small domestic setups, and their primary function is to convert solar energy into electrical power. Renewable sources, including solar radiation, are less harmful to the environment and are recognized as one of the most promising future energy sources [5,6]. However, the intermittent power supply of solar systems can pose challenges to their integration. Various factors, particularly solar radiation, contribute to the variability in energy output [7]. Environmental conditions, such as cloudiness, visibility, etc. directly impact solar irradiance. For example, in regions prone to frequent sandstorms and high particle levels, developing an irradiation prediction model that incorporates dust phenomena is essential, as dust accumulation on PV panels affects the efficiency of solar modules [8,9]. Accurate estimation of these climatic characteristics is essential for developing precise models of solar irradiation. Additionally, connecting large-scale renewable power to the grid presents challenges [10]. The imbalance between supply and demand can cause instability and blackouts. Load balancing, which involves controlling the proportion of energy generated and consumed, is a complex task typically achieved by adjusting output energy and increasing energy production [11,12]. That's why we must ensure the maximum possible production from solar to mitigate the challenge. The variability of solar photovoltaic output power across geographic regions and climatic variables introduces volatility and unpredictability, underscoring the need for accurate solar PV prediction to ensure the reliability of the entire power grid [13]. Precise predictions can assist utility administrations and corporate workers in promptly adjusting and optimizing power generation plans, thereby enhancing the use and economic productivity of new energy sources [14,15]. PV forecast algorithms primarily focus on predicting photovoltaic generation or solar irradiation [16]. Solar forecasting involves creating prediction models that utilize historical data and adhere to data science methodologies [17]. Accurate forecasting of solar resources and photovoltaic power production is of interest to electricity network operators and energy generators due to its impact on power grid maintenance, market structure, and cost reduction. As the popularity of photovoltaics continues to grow, companies are investing heavily in power management systems to improve data collection and enable autonomous resource management [18].

Solar irradiance forecasting has progressed with advancements in forecasting theories and machine learning. With an emphasis primarily on short-term or day-ahead forecasts, several methodologies, including statistical and machine learning approaches, predict solar irradiance at different time horizons [19]. These models can only capture linear relationships and need stationary input data. Some of the statistical methods used include persistence forecasting, Autoregressive (AR), Autoregressive Integrated Moving Average (ARIMA), and Exponential Smoothing Models [20,21]; however, these techniques do not make use of multivariate data, such as relevant meteorological variables. Machine learning-based methods, like Artificial Neural Networks (ANNs) [22], Support Vector Machine (SVM) [23], and K-Nearest Neighbor (KNN) are widely used and show superior accuracy in short-term predictions. Without the complexity of mathematical and physical relationships, ANNs can learn any nonlinear information and produce accurate short-term predictions [24]. In time series forecasting, they do have certain drawbacks. Time series data contain sequential information and have a time order. When dealing with sequential data, the ANN model does not preserve sequential information effectively. Deep Learning techniques like Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) [25] are popular for solar forecasting due to their capacity to characterize high-dimensional nonlinear complex relationships between inputs and outputs [26,27]. Sequential models such as RNN, LSTM, and GRU have a recurrent connection that can capture the sequential relationship of the data during forecasting [28]. RNN-based methods provide better results in comparison to other machine learning models; however, they struggle with multi-step forward prediction. This issue is better served by encoder-decoder architecture, which is used in the fields of machine translation and natural language processing [29]. This architecture is also employed in several time series forecasting tasks. In order to accurately forecast the weather and stock prices, Qin employs a two-stage encoder-decoder method [30]. Using seq2seq models, Bottieau was able to make probabilistic predictions about the cost of various imbalances in the European power markets [31].

Because of the wide range of meteorological variables included in the input data, solar irradiance provides a unique forecasting problem. This multivariate time series data encapsulates a spectrum of input attributes, making it difficult for the existing forecasting models to extract the complex feature correlations and long temporal dependencies of these input features from nonlinear and non-stationary data. Additionally, for multi-step forecasting, the output sequence's temporal dependency coupled with external factors like seasonality makes prediction more challenging. The encoder in the encoder-decoder architecture struggles to capture long temporal relationships for particularly lengthy input sequences since the encoder converts the input sequence into a fixed-length context vector, which could lead to information loss. To address this problem, we present a modeling approach for time series data using the attention mechanism and transformer model in our study. The Attention mechanism was first introduced in the machine translation problem to solve the long-range dependency problem of the encoder-decoder [32]. The Transformer model has recently revolutionized the field of natural language processing by pushing the state-of-the-art and being used for a wide range of tasks, including conversational chatbots, vision-language tasks, and machine translation [33]. It is possible to model time series data with complex temporal relations using transformer-based models. Temporal Fusion Transformer (TFT) is an attention-based transformer model for time series forecasting with a high degree of flexibility and the capacity for multi-step prediction [34]. TFT's attention mechanisms empower it to learn the complex temporal dynamics of time sequences and its capacity to deal with seasonality makes TFT a strategic choice for our study's goals. TFT can take into account a variety of input variables and provide insights on relevant time phases.

In this work, we present the application of attention-based models in multivariate time series forecasting for 24-step forward prediction with a resolution of 30 min with improved accuracy and interpretability. By leveraging attention mechanisms, our approach aims to address critical problems faced by conventional forecasting methods by dynamically emphasizing essential spatiotemporal elements in solar irradiance time series data. Furthermore, the research intends to contribute to the field by offering insights into the interpretability of the attention-based model, resulting in more reliable predictions and therefore increasing the model's adaptability in real-world applications. The key contribution of this paper lies in the application of the Temporal Fusion Transformer (TFT) and attention-based models to the task of solar irradiance forecasting within the particular context of our area, Dhaka and Cox's Bazar, two

places in Bangladesh. Our study includes thorough data preprocessing, model construction, and parameter tuning to improve the performance of TFT and other models, as well as the practicality of TFT by customizing it to our region's distinct geographical and climatic characteristics. We demonstrate the efficiency and applicability of attention-based models in addressing the complex nature of solar forecasting in our region-specific solar data through comprehensive experimentation and comparisons of prediction accuracy between the proposed model and other benchmark forecasting models. The following is how the paper is organized. Section 2 discusses relevant work on deep learning models. Section 3 discusses methodologies, data preparation, and key terminology. Section 4 provides training setups, detailed experimental findings, and further discussions. Section 5 concludes the paper.

2. Related work

Recent advances in the fields of artificial intelligence and deep learning have led to the development of a variety of deep learning models for time series forecasting problems. For such time-series analyses, conventional statistical analysis approaches were previously employed. Due to the availability of relatively large amounts of energy and meteorological data, the use of deep learning algorithms in solar irradiance forecasting over different time horizons, including short, medium, and long-term, is growing increasingly appealing. P. Bendiek et al. [35] introduce DCF, a solar irradiation forecasting algorithm with improved accuracy in three cities (Seattle, Denver, and Boston). The algorithm uses two components: precise ML algorithms (SVM and FBP) and contextual information. SVM performs better for short-term 1-h projections, while FBP is used for longer-term forecasts beyond 3 h due to stability. M. Abdel-Nasser et al. [36] suggested HIFA, a solar irradiation forecasting technique that uses LSTM and GRU networks. It was tested in three Finnish locales and showed better performance compared to three other ensemble techniques with low site RMSE values. N. Yogambal et al. [37] introduce a CSO-GWO optimizer algorithm for multi-timescale solar irradiance predictions using an LSTM-based deep recurrent neural network that outperforms other models in single and multi-timescale forecasting with low MSE and MAPE values.

M. Abdel-Nasser [38] performed a solar irradiance forecasting approach based on LSTM models aggregated by the Choquet integral which provides accurate forecasts and eliminates the need for costly meteorological equipment. X. Huang et al. [39] presented a two-branch input LSTM-MLP structure for solar irradiance forecasting, which includes main output, main input, auxiliary input, and auxiliary output, as well as LSTM layers that use irradiance history and meteorological parameters. Model II-BD outperforms other models by using historical irradiance and meteorological features as main inputs and next-instant meteorological data as auxiliary inputs. G. Guariso et al. [40] validated the accuracy of FF and LSTM networks for predicting environmental variable time series, emphasizing the effect of null values and midnight samples on performance metrics. J. Wojtkiewicz et al. [41] employ univariate and multivariate GRU and LSTM models to predict Phoenix, Arizona's solar irradiance based on historical data, weather variables, and cloud cover data.

GRU attention, a hybrid deep learning model built on Keras, was introduced by K. Yan et al. [42] for solar irradiance prediction and has shown good prediction accuracy, quick modeling, and high portability. The authors emphasized the advantages of utilizing deep learning to estimate power generation stability, dependability, and precision. Y. Yu et al. [17] developed a short-term LSTM model to forecast solar irradiance and tested it in Atlanta, New York, and Hawaii 1 h and one day ahead. With low MAPE values in all three cities, LSTM outperforms other models, particularly on cloudy and mixed days. M. Husein et al. [43] proposed a deep LSTM RNN for solar irradiance forecasting using external features such as dry bulb temperature, dew point temperature, and relative humidity. The model showed an average root mean square error of 80.07 W/m^2 across six datasets, outperforming traditional feedforward neural networks (FFNN). S. Dev et al. [44] proposed a solar irradiance forecasting approach based on clearness index data and triple exponential smoothing to accurately reflect seasonality.

Tong et al. [45] propose an encoder-decoder deep hybrid model combining TCN, LSTM, and MLP, enhanced by dynamic error compensation, achieving balanced multi-step forecasting through unique loss functions. Li et al. [46] suggest a two-channel method employing LSTM, WGAN, and CEEMDAN, splitting solar output into frequency-based subsequences for prediction, and integrating their values for final output. Hou et al. [47] introduce CNN-A-LSTM, employing comparable day analysis and attention processes, surpassing various models on the NSRDB dataset for accurate solar irradiance prediction, particularly excelling in unclouded and partly cloudy conditions. Munsif et al. [48] explore the CT-NET model, a transformer variation combining CNN and multi-head attention for both local and global information utilization, outperforming CNN-RNN, CNN-GRU, and CNN-LSTM across seasons using the Alice Springs dataset. Yang et al. [49] developed a model with RACB, DIFM, and TSAM components, demonstrating improved accuracy and resilience in multi-step forecasting compared to TCN, LSTM, LSTM-Attention, CNN-LSTM, and Transformer models across various locations. Kong et al. [50] utilize EMD, GRU-A with attention, and Kalman filtering for accurate solar radiation forecasting, proving its effectiveness against RNN, GRU, EMD-GRU, and GRU-A models.

Previous research has primarily focused on traditional approaches such as statistical models, Artificial Neural Networks (ANN), and sequence models such as Long Short-Term Memory (LSTM) networks. While these techniques provided useful insights and advances, their difficulties in dealing with multivariate time series data and capturing complex temporal correlations in solar irradiance data still need to be addressed. Moreover, the existing literature reveals challenges in achieving optimal forecasting accuracy, particularly when dealing with volatility and unpredictability, as well as the inability to demonstrate good generalization across different geographical locations, which pose barriers to achieving robust and accurate predictions. Transformer models have recently been integrated into time series forecasting problems, even though there is a discussion about whether or not transformers are effective for time series data [51]. There are very limited works utilizing the advantages of attention-based models and transformers while some prior studies used transformer models to estimate direct PV power using historical power generation data [52]. Considering these limitations, our study aims to address them by introducing the Temporal Fusion Transformer (TFT) to the area of solar irradiance forecasting and applying

this model directly to a real-world scenario, especially forecasting solar irradiance at two specific sites in Bangladesh: Dhaka and Cox's Bazar. These two locations have different geographical features, such as climate, distance from the sea, and seasonality, that affect the availability and variability of solar resources. This study focuses on solar irradiance data as the input and output to our model with other meteorological variables to increase the applicability to different regions and enhance our understanding of the dynamic patterns and complexity driving energy output. In addition, we examine and compare the effectiveness of the TFT, transformer, and attention-based models in comparison to other well-established models, offering enhanced accuracy and adaptability in solar irradiance predictions, particularly in our specific geographical and climatic setting.

3. Methodology

3.1. Seq2seq encoder-decoder

The Sequence-to-Sequence encoder-decoder architecture was developed [29,53] to encode and produce a sequence of any length for machine translation tasks with sequential input and output. The architecture has two RNN networks called encoder and decoder. After recursively processing the input sequence $(x_1, x_2, \ldots x_\tau)$ of length τ, the encoder RNN computes a fixed-length representation h_τ of the final hidden state vector which recapitulates the entire input sequence. The decoder is another RNN network that produces a target sequence $(s_1, s_2, \ldots, s_{\tau'})$ of length τ' that employs the encoder's hidden state as its initial state. The decoder generates the target iteratively, and at each step, it utilizes the previous step's output as well as the previous hidden state as input. It should be noted here that the lengths of the input and output sequences may differ. Either a basic RNN, an LSTM [54], or a GRU [55] may be used as the RNN in the encoder and decoder. Each hidden state of the encoder in a basic RNN is calculated using equation (1).

$$h_t = \delta(w_{hh} * h_{t-1} + w_{hx} * x_t) \tag{1}$$

Weight matrices w_{hh} and w_{hx} link the input and the encoder's hidden states, respectively, where δ is the activation function and h_t stands for the encoder's hidden states.

Given an input sequence $(x_1, x_2 \ldots x_\tau)$ whose fixed length hidden state representation is h_τ, the conditional probability of the output sequence $p(s_1, s_2 \ldots, s_{\tau'} | x_1, x_2 \ldots x_\tau)$ is formulated in equation (2).

$$p(s_1, s_2 \ldots, s_{\tau'} | x_1, x_2 \ldots x_\tau) \approx \prod_{t=1}^{\tau'} p(s_t | h_\tau, s_1, \ldots s_{t-1}) \tag{2}$$
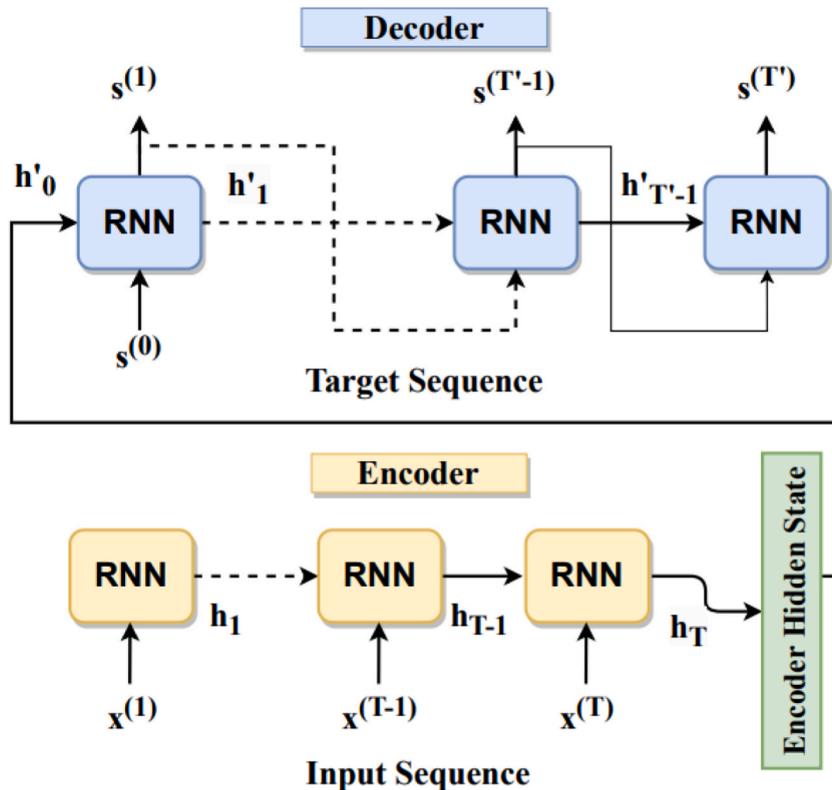
Fig. 1. The RNN encoder-decoder architecture.

The encoder-decoder model's architecture was designed for language modeling, and the input and output sequences are both represented as word embeddings, which are learned numerical vector representations for text. The decoder initializes with a start token or a dummy input to begin the prediction. However, the preceding value to the target sequence is known to our time series task. Additionally, the input and output sequences don't share the same size of feature representation. The dataset we're using here has multiple features in each sequence hence it is called multivariate time series forecasting, whereas the output sequence only has one feature. Therefore, we adapt the model to our problem in that manner. Here, the prior true output value shown in Fig. 1 is not known by the decoder; instead, it only has access to the initial target value $s^{(0)}$ during the prediction phase. So, the decoder updates the sequence $(s^{(1)}, s^{(2)}, \ldots, s^{(T)})$ using the probability distribution it obtained from the prior state. There are several methods for updating decoder predictions during training. Recursive prediction is one way. That is, the previously predicted decoder outputs feed into the decoder recurrently until we obtain an output of the desired target length. One disadvantage of this strategy is that if the predictions are too poor in the early stages of training, the errors will accrue over the sequence length, making it harder for the model to learn and converge rapidly. Another method is using teacher forcing [56,57]. In teacher forcing, the model's decoder makes predictions based on the true previous target value. It forces the sequence model to stay near the true sequence. This approach has one drawback: there is no true target value during inference. We need to forecast recursively during inference, resulting in a discrepancy between training and inference. So, we adopted a hybrid of the two approaches. Using a ratio, we combined two approaches by giving the decoder the projected value in some steps and providing it with the true value at other times. This ratio is designated as *TFR*.

3.2. Encoder-decoder with attention mechanism

In encode-decoder attention model, the time series input sequence is read by the encoder, which then transform into hidden states (h_{en}) to create a fixed-size context vector (c_i) representation of the data. The context vector is then utilized by the decoder to generate an output sequence based on the previously generated output (y_{t-1}) and the previous hidden state $(h_{de,i-1})$. The attention mechanism is used at each decoding step to continuously select information from the hidden states, adjusting the context vector based on the decoder's current state. The attention mechanism starts by generating an alignment score using the decoder's hidden states and each of the encoder's hidden states, which is then transformed into attention weights. Then, the context vector c_i is generated by using the attention weights and weighted-summarizing of encoder hidden states h_{en} which is displayed in equation (3).
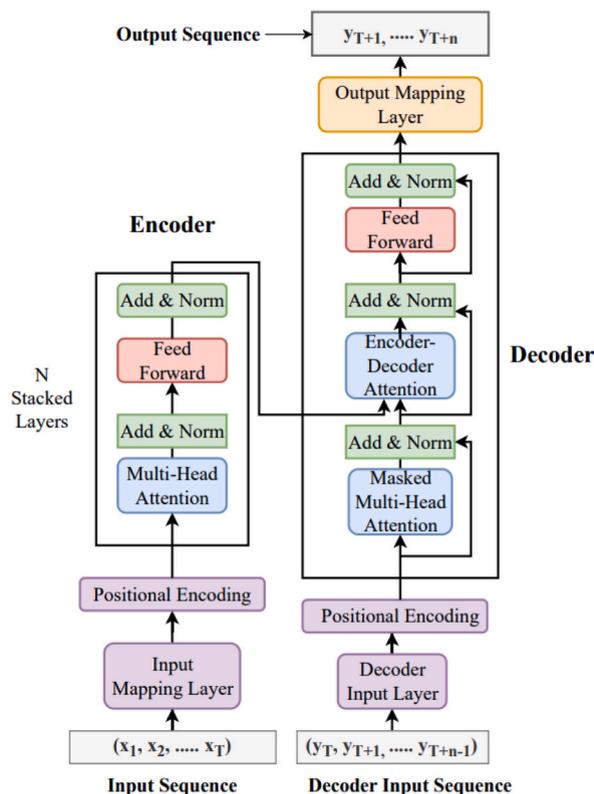
Fig. 2. Transformer encoder decoder layer.

$$c_i = \sum_{j=1}^{T_X} a_{ij} h_{en,j} \tag{3}$$

Using equations (4) and (5) each annotation's value a_{ij} is determined.

$$e_{ij} = a\big(h_{de,i-1}, h_{en,j}\big) \tag{4}$$

$$a_{ij} = \frac{\exp\big(e_{ij}\big)}{\sum\limits_{k=1}^{T_X} exp(e_{ik})} \tag{5}$$

The GRU and LSTM layers used in the encoder of the attention-based model are bidirectional. Mixed recursive and teacher-forcing methods were used for the training phase as mentioned in the preceding section.

3.3. Transformer for time series

In 2017, researchers from Google Brain unveiled the first-ever transformer [33]. To adapt the transformer model for time series forecasting, Neo et al. [58] created a variant that maintains the original structure of encoder-decoder layers. In the original transformer model, which was developed to solve the machine translation issue, the embedding size is utilized as the D_{model} dimensional vector value throughout the encoder and the decoder. This ensures that the feature size of the input and output text data is the same. In this scenario, input and output time series data may have different characteristics. Fig. 2 depicts the input layer of the encoder, which is a fully connected neural network used to map the input data's attributes onto a D_{model} dimensional vector. The decoder also has a layer like an encoder to translate the output data to the D_{model} dimensional vector.

In multi-headed attention, the time series data is linearly transformed to obtain query vectors (Q), key vectors (K), and value vectors (V) and each of these transformed vectors is split into multiple heads. Using the scaled dot-product attention mechanism, each attention head separately computes attention scores. To generate attention output, the outputs of all attention heads are concatenated and linearly transformed, as presented in equation (6).

$$Attention(Q, K, V) = softmax\frac{\big(Q.K^T\big)}{\sqrt{d_k}}.V \tag{6}$$

Positional encoding is used to capture the sequential information of the input data since our model does not include a sequential unit like an RNN. In addition, masking is used in the decoder's output sequence to ensure that only preceding data points in the time series are included in the prediction. A normalizing layer is included underneath each sublayer.

3.4. Temporal Fusion Transformer (TFT)

Temporal Fusion Transformer (TFT) [34] provides a neural network design that combines the features of other networks, such as LSTM layers and Transformers' attention heads. TFT is able to accommodate three distinct kinds of features. They are temporal data with known inputs into the future, temporal data known only up to the present, and external categorical or static variables, which are also referred to as time-invariant features. The model has a high degree of adaptability with the capability of multi-step prediction. Certain time sequences might be rather complicated or noisy, but others can be easily modeled using seasonal naive predictors and require very little effort. In an ideal world, the model would be able to distinguish between these distinct kinds of situations. There is also the possibility of success with one-step-ahead prediction models that recursively feed forecasts.

In order to adapt to a broad variety of datasets and use cases, the architecture may be equipped with gating mechanisms that allow data to bypass unused parts of the network, as shown in equations (7)–(11).

$$GRN_w(a, c) = LN(a + GLU_{w1}(n_1)) \tag{7}$$

$$n_1 = [TCN_1(\text{ŋ}_2), TCN_2(\text{ŋ}_2)] \tag{8}$$

$$\text{ŋ}_2 = GLU_{w2}(\text{ŋ}_3) \tag{9}$$

$$\text{ŋ}_3 = W_{1,w}\text{ŋ}_4 + b_{1,w} \tag{10}$$

$$\text{ŋ}_4 = ELU\big(W_{2,w}a + W_{3,w}c + b_{2,w}\big) \tag{11}$$

In these equation, ELU is represented as the Exponential Linear Unit activation function, $\text{ŋ}_1, \text{ŋ}_2, \text{ŋ}_3, \text{ŋ}_4$ are represented as intermediate layers, LN is represented as standard layer normalization, ŋ_1 is the result of concatenating $TCN_1(\text{ŋ}_2)$ and $TCN_2(\text{ŋ}_2)$ and w is represented as weight sharing.

At each time step, variable selection networks choose the right set of input variables. In order to include static characteristics in the network, context vectors are encoded and used to condition the temporal dynamics using static covariate encoders. For the purpose of local processing, a sequence-to-sequence layer is used, and for the purpose of capturing long-term dependencies, an innovative

interpretable multi-head attention block is provided. Quantile forecasting intervals are used to determine the probable range of goal values at each time step in the forecasting process.

3.5. Data Description

The historical irradiance data utilized for the system modeling and validation for this study came from the National Solar Radiation Database (NSRDB) [59] over the period of January to December from two consecutive years 2019 and 2020. To assess the robustness of the models, it is necessary to investigate data from several locations. Dhaka (23.8° N, 90.41° E) and Cox's Bazar (21.46° N, 92.01° E) are the two locations in Bangladesh that were utilized in this study. Table 1 below shows the statistical characteristics of the data for these two locations.

The dataset contains a total of 70,176 data points from two locations with a temporal resolution of 30 min and has no missing values. Global Horizontal Irradiation (GHI), one of the three solar irradiation components included in this database, is chosen as the target variable for our experiment. Fig. 3 displays the Global Horizontal Irradiation distribution for Dhaka for different months in 2019. The figure shows that solar irradiance varies between the hours of each day and that each month has a different peak.

Due to various weather conditions, the distribution of solar irradiance in different locations varies substantially. In cloudy or rainy conditions, the solar irradiation value is highly uncertain and variable.

Fig. 4(a and b) shows the solar irradiance for two different weather scenarios: a clear sky and cloud cover, during the course of the day. Data exhibits a pattern on days with a clear sky. However, when there is cloud cover, GHI readings become extremely irregular and exhibit a sharp drop in the curve.

To enhance the forecasting ability of our model, we incorporate meteorological data, which is also provided by the National Solar Radiation Database, along with the solar irradiance data. The properties of the meteorological data are shown in Table 2.

3.6. Feature selection

Numerous meteorological factors can be thought of as possible factors that can have an impact on the solar radiation that a surface receives from above. In order to choose an optimum feature subset as the model input, it is necessary to differentiate the particular features linked to weather conditions into those that are useful to the model and those that are irrelevant. Pearson's correlation coefficient is the measure of the statistical relationship between two continuous variables. To decide which factors should be used as inputs, the correlation between GHI and other meteorological variables was examined. Table 3 displays the dataset's solar irradiance and weather variables' Pearson correlation coefficients.

The correlation between GHI and the various weather variables differs by location, indicating that the climate condition has an impact on these parameters. A minimum value of 0.2 for the absolute value of Pearson's correlation coefficients in either location was chosen to determine the inclusion of the features. From the table, it can be seen that Temperature, Humidity, Solar Zenith Angle, and Wind Speed were deemed to be critical for the model and that the remaining parameters were excluded since they showed no significant correlation with the GHI.

3.7. Feature transform and encoding

Cloud type is a categorical feature that represents different cloud conditions and weather types. It is an important feature since cloud condition is responsible for the abrupt change in radiation received at the surface. One-hot encoding is used since this feature doesn't have any ordinal relationships. DateTime variable is also an important feature as there is a strong correlation between GHI and time which can be seen in Fig. 3. One-hot encoding is not suitable for this feature as there are too many categories. Moreover, the variables have a cyclical relationship that one-hot encoding can't address. For instance, although appearing to be separated by 11 months in categorical value, December and January are only 1 month apart. To resolve this problem, we encoded the cyclic feature using sine and cosine transformations, as shown in equations (12) and (13).

$$T_{sin} = sin\left(\frac{2.\pi.T}{max(T)} \right) \tag{12}$$

$$T_{cos} = cos\left(\frac{2.\pi.T}{max(T)} \right) \tag{13}$$

Table 1
Statistical features of the solar irradiance data.

Location	GHI(W/m^2)		
	Max	Mean	Std.
All samples	1017	207.23	287.50
Dhaka	994	200.24	278.47
Cox's Bazar	1017	214.23	296.09
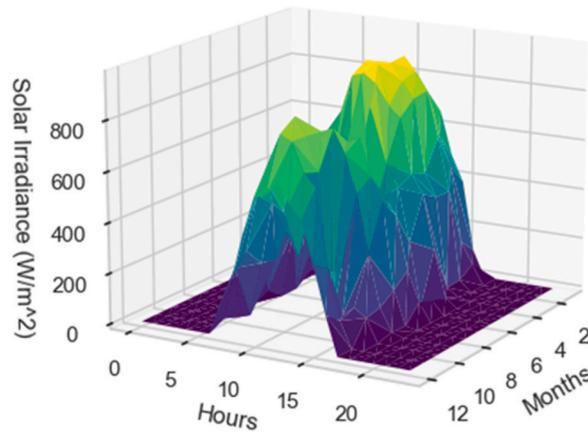
Fig. 3. Solar irradiation data in Dhaka during 2019.
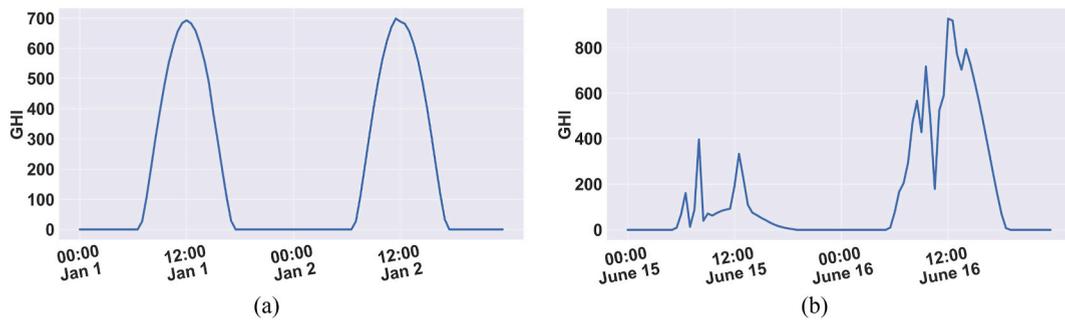
Fig. 4. Global Horizontal Irradiation during (a) clear-sky (b) cloudy day.

Table 2
Meteorological parameters.

Variable Name	Unit
Global Horizontal Irradiance	W/m^2
Ozone	
Solar Zenith Angle	Degree
Precipitable Water	cm
Temperature	°C
Dew Point	°C
Relative Humidity	%
Pressure	mbar
Wind Direction	Degree
Wind Speed	m/s

Table 3
Pearson's correlation coefficients between meteorological parameters and GHI.

Weather Variables	Dhaka	Cox's Bazar
Ozone	0.064	0.047
Solar Zenith Angle	−0.815	−0.817
Precipitable Water	−0.002	−0.048
Temperature	0.510	0.271
Dew Point	0.018	−0.021
Relative Humidity	−0.547	−0.470
Pressure	−0.007	0.057
Wind Direction	0.054	0.093
Wind Speed	0.227	−0.033

3.8. Data scaling and splitting

Different continuous input variables' scales may result in slow learning or cause it to become trapped in local optimums. If the scale or distribution of the time series data is constant, gradient descent-based algorithms, such as neural networks, would perform better. This necessitates that we need to normalize the data such that each feature has the same scale and significance. Standardization (z-score), a technique that rescales the distribution of values with a zero mean and a standard deviation of 1, is used in this study to rescale the data. The z-score normalization formula is as follows in equation (14):

$$x_z = \frac{x_i - \overline{x}}{\sigma} \tag{14}$$

where x_i is the input data, $\overline{x}$ denotes the mean of the feature vector, and σ denotes the feature vector's standard deviation.

For training purposes, the complete dataset is split into three sets: train, validation, and test sets. 75% of the data, covering the first year (2019) and the first six months of 2020, are in the training set, which is used to fit the models. The remaining six months are split between the test (12.5%) and validation (12.5%) sets. The validation set is used to provide an unbiased assessment of a fitted model while fine-tuning its hyperparameters whereas the test set is used to evaluate the final model. Since it is necessary to preserve the temporal order of time series data, data points are not shuffled while splitting.

3.9. Performance criterion

Four performance metrics, including the mean square error (MSE), the mean absolute error (MAE), the mean absolute scaled error (MASE), and the coefficient of determination (R^2) are used in the forecasting experiments to assess the forecasting accuracy of our models.

MSE stands for Mean Squared Error which is shown in equation (15). It measures the average of the squared differences between the actual and estimated values.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 \tag{15}$$

MAE stands for Mean Absolute Error which is presented in equation (16). It calculates the sum of the absolute differences between the actual and predicted values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i| \tag{16}$$

MASE stands for Mean Absolute Scaled Error which is exhibited in equation (17). It evaluates the accuracy of forecasts by comparing the mean absolute error of the forecast values with the mean absolute error of a naive model. A Naive model is a simple baseline model that forecasts the future value to be the same as the previous one.

$$MASE = \frac{MAE}{MAE_{naive}} \tag{17}$$

R^2 is a coefficient of determination which is shown in equation (18). It indicates how well the model fits the data by comparing the total variance explained by the model and the total variance in the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{18}$$

here, y_i *and* $\widehat{y}_i$ represent the actual and predicted values, respectively, while $\overline{y}$ indicates the mean of the actual values.

4. Results and analysis

From the datasets of two different locations, multi-step solar irradiance is forecasted using different sequence-to-sequence

Table 4
Selected parameters for Encoder-Decoder & Attention-Based GRU and LSTM model.

Parameter	GRU-ED	LSTM-ED	GRU-attn	LSTM-attn
Layers	1	1	1	1
Encoder hidden size	64	48	32	32
Decoder hidden size	64	48	32	32
Learning rate	0.0005	0.0005	0.0005	0.0005
Input sequence length	48	48	48	48
TFR	0.6	0.5	0.6	0.5
Dropout	0	0	0	0
Batch size	256	256	256	256

attention-based models. As a multi-step ahead time series forecasting, the model predicts 12 h ahead of the Global Horizontal Irradiance(GHI) value using the last 24 h of data as the input sequence. According to the methods described in the preceding section, Transformer, GRU and LSTM Encoder-Decoder (GRU-ED, LSTM-ED), GRU and LSTM Encoder-Decoder with attention (GRU-attn, LSTM-attn) models were developed and trained in Pytorch. The TFT model was trained using the Pytorch implementation in Pytorch Forecasting [60]. As the various hyperparameters, like learning rate and hidden units, significantly impact the model's performance, we tuned the hyperparameters of the models using Optuna [61]. The optimization method used in this experiment is the Adam optimizer. The selected hyperparameters for our forecasting models are presented in Tables 4 and 5.

The performance of the sequence-to-sequence models is also compared with the simple MLP and Naive models. The Naive model uses the previous value or period to forecast the next value/period. Because we are forecasting sequences, the naive model will anticipate the following day's irradiance based on the value from the previous day. To compare our sequence models, we also construct a simple MLP model that predicts sequence recursively. Sometimes MLP model performs well on several occasions in time series forecasting [62,63]. The MLP model used in this experiment has 2 hidden layers, each with 64 hidden units.

The evaluation metrics of these forecasting models for the two different locations are shown in Table 6.

As seen in the table, almost all forecasting models can forecast with reasonable accuracy when compared to the naive model. The table also shows that TFT outperforms the other models for most of the metrics in both locations. After the Naive model, ANN and Transformer perform worse compared to other models overall.

In time series forecasting, sequential models generally outperform MLP because they contain recurrent structures that can store sequential data. Here, at Cox's Bazar location, GRU-ED and LSTM-ED outperform MLP across all parameters, with GRU-ED doing the best. MLP outperforms LSTM-ED in Dhaka in terms of MSE and MASE values, however, LSTM-ED is more effective in terms of MAE and R^2. In this case, GRU-ED also gives superior results than MLP and LSTM-ED. GRU-ED model has shown better results in Cox's Bazar location than attention models, with MSE and MAE values of 0.152 and 0.219, respectively. In Dhaka, GRU and LSTM attention models beat MLP and encoder-decoder models, while the GRU-attn model performs the best and even outperforms TFT in terms of MSE and R^2 score. The effectiveness of the attention mechanism is evident as it facilitates attention-based models in retaining all prior information in long sequences. The attention mechanism assesses all hidden states from the encoder sequence and also assigns relative importance to the time steps and features that affect output when formulating predictions, thus improving the prediction accuracy.

The Transformer model performs the worst in both locations, slightly outperforming the Naive model. Although the Transformer model does well throughout the training phase, it does poorly in the testing data. Finally, the TFT model beats all other models in Cox's Bazar location with the lowest MSE, MAE, and MASE loss and high R^2 value. Only GRU-attn has a better MSE and R^2 value than TFT with values of 0.153 and 0.809 in Dhaka. TFT has the best MAE and MASE scores in this location. The TFT model can handle a variety of input data, including static covariates, future known inputs, and temporal variables known just up to the present. The model can also be trained on multiple time series. This algorithm combines a temporal self-attention decoder with a novel Multi-head attention mechanism that, when evaluated, gives additional insight into feature importance in order to capture long-term dependencies.

The actual data and predicted outcomes for the various models in both locations and for the two weather conditions are shown in Fig. 5(a and b) and 6(a,b). Our forecasting algorithms predict 24 steps ahead of the data. On days with cloud cover, as shown in Figs. 5 (b) and 6(b), algorithms can capture the uncertainty and volatility in solar data. Due to the high level of weather unpredictability on cloudy days, models work better when the sky is clear than when it is cloudy.

Better performance in forecasting is achieved in the location of Cox's Bazar. Almost every forecasting model performs better in this location. This might be because the seasonality pattern is more consistent in this location and there is less residual or randomness owing to the cloudy and variability in weather conditions. Moreover, the same information can be observed through the Naive model, where the error values are smaller in Cox's Bazar than in Dhaka. We may infer that Cox's Bazar data follow seasonality with less unpredictability since the Naive model predicts the upcoming period using the prior period. The TFT model shows more consistency in both locations with MSE values of 0.154 and 0.147 and MAE values of 0.215 and 0.210 respectively. Attention models also perform well in both locations although they have better values in the Dhaka location. All of the other models projected inconsistently for the two separate locations. TFT's ability to maintain consistent performance levels across varying contexts implies that it is a robust choice for diverse patterns.

To provide a thorough assessment of our solar prediction models, the test datasets from two locations are combined to compute the error metrics of the total test datasets, as shown in Table 7. The combination of results allows for a comparative analysis, which provides insights into the models' overall performance under two distinct environmental settings. Table 7 demonstrates TFT's superior

Table 5
Selected parameters for the Transformer and Temporal Fusion Transformer (TFT) model.

Transformer		Temporal Fusion Transformer	
Parameter	Value	Parameter	Value
Layers	3	Layers	1
D_{model}	24	Hidden size	32
D_{ff}	16	Hidden continuous size	16
Attention heads	8	Attention heads	4
Learning rate	0.0005	Learning rate	0.0001
Input sequence length	48	Input sequence length	48
Dropout	0.2	Dropout	0.2
Batch size	256	Batch size	256

Table 6

Forecasting metrics for the different models in two locations.

	Dhaka				Cox's Bazar			
Model	MSE	MAE	MASE	R^2	MSE	MAE	MASE	R^2
Naive	0.302	0.283		0.622	0.277	0.263		0.668
MLP	0.180	0.243	0.858	0.775	0.171	0.241	0.916	0.796
GRU-ED	0.179	0.232	0.819	0.776	0.152	0.219	0.833	0.818
LSTM-ED	0.183	0.236	0.834	0.770	0.156	0.227	0.863	0.814
GRU-attn	0.153	0.231	0.816	0.809	0.160	0.242	0.920	0.809
LSTM-attn	0.160	0.219	0.773	0.799	0.164	0.236	0.897	0.804
Transformer	0.1945	0.271	0.957	0.757	0.1865	0.296	1.125	0.777
TFT	0.154	0.215	0.759	0.806	0.147	0.210	0.798	0.824
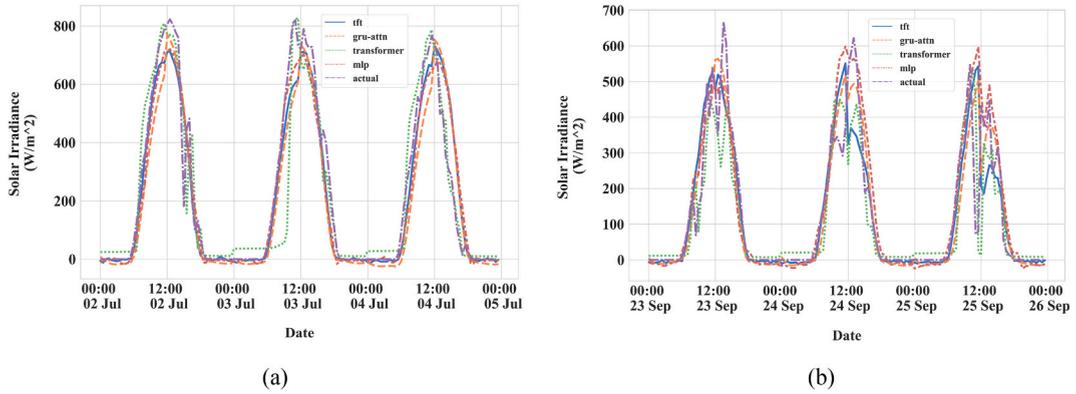
Fig. 5. Predicted solar irradiance for different models in Dhaka during (a) clear-sky (b) cloudy days.
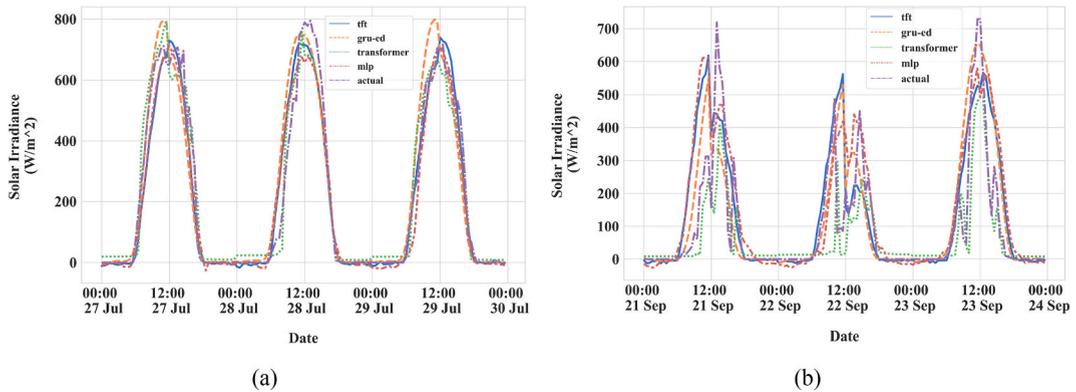
Fig. 6. Predicted solar irradiance for different models in Cox's Bazar during (a) clear-sky (b) cloudy days.

Table 7

Overall Forecasting metrics for the different models in both locations.

Model	MSE	MAE	MASE	R^2
Naive	0.290	0.273		0.646
MLP	0.176	0.242	0.886	0.785
GRU-ED	0.165	0.226	0.828	0.798
LSTM-ED	0.169	0.231	0.846	0.794
GRU-attn	0.157	0.236	0.864	0.808
LSTM-attn	0.162	0.227	0.831	0.802
Transformer	0.190	0.270	0.989	0.767
TFT	0.151	0.212	0.776	0.815

performance in comparison to other forecasting models, with TFT having a better value in all error metrics, with a 0.151 MSE and 0.212 MAE value while the 0.776 MASE and 0.815 R^2 scores further corroborate its superior performance. Overall experimental results show that the TFT's performance is on par with the attention models and outperforms Encoder-Decoder models and a simple estimator (Naive model). In contrast to the encoder-decoder architecture, which fails to capture information because of its fixed-length context vector representation, attention-based models are able to collect information in long input sequences. Particularly, we illustrate the benefits of the attention mechanisms which provide a clear view into the decision-making process, allowing models to gain insights into specific meteorological components and temporal patterns influencing solar irradiance forecasts. We also observed that the GRU and LSTM architecture in the Encoder-Decoder and Attention models function similarly despite having different architectural designs, with GRU marginally outperforming LSTM. Our results demonstrate that the TFT consistently surpasses traditional sequential models and other attention-based architectures in both locations, showcasing its robustness and effectiveness in capturing the intricate patterns inherent in our region's solar data. However, since TFT is more computationally expensive due to containing significantly more parameters, a careful trade-off between model complexity and training efficiency is required.

5. Conclusion

In this paper, we presented an Attention-based deep learning framework to address the multivariate multistep Time Series Forecasting problem. Attention-based encoder-decoder, transformer, and Temporal Fusion Transformer (TFT) models are evaluated to forecast 24 steps forward solar irradiance at two different locations in Bangladesh. The dataset with an interval of 30 min includes information on cloud cover, meteorological variables, and historical solar irradiance values. The unpredictable nature of the weather makes it challenging to forecast solar irradiance, which leads to imbalances in the interconnected grid. Our primary motivation was to assess the attention mechanism's capabilities to address the complicated and dynamic nature of solar irradiance patterns, therefore contributing to the grid and optimizing renewable energy utilization. According to the results, the TFT model had superior outcomes than other existing models such as MLP and sequential encoder-decoder models, across all performance measures. Attention-based GRU Encoder-Decoder, which has the best MSE and R^2 score in the Dhaka location, was the second-best method after TFT. The Transformer model for the Time Series performed the worst out of all the models used. In comparison to the other models' inconsistent predictions, the empirical results exhibit a significant decrease in forecasting errors, as well as the consistency and robustness of TFT in two separate locations in our specific region, proving its usefulness in real-world applications. As the need for clean and renewable energy sources increases, our research contributes to assisting energy management in making informed decisions for sustainable energy integration into the grid and more reliable and efficient utilization of solar energy. It is important to recognize several limitations of our study. Firstly, our work primarily focuses on a specific time horizon for solar radiation predictions; future studies could investigate multiple time horizons to further assess the robustness of forecasting methodologies. Furthermore, the training period for TFT and other attention models is relatively high, which could lead to potential practical issues in situations when quick model response is necessary. Despite these limitations, our research demonstrates the importance of the application of the TFT model and incorporating the attention mechanism to overcome the issues associated with solar irradiation variability.

Data availability

Solar Irradiance Forecasting: Dataset from NSRDB (National Solar Radiation Database) was used in order to support this study and is available at "https://nsrdb.nrel.gov/". The dataset is cited at relevant places within the text as Ref [59].

CRediT authorship contribution statement

Sadman Sakib: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mahin K. Mahadi:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Samiur R. Abir:** Software, Methodology, Investigation, Formal analysis. **Al-Muzadded Moon:** Validation, Resources, Methodology. **Ahmad Shafiullah:** Writing – review & editing, Supervision, Software, Project administration, Methodology. **Sanjida Ali:** Writing – review & editing, Validation. **Fahim Faisal:** Writing – review & editing, Supervision, Project administration. **Mirza M. Nishat:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

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

References

[1] L.G. Thompson, Climate change: the evidence and our options, Behav. Anal. 33 (2) (Oct. 2010) 153–170, https://doi.org/10.1007/BF03392211.
[2] P. Newell, A. Simms, How Did We do that? Histories and political economies of rapid and just transitions, New Polit. Econ. 26 (6) (Nov. 2021) 907–922, https://doi.org/10.1080/13563467.2020.1810216.
[3] F. Wang, Z. Zhen, Z. Mi, H. Sun, S. Su, G. Yang, Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting, Energy Build. 86 (Jan. 2015) 427–438, https://doi.org/10.1016/j.enbuild.2014.10.002.
[4] S.A. Reinders, P. Verlinden, A. Freundlich, Photovoltaic Solar Energy: from Fundamentals to Applications, John Wiley, 2017.

[5] S. Jiang, C. Wan, C. Chen, E. Cao, Y. Song, Distributed photovoltaic generation in the electricity market: status, mode and strategy, CSEE J. Power Energy Syst. 4 (3) (Sep. 2018) 263–272, https://doi.org/10.17775/CSEEJPES.2018.00600.

[6] P. Hanser, R. Lueken, W. Gorman, J. Mashal, The practicality of distributed PV-battery systems to reduce household grid reliance, Util. Pol. 46 (Jun. 2017) 22–32, https://doi.org/10.1016/j.jup.2017.03.004.

[7] M.Q. Raza, M. Nadarajah, C. Ekanayake, On recent advances in PV output power forecast, Sol. Energy 136 (Oct. 2016) 125–144, https://doi.org/10.1016/j.solener.2016.06.073.

[8] T. Sarver, A. Al-Qaraghuli, L.L. Kazmerski, A comprehensive review of the impact of dust on the use of solar energy: history, investigations, results, literature, and mitigation approaches, Renew. Sustain. Energy Rev. 22 (Jun. 2013) 698–733, https://doi.org/10.1016/j.rser.2012.12.065.

[9] S.A. Sulaiman, A.K. Singh, M.M.M. Mokhtar, M.A. Bou-Rabee, Influence of dirt accumulation on performance of PV panels, Energy Proc. 50 (2014) 50–56, https://doi.org/10.1016/j.egypro.2014.06.006.

[10] Y. Jia, X. Lyu, C.S. Lai, Z. Xu, M. Chen, A retroactive approach to microgrid real-time scheduling in quest of perfect dispatch solution, J. Mod. Power Syst. Clean Energy 7 (6) (Nov. 2019) 1608–1618, https://doi.org/10.1007/s40565-019-00574-2.

[11] K.S. Perera, Z. Aung, W.L. Woon, Machine Learning Techniques for Supporting Renewable Energy Generation and Integration, A Survey, 2014, pp. 81–96.

[12] A. Fouilloy, et al., Solar irradiation prediction with machine learning: forecasting models selection method depending on weather variability, Energy 165 (Dec. 2018) 620–629, https://doi.org/10.1016/j.energy.2018.09.116.

[13] F. Wang, Y. Yu, Z. Zhang, J. Li, Z. Zhen, K. Li, Wavelet decomposition and convolutional LSTM networks based improved deep learning model for solar irradiance forecasting, Appl. Sci. 8 (8) (2018) 1286, https://doi.org/10.3390/app8081286. Aug.

[14] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan, Y. Du, Short-term photovoltaic power forecasting based on long short term memory neural network and attention mechanism, IEEE Access 7 (2019) 78063–78074, https://doi.org/10.1109/ACCESS.2019.2923006.

[15] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de-Pison, F. Antonanzas-Torres, Review of photovoltaic power forecasting, Sol. Energy 136 (Oct. 2016) 78–111, https://doi.org/10.1016/j.solener.2016.06.069.

[16] J. Kleissl, Solar Energy Forecasting and Resource Assessment, Acad. Press, 2013.

[17] Y. Yu, J. Cao, J. Zhu, An LSTM short-term solar irradiance forecasting under complicated weather conditions, IEEE Access 7 (2019) 145651–145666, https://doi.org/10.1109/ACCESS.2019.2946057.

[18] R.B. Melton, et al., Leveraging standards to create an open platform for the development of advanced distribution applications, IEEE Access 6 (2018) 37361–37370, https://doi.org/10.1109/ACCESS.2018.2851186.

[19] R. Baños, F. Manzano-Agugliaro, F.G. Montoya, C. Gil, A. Alcayde, J. Gómez, Optimization methods applied to renewable and sustainable energy: a review, Renew. Sustain. Energy Rev. 15 (4) (May 2011) 1753–1766, https://doi.org/10.1016/j.rser.2010.12.008.

[20] G. Reikard, Predicting solar radiation at high resolutions: a comparison of time series forecasts, Sol. Energy 83 (3) (Mar. 2009) 342–349, https://doi.org/10.1016/j.solener.2008.08.007.

[21] Z. Dong, D. Yang, T. Reindl, W.M. Walsh, Short-term solar irradiance forecasting using exponential smoothing state space model, Energy 55 (Jun. 2013) 1104–1113, https://doi.org/10.1016/j.energy.2013.04.027.

[22] S.P. Durrani, S. Balluff, L. Wurzer, S. Krauter, Photovoltaic yield prediction using an irradiance forecast model based on multiple neural networks, J. Mod. Power Syst. Clean Energy 6 (2) (2018) 255–267, https://doi.org/10.1007/s40565-018-0393-5. Mar.

[23] M. Pan, et al., Photovoltaic power forecasting based on a support vector machine with improved ant colony optimization, J. Clean. Prod. 277 (2020) 123948, https://doi.org/10.1016/j.jclepro.2020.123948. Dec.

[24] M. Marzouq, H. El Fadili, K. Zenkouar, Z. Lakhliai, M. Amouzg, Short term solar irradiance forecasting via a novel evolutionary multi-model framework and performance assessment for sites with no solar irradiance data, Renew. Energy 157 (Sep. 2020) 214–231, https://doi.org/10.1016/j.renene.2020.04.133.

[25] S.M.J. Jalali, S. Ahmadian, A. Kavousi-Fard, A. Khosravi, S. Nahavandi, Automated deep CNN-LSTM architecture design for solar irradiance forecasting, IEEE Trans. Syst. Man, Cybern. Syst. 52 (1) (Jan. 2022) 54–65, https://doi.org/10.1109/TSMC.2021.3093519.

[26] P. Kumari, D. Toshniwal, Deep learning models for solar irradiance forecasting: a comprehensive review, J. Clean. Prod. 318 (Oct. 2021) 128566, https://doi.org/10.1016/j.jclepro.2021.128566.

[27] Z. Pang, F. Niu, Z. O'Neill, Solar radiation prediction using recurrent neural network and artificial neural network: a case study with comparisons, Renew. Energy 156 (Aug. 2020) 279–289, https://doi.org/10.1016/j.renene.2020.04.042.

[28] P. Kumari, D. Toshniwal, Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting, Appl. Energy 295 (Aug. 2021) 117061, https://doi.org/10.1016/j.apenergy.2021.117061.

[29] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Adv. Neural Inf. Process. Syst. 4 (2014) 3104–3112. January.

[30] G. C. Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, "A Dual-Stage Attention-Based Recurrent Neural Network for Time Series PredictionDec" doi: https://doi.org/10.48550/arXiv.1704.02971..

[31] J. Bottieau, L. Hubert, Z. De Greve, F. Vallee, J.-F. Toubeau, Very-short-term probabilistic forecasting for a risk-aware participation in the single price imbalance settlement, IEEE Trans. Power Syst. 35 (2) (Mar. 2020) 1218–1230, https://doi.org/10.1109/TPWRS.2019.2940756.

[32] D. Bahdanau, K.H. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. (2015) 1–15.

[33] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 5999–6009, 2017-Decem, no. Nips.

[34] B. Lim, S. Arık, N. Loeff, T. Pfister, Temporal Fusion Transformers for interpretable multi-horizon time series forecasting, Int. J. Forecast. 37 (4) (2021) 1748–1764, https://doi.org/10.1016/j.ijforecast.2021.03.012.

[35] 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, https://doi.org/10.3390/app12010134. Dec.

[36] M. Abdel-Nasser, K. Mahmoud, M. Lehtonen, HIFA: promising Heterogeneous solar irradiance forecasting approach based on Kernel Mapping, IEEE Access 9 (2021) 144906–144915, https://doi.org/10.1109/ACCESS.2021.3122826.

[37] N.Y. Jayalakshmi, et al., Novel multi-time scale deep learning algorithm for solar irradiance forecasting, Energies 14 (9) (Apr. 2021) 2404, https://doi.org/10.3390/en14092404.

[38] M. Abdel-Nasser, K. Mahmoud, M. Lehtonen, Reliable solar irradiance forecasting approach based on Choquet integral and deep LSTMs, IEEE Trans. Ind. Inf. 17 (3) (Mar. 2021) 1873–1881, https://doi.org/10.1109/TII.2020.2996235.

[39] X. Huang, C. Zhang, Q. Li, Y. Tai, B. Gao, J. Shi, A comparison of hour-ahead solar irradiance forecasting models based on LSTM network, Math. Probl Eng. (2020) 1–15, https://doi.org/10.1155/2020/4251517. Aug. 2020.

[40] G. Guariso, G. Nunnari, M. Sangiorgio, Multi-step solar irradiance forecasting and Domain adaptation of deep neural networks, Energies 13 (15) (2020) 3987, https://doi.org/10.3390/en13153987. Aug.

[41] J. Wojtkiewicz, M. Hosseini, R. Gottumukkala, T.L. Chambers, Hour-ahead solar irradiance forecasting using multivariate gated recurrent units, Energies 12 (21) (Oct. 2019) 4055, https://doi.org/10.3390/en12214055.

[42] K. Yan, H. Shen, L. Wang, H. Zhou, M. Xu, Y. Mo, Short-term solar irradiance forecasting based on a hybrid deep learning methodology, Information 11 (1) (Jan. 2020) 32, https://doi.org/10.3390/info11010032.

[43] M. Husein, I.-Y. Chung, Day-ahead solar irradiance forecasting for Microgrids using a long short-term memory recurrent neural network: a deep learning approach, Energies 12 (10) (May 2019), https://doi.org/10.3390/en12101856, 1856.

[44] S. Dev, T. AlSkaif, M. Hossari, R. Godina, A. Louwen, W. van Sark, Solar irradiance forecasting using triple exponential smoothing, in: 2018 International Conference on Smart Energy Systems and Technologies (SEST), Sep. 2018, pp. 1–6, https://doi.org/10.1109/SEST.2018.8495816.

[45] J. Tong, L. Xie, S. Fang, W. Yang, K. Zhang, Hourly solar irradiance forecasting based on encoder–decoder model using series decomposition and dynamic error compensation, Energy Convers. Manag. 270 (Oct. 2022) 116049, https://doi.org/10.1016/j.enconman.2022.116049.

[46] Q. Li, D. Zhang, K. Yan, A solar irradiance forecasting framework based on the CEE-WGAN-LSTM model, Sensors 23 (5) (Mar. 2023) 2799, https://doi.org/10.3390/s23052799.

[47] X. Hou, C. Ju, B. Wang, 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) (Nov. 2023) e21484, https://doi.org/10.1016/j.heliyon.2023.e21484.

[48] M. Munsif, F. U Min Ullah, S. Ullah Khan, N. Khan, S. Wook Baik, "CT-NET: a novel convolutional transformer-based network for short-term solar energy forecasting using climatic information,", Comput. Syst. Sci. Eng. 47 (2) (2023) 1751–1773, https://doi.org/10.32604/csse.2023.038514.

[49] Y. Yang, Z. Tang, Z. Li, J. He, X. Shi, Y. Zhu, Dual-path information fusion and twin attention-driven global modeling for solar irradiance prediction, Sensors 23 (17) (2023) 7469, https://doi.org/10.3390/s23177469. Aug.

[50] X. Kong, X. Du, G. Xue, Z. Xu, Multi-step short-term solar radiation prediction based on empirical mode decomposition and gated recurrent unit optimized via an attention mechanism, Energy 282 (Nov. 2023) 128825, https://doi.org/10.1016/j.energy.2023.128825.

[51] A. Zeng, M. Chen, L. Zhang, Q. Xu, "Are transformers effective for time series forecasting?," [Online]. Available: http://arxiv.org/abs/2205.13504, 2022.

[52] M. López Santos, X. García-Santiago, F. Echevarría Camarero, G. Blázquez Gil, P. Carrasco Ortega, Application of temporal fusion transformer for day-ahead PV power forecasting, Energies 15 (14) (2022) 5232, https://doi.org/10.3390/en15145232. Jul.

[53] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, EMNLP 2013 - 2013 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf. (2013) 1700–1709. October.

[54] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Network. 5 (2) (Mar. 1994) 157–166, https://doi.org/10.1109/72.279181.

[55] K. Cho, et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation, EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf. (2014) 1724–1734, https://doi.org/10.3115/v1/d14-1179.

[56] R.J. Williams, D. Zipser, A learning algorithm for continually running fully recurrent neural networks, Neural Comput. 1 (2) (Jun. 1989) 270–280, https://doi.org/10.1162/neco.1989.1.2.270.

[57] A. Goyal, A. Lamb, Y. Zhang, S. Zhang, A. Courville, Y. Bengio, Professor forcing: a new algorithm for training recurrent networks, in: Advances in Neural Information Processing Systems, 2016, pp. 4608–4616. Nips 2016.

[58] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case," 2020, [Online]. Available: http://arxiv.org/abs/2001.08317..

[59] "NSRDB: National Solar Radiation Database." https://nsrdb.nrel.gov/..

[60] "PyTorch Forecasting Documentation." https://pytorch-forecasting.readthedocs.io/en/stable/index.html.

[61] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Jul. 2019, pp. 2623–2631, https://doi.org/10.1145/3292500.3330701.

[62] T. Zhang, et al., Less is more: fast multivariate time series forecasting with light sampling-oriented MLP structures, Proc. ACM Conf. 1 (1) (2022) [Online]. Available: http://arxiv.org/abs/2207.01186.

[63] P.H. Borghi, O. Zakordonets, J.P. Teixeira, A COVID-19 time series forecasting model based on MLP ANN, Procedia Comput. Sci. 181 (2021) 940–947, https://doi.org/10.1016/j.procs.2021.01.250.