



Article

Improving Solar Radiation Forecasting Utilizing Data Augmentation Model Generative Adversarial Networks with Convolutional Support Vector Machine (GAN-CSVR)

Abbas Mohammed Assaf^{1,2,*} , Habibollah Haron^{1,*}, Haza Nuzly Abdull Hamed¹, Fuad A. Ghaleb¹ , Mhassen Elnour Dalam³ and Taiseer Abdalla Elfadil Eisa⁴

- ¹ Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia; haza@utm.my (H.N.A.H.); abdulgaleel@utm.my (F.A.G.)
² State Company of North Distribution Electricity, Ministry of Electricity, Mosul 10013, Iraq
³ Department of Mathematics-Girls Section, King Khalid University, Mahayil 62529, Saudi Arabia; mdalam01@kku.edu.sa
⁴ Department of Information Systems-Girls Section, King Khalid University, Mahayil 62529, Saudi Arabia; teisa@kku.edu.sa
* Correspondence: assaf.abbas@graduate.utm.my (A.M.A.); habib@utm.my (H.H.)

Abstract: The accuracy of solar radiation forecasting depends greatly on the quantity and quality of input data. Although deep learning techniques have robust performance, especially when dealing with temporal and spatial features, they are not sufficient because they do not have enough data for training. Therefore, extending a similar climate dataset using an augmentation process will help overcome the issue. This paper proposed a generative adversarial network model with convolutional support vector regression, which is named (GAN-CSVR) that combines a GAN, convolutional neural network, and SVR to augment training data. The proposed model is trained utilizing the Multi-Objective loss function, which combines the mean squared error and binary cross-entropy. The original solar radiation dataset used in the testing is derived from three locations, and the results are evaluated using two scales, namely standard deviation (STD) and cumulative distribution function (CDF). The STD and the average error value of the CDF between the original dataset and the augmented dataset for these three locations are 0.0208, 0.1603, 0.9393, and 7.443981, 4.968554, and 1.495882, respectively. These values show very significant similarity in these two datasets for all locations. The forecasting accuracy findings show that the GAN-CSVR model produced augmented datasets that improved forecasting from 31.77% to 49.86% with respect to RMSE and MAE over the original datasets. This study revealed that the augmented dataset produced by the GAN-CSVR model is reliable because it provides sufficient data for training deep networks.

Keywords: convolutional neural network; data augmentation; generative adversarial network; loss function; support vector regression; solar radiation data



Citation: Assaf, A.M.; Haron, H.; Abdull Hamed, H.N.; Ghaleb, F.A.; Dalam, M.E.; Elfadil Eisa, T.A. Improving Solar Radiation Forecasting Utilizing Data Augmentation Model Generative Adversarial Networks with Convolutional Support Vector Machine (GAN-CSVR). *Appl. Sci.* **2023**, *13*, 12768. <https://doi.org/10.3390/app132312768>

Academic Editors: Seung-Hoon Yoo and Luisa F. Cabeza

Received: 18 July 2023

Revised: 10 November 2023

Accepted: 14 November 2023

Published: 28 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Many photovoltaic power (PV) installations have several problems in forecasting because they do not have sufficient historical data. Insufficient training data makes forecasting a significant challenge, and to overcome this challenge, it is vital to develop a trustworthy and dependable generative model capable of simulating actual historical records of various weather conditions [1]. According to [2,3], lowering the amount of training data would surely degrade the performance of the forecasting model, and its performance would be poor with insufficient training data because it imposes a bias in favor of the majority class. Small datasets may not provide suitable or adequate representative training data, resulting in model overfitting. Consequently, running tests on the models produces inaccurate results [4]. Deep learning forecasting models alone will not provide reliable PV power

estimates; researchers must also consider the availability of solar radiation data. As a result, they are becoming more interested in studies that gather weather variables and the availability of photovoltaic forecasting models to effectively estimate the volatility and uncertainty of photovoltaic power output owing to varied and changing climatic conditions [5–8].

Accurate and reliable data are essential for training and validating the models to provide accurate forecasting. Missing data in the historical dataset might lead to incomplete training, resulting in lower forecasting accuracy. Erroneous data, such as faulty sensor readings or outliers, can alter the model's grasp of patterns and correlations, resulting in false forecasting. PV power forecasting models consider various input features such as solar irradiance, temperature, cloud cover, and humidity. If any of these features have missing or incorrect data, the model may be unable to effectively represent the underlying relationship between the inputs and the PV power output. This missing data can lead to biased or untrustworthy forecasts. For instance, gaps in time series analysis due to missing data might reduce the precision of identifying trends and detecting seasonality. It is possible to impute or replace missing data using the right procedures, but it might be difficult to do so accurately if there are many missing data points [9]. Therefore, data augmentation has indeed been required when implementing deep learning techniques with a restricted training dataset.

Data augmentation is a technique for boosting deep learning performance, particularly deep networks. It produces new data generated from present data or new copies of old data that have been altered to make more training data available. Artistically, data augmentation contributes a kind of disturbance or noise to the datasets, which are both considered undesirable features in statistical modeling and must be removed using filters [10,11]. The impact of data augmentation in deep learning is to regularize the model and reduce overfitting throughout deep training, which increases the generalizability and prevalence of the learned models. The enhancement in performance occurs when a model is tested using data from the training set compared to data from the testing set that the model has never been exposed to before [12].

Numerous advanced generative models were introduced in previous publications; for instance, Random Oversampling (RO), a basic generative model, generates new data by replicating random minority class specimens [3]. Although the RO process is straightforward, its perfect reproduction of training instances might lead to overfitting because the model is repeatedly exposed to the same data. A different strategy, Synthetic Minority Oversampling Technology (SMOTE) [13], tries to avoid overfitting by generating artificial data connecting specimens of minority classes. Moreover, SMOTE may produce noisy samples because the distinction between minority and majority category clusters is not always apparent [14]. There have been some previous efforts to build generative models using an agent-based or Markov chain approach, but these methods have strict requirements for the data to satisfy; for instance, the first-order Markov property [15]. Auto-encoders are generative models consisting of an encoder and a decoder. They are frequently used to reconstruct images in image processing as close to the original [16]. Using an auto-encoder may result in the same issue as RO since the recently founded data only retain the distribution of specimens from the pristine data and do not conserve the variety of samples produced.

A new robust generative model known as the “generative adversarial network” was recently developed by [17]; it is an excellent resolution to solve difficulties in the preceding techniques and is a well-known technique for data augmentation that creates new data by examining the distribution of the existing data. Additionally, the size or number of training data categories is not a restriction for GAN-based data augmentation. It is worth noting that although typical GAN models prioritize data production with adequate variability, they also prioritize consistency [18]. Since then, numerous GAN variants have been proposed for use in diverse situations. While GANs were first developed for picture generation, they are now used to generate time series data. Adversarial networks can now process intricate time-series data patterns. Since their inception, GANs have been used in many different

study areas, from speech and language processing to image and vision computing and beyond [19]. The capabilities of excellent GAN modeling allow it to generate new data with a distribution identical to pristine data while preserving the diversification of produced data to create new specimens with a distribution similar to the pristine information while preserving diversification of produced data [19]. Furthermore, modeling complex data with certain implicit distributions does not require the data to conform to any pre-existing assumptions [15]. Convolutional neural networks and generative adversarial networks have been proposed in [1] to classify the weather into ten categories. Generative adversarial networks augment the training dataset for all weather categories. The simulation results show that the generative adversarial networks have the ability to generate high-quality data that capture the underlying features of the original data. The data augmentation based on convolutional generative adversarial networks optimizes whole categorization models. Thus, a convolutional GAN is ideal for data generation and augments the training dataset, which is the basis for our study. In this study, the core of the generative model is understanding the distributions of previous solar radiation data under various weather circumstances to generate sufficient training data for the forecasting models; this will also be explained in Section 2.5. On the other hand, researchers have rarely used convolutional GANs to create time series solar radiation data. Indeed, there is a significant necessity for sufficient training data for PV power models, which might come from either an actual dataset or augmentations.

Short-term solar radiation estimates have been shown to enhance yearly energy consumption for commercial operations in building microgrids. Sudden changes in solar radiation, known as ramp events, hold particular relevance for short-term forecast horizons, which can be captured with greater precision. Changes in solar radiation that happen quickly and are very strong could make PV power become less reliable and reduce in quality. As a result, the results of short-term forecasting may be utilized to determine the I confirm the meaning is retained most effective PV power ramp rates [20]; therefore, the forecasting period in this study was set at 30 min in advance.

Forecasting solar radiation facilitates the integration of PV power units into the electrical grid, scheduling energy storage systems, and optimizing energy transmission, thus reducing energy loss [21]. By anticipating PV power generation, it reduces reserve capacity and generation costs, thus preventing electrical energy system disruption [22]. Throughout the years, solar radiation forecasting has gained much interest from academics and businesses. Recently, machine learning methods have been used to investigate the problem of estimating solar radiation forecasting. Among these approaches are the boosted decision tree regression (BDTR) model, which was used to predict solar radiation based on data gathered in Malaysia [23], a multi-objective shark algorithm, and a fuzzy method for forecasting solar radiation [24]. In [25], predicted solar radiation also has used artificial neural network (ANN) and random forest (RF) models. However, deep learning techniques are rapidly developing, particularly in unsupervised feature extraction. Deep learning models are crucial for estimating solar radiation, especially when dealing with a complex problem with a huge amount of data. In contrast to machine learning models, their efficiency remains constant as the amount of input data increases. Technological advancements in data collection and generation have enabled meteorological stations and photovoltaic power plants to collect massive amounts of data samples [26]. While some machine learning algorithms cannot manage high-dimensional inputs or huge datasets, others cannot handle them. The hybrid model focuses on combining standalone methods to exploit the benefits of individual forecasting models, which enhance the accuracy of deep learning networks [27]. For example, ref. [28] proposed a hybrid model consisting of a long short-term memory neural network (LSTM), multi-layer perceptron model (MLP), convolutional neural network (CNN), Gaussian process model (GPR), and graph convolutional network (GCN) to forecast ozone concentration. Ref. [29] also introduced a hybrid deep learning model that consists of a support vector machine (SVM), Gaussian regression process (GPR), and convolutional neural network (CONN) for rainfall forecasting. What distinguishes this study

from previous studies and motivates us to investigate the generation of solar radiation data and its effect on the forecasting accuracy of PV power is that most previous studies focus on developing forecasting models utilizing small datasets, which suffer from inaccuracy, and small datasets do not provide enough specimens to train deep neural networks. To the best of our knowledge, no research has been carried out that considers the generation of solar radiation data that may be trained to improve the performance accuracy of forecasting models. Since the proposed method focuses on generating solar radiation data to improve forecasting performance, this work also paves the way for future studies in a wide array of areas by expanding and embracing the modern technique proposed to augment data, whereas most studies of photovoltaic power forecasting, wind power forecasting, and energy consumption concentrate on the short-term forecast, disregarding medium-term and long-term forecasting because of a lack of training data for forecasting models.

The following are the significant contributions of our current work:

1. Convolutional GAN (Conv-GAN) that combines GAN and CNN will be enhanced by replacing the fully connected CNN layer with a more superficial linear SVR layer; this linear SVR assists in restricting the deviation of generated specimens, is robust in discarding outliers, and has excellent generalization capability; and the model is trained using a Multi-Objective loss function that combines Mean Square Error (MSE) and Binary Cross Entropy (BCE). The MSE loss function was used to determine how similar the produced samples were to the original samples, and the BCE loss function was used to stabilize the training process and confirm that the generated samples were structurally fairly similar to the training data, which led to obtaining data which were identical to the original data. This model has indeed been trained to create meteorological data that include both spatial and temporal data, which lead to better forecasting.
2. The new augmented solar radiation dataset via the GAN-CSVR model is evaluated by two effective indices: the standard deviation (STD) and the cumulative distribution function (CDF).
3. To validate the impact of augmented data on the accuracy of forecasting models, solar radiation forecasting is rigorously evaluated on the original datasets and augmented datasets of three different locations.

The materials and methods are described in Section 2. In Section 3, the experimental design is presented. The evaluation and discussion are presented in Section 4. Section 5 gives the conclusion.

2. Materials and Methods

To build meaningful deep learning models, testing and training mistakes must be continually decreased, which happens due to the scarcity of training datasets, particularly for extreme weather events. Data augmentation is a remarkably sturdy strategy for achieving this goal. The data augmentation technique rotates and tunes the data without altering the temporal sequence of the pristine data and ensures that the generated data and preceding data are statistically uniform. The training samples are supplemented in this study and their quantity is increased to two times the initial sample size.

2.1. Generative Adversarial Networks

The GAN was invented by [17], and its other version, the convolutional GAN (Conv-GAN), which was developed by [30], demonstrates significant potential in producing pictures that are incredibly similar to a given collection of training images. The GAN is a robust category of generative models that perform their task by implicitly modeling high-dimensional data distributions [31]. In image processing, the GAN has shown superiority over other generating approaches in its capacity to create realistic synthetic pictures [16,30,32]. In terms of applicability in our study endeavor, the GAN is utilized to understand the distribution of solar radiation data and augment it.

The training technique in the GAN algorithm is to build two antagonistic neural networks that compete with each other, which are typically the generator (G) and the discriminator (D), which may be trained using a traditional backpropagation approach.

The two networks collaborate to enhance each other but in an aggressive manner. During the training phase, G attempts to learn and make “fake” samples of input noise Z to trick D, while D endeavors to identify “fake” or “genuine” inputs accurately. The model converges until the discriminator no longer distinguishes between the samples. It is worth noting that the input noise Z typically keeps track of a Gaussian distribution across G to provide the sample $I = G(z)$, where $Z = N(\mu, \sigma^2)$, and D is a fundamental neural network filter for bilateral categorization. This work gives a fantastic possibility for supplementing training samples with a wide range of specimens to increase the generalization of deep learning algorithms.

2.2. Convolutional Neural Networks

The design of the CNN algorithm contains feature extraction and classification, which make it possible for the CNN to learn how to improve features based on the fundamental data it receives while being trained. The neurons in the CNN are linked with the layer before it, the filter weights are shared, and they can react more efficiently with massive datasets. The output of the convolutional layer may be computed as shown in Equation (1) [33]:

$$y^k \cdot i_k = f \cdot \left((w^k \cdot *h) i_j + b_k \right) \quad (1)$$

where f refers to an activation function, $*$ represents a convolutional process operator, and w^k means the weight of the kernel.

In general, the CNN model has two steps: The initial step of the architecture involves the convolutional layer and the pooling layer. In contrast, the final step comprises the fully connected layers. Although the CNN algorithm is frequently employed for image distinction, one-dimensional CNN models for forecasting and classification tasks using time series have just recently been proposed. Another noteworthy aspect of the 1D CNN is that it can be implemented effectively and inexpensively, owing to its straightforward and compact architecture, which performs one-dimensional convolutions [33]. A typical CNN consists of two merged layers: the fully connected layer and the feature extraction layer [34]. The convolutional layer (feature extraction layer) follows the input layer in the structure and comprises two types of layers: convolution layers and pooling layers [35]. The activation maps for the filters are generated to help the convolutional layer, which employs a set of filters to convolute over the data. In every filter, neurons are directly linked to the data points that are being input, which results in multiplying the weights by the data points. Sharing the weights of the neurons included inside one filter helps reduce the amount of time and complexity required for the CNN’s optimization.

The pooling layer is intended to reduce the overall magnitude of the matrix. The two types of pooling layers are called maximum pooling and average pooling, described below. Max pooling is a method of reducing the magnitude of a matrix by selecting the most significant value included in it. Average pooling is a method of reducing the volume of an array by specifying the median value included in it [36]. Both are utilized in the methods described in Equation (2).

$$f(x) = \max(0, x) \quad (2)$$

2.3. Support Vector Regression

The SVR was created based on the statistical theory of learning, and it has been widely applied to tackle challenges related to high-dimensional regression. It performs effectively with limited computational resources and few training samples [37,38]. The core idea behind the SVR model is to transform the input space of pristine data points into a higher-dimensional or infinite-dimensional space feature in which an ideal separating hyperplane has been constructed and the distance between the constructed hyperplane and all data

points is the shortest possible distance [39]. The mathematical formula for the SVR is shown in Equation (3).

$$f(x) = \phi(x) \times W^T + b \quad (3)$$

where $f(x)$ symbolizes the generated feature, W represents the weight of the feature, and b is the bias; $\phi(x)$ represents a mapping of the input space to the high-dimensional feature space as well as the structure of linear regression for new features [39].

2.4. Loss Function

Two loss functions, which are the Mean Square Error (MSE) and Binary Cross Entropy (BCE), have been used. The MSE loss function computes the similarity between the produced and original specimens. At the same time, the MSE loss function learns to decrease the mistake by calculating the difference between the original and produced values using an Adam optimization technique. The MSE loss is defined as the squared L2-norm, often known as the least squares error (LSE). It reduces the sum of the squares of the variations between the genuine and synthetic specimens in meteorological data. The MSE loss function is defined in Equation (4) [40]:

$$\text{Loss}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (4)$$

where y_i and \hat{y}_i are real and augmented data. The MSE loss is used to pinpoint the Euclidean distance between the target vector and its assessment in the complex domain, which helps the model generate samples that are structurally reasonably similar to the training data.

The BCE loss function is used to stabilize the training process and verify that the augmented specimens have the appropriate qualities in terms of their overall structure. Equation (5) [41] shows the BCE loss function.

$$\text{Loss}_{\text{BCE}} = - \sum_{k=1}^c y_k \log(f(s)_k) \quad (5)$$

where y_k and s_k are real data and augmented data for every feature (k) in solar radiation data (c).

During the training process, the model's parameters are adjusted iteratively to minimize the value of the loss function until the convergence. This loss function convergence is a steady state, which means that the model has learned to minimize the loss and has reached a point where further updates to the parameters of the model do not significantly reduce the loss function value anymore.

Multi-Objective loss: It is a loss function that incorporates multiple objectives by combining the individual components. It combines two loss functions, (MSE) and (BCE). The combined loss function can be defined as

$$\text{Combined Loss} = \alpha * \text{MSE} + \beta * \text{BCE}$$

where both α and β are set to 0.5, and the combined loss function equally weighs the MSE and BCE, simultaneously optimizing both objectives.

2.5. The Proposed GAN-CSVR Model

This Section describes the GAN-CSVR model, which consists of three steps: partition of data for training and validation, augmenting data utilizing the GAN-CSVR algorithm, and validation of the augmented data. Figure 1 shows the three steps and the details of each step for the GAN-CSVR model. The following subsections describe the GAN-CSVR model, especially how it is built, trained, and validated.

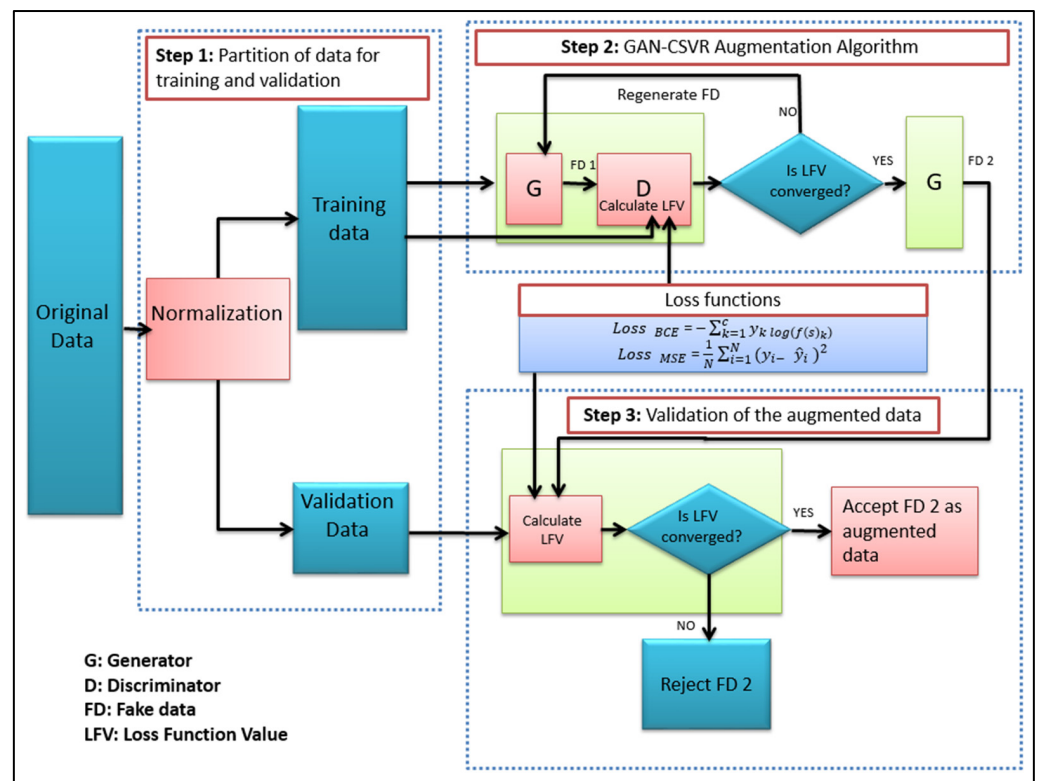


Figure 1. The structure of the proposed GAN-CSVR model.

2.5.1. Step 1: Partition of Data for Training and Validation

In this step, the original dataset has been partitioned into a training set and validation; the validation set is used to evaluate the augmented data. The original data split is 80% for training and 20% for validation. This step is crucial for maintaining the representative nature of the data. The Min-Max scalar normalization has been used to normalize all input solar irradiance data to the scope [0, 1] to reduce data dispersion during the training phase. It is beneficial to be utilized as it retains the distribution pattern of the original data and the information embedded in it remains unchanged after conversion. Equation (6) was used to normalize the data [42]:

$$x_n = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

where x_n , x_i , x_{\max} , and x_{\min} are the normalized, original, maximum, and minimum values of the time series input, respectively.

2.5.2. Step 2: Augmenting Data Utilizing GAN-CSVR Algorithm

In this step, a new algorithm is proposed that combines a GAN, CNN, and SVR, which is called the GAN-CSVR algorithm, to augment data. As shown in Figure 1, the GAN consists of two neural networks competing with each other; the generator (G) and discriminator (D) have been trained using a backpropagation technique. These two networks work together aggressively to improve each other. During the training phase, G tries to learn and create “fake” samples of input in order to fool D, while D attempts to accurately identify “fake” or “genuine” inputs. The model converges until the discriminator can no longer detect the difference between the samples. The convolutional GAN (Conv-GAN) [30] has successfully been utilized to create specimens for data augmentation, which is one major mission for the GAN, and serves as the foundation of this model. However, when applied to solar irradiance data generation, it often suffers from several issues, such as the loss of critical discriminative features and big variations between generated specimens and input specimens, which might not help train a deeper network. Consequently, the traditional Conv-GAN is enhanced by removing the fully connected CNN layer and substituting it

with a basic linear layer SVR to decrease the number of features that need to be learned without a significant number of training examples to produce the output specimens. This kind of linear layer SVR is effective for limiting the variance of the samples produced, is robust in discarding outliers, and has excellent generalization capability. The SVR is utilized in this study as the ultimate layer of the GAN-CSVR model because of its own recent superiority in solar irradiance forecasting and energy demand problems [43–47].

The combination of the GAN, CNN, and SVR in the hybrid model constructed to augment the dataset allows the CNN to capture domestic pattern features and popular properties that repeat in the time series at various periods. Thus, the proposed GAN-CSVR model can produce data precisely by extracting features from meteorological data that influence the production of specimens. The generator and the discriminator use a neural network with two convolutional layers (conv1D) together with two Max pooling layers, one convolutional layer (conv1D) with RELU activation function, a single flatten layer, and SVR as the last layer, as shown in Figure 2. The 1D convolutional layers start learning the input time series data to best grasp the dependence between features to create a symmetric feature map and increase the number of rows of the input twofold by a simple linear operation simultaneously. After these three 1D convolutional layers, the output is filtered by an SVR layer that linearly separates the patterns of the original input features, which can augment a dataset to fit as many features as possible.

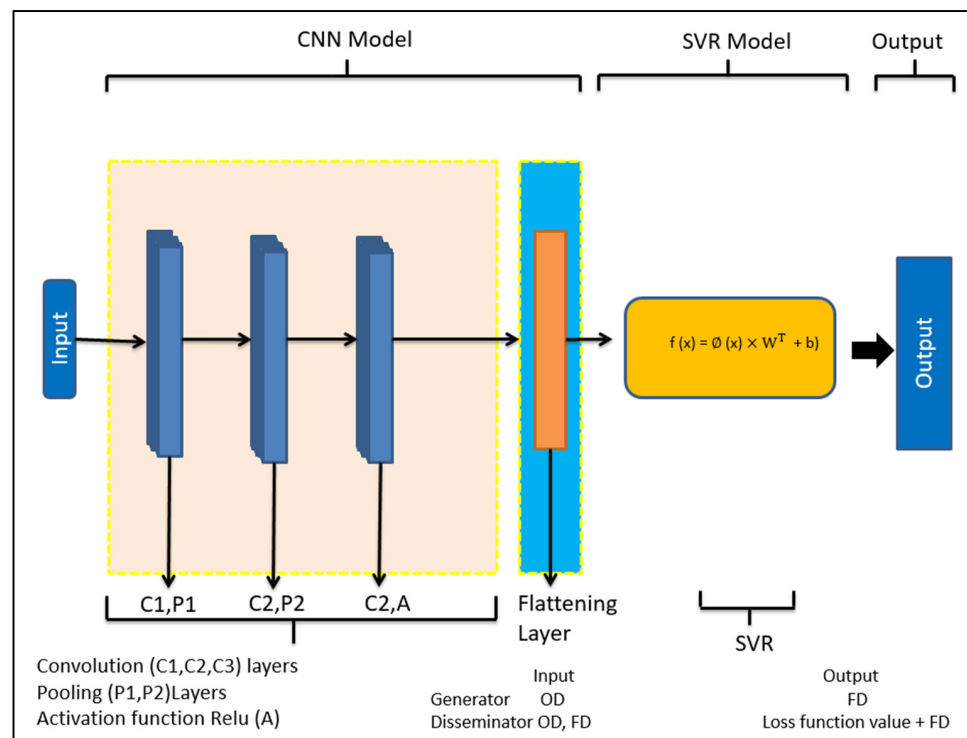


Figure 2. Components of generator and discriminator.

The discriminator is updated using training data and augmented data, while the generator is updated using the output of the discriminator as feedback. This iterative process allows the generator to improve its ability to produce realistic data. The generator is compiled with the optimizer, while the discriminator is compiled with the optimizer and the Multi-Objective loss function to eliminate a gradient vanishing problem in the generator. The loss function convergence helps the discriminator to reach a stable state where it can effectively distinguish between real and fake samples, which leads to producing extremely identical samples that have the diversity representing the underlying properties of the original data. Adam is used as an optimizer for the generator and discriminator. The pseudocode of GAN-CSVR is shown in Figure 3.

| Algorithm: GAN-CSVR |
|---|
| Input: Original-Dataset Output: Augmented-Dataset Start // Step 1: Partition, Initialize, and Normalize Original Data 1.1 Normalize the dataset using min_max scaler . 1.2 Use the dataset and split it to train and validation sets with different sizes for each set. // Step 2: Define the architecture of the Generator and Discriminator Functions. 2.1 Let G () is the define function of Generator. 2.2 Create 1D-CNN component to the Generator. 2.3 Create SVR component to the Generator. 2.4 Let D() is the define function of Discriminator. 2.5 Create 1D-CNN component to the Discriminator 2.6 Create SVR component to the Discriminator. // Step 3: Initialize the input and output for the Generator and Discriminator. 3.1 Define GAN-CSVR_input which is the Original_dataset. 3.2 Define Generator_input \leftarrow Original Data 3.3 Define Discriminator_input \leftarrow Fake_Input, Original_Data 3.4 Define Fake_Input of Discriminator \leftarrow Output of Generator. 3.5 Define Discriminator_Output \leftarrow (Fake_Output). 3.6 Define GAN-CSVR_Output is the output of the Discriminator. 3.7 Define New Fake_Input \leftarrow Discriminator (Fake_Output). // Step 4: Training the Discriminator with the multi-objective loss function 4.1 Compile the discriminator: 4.2 Compile discriminator with multi-objective loss function. 4.3 Freeze the weights of the discriminator while training the generator. // Step 5: Training GAN-CSVR model with the multi-objective loss function to validate augmented data. 5.1 Compile GAN-CSVR model with multi-objective loss 5.2 Define GAN-CSVR_Input \leftarrow "Validation data input, New Fake_output" 5.3 Define GAN-CSVR_Output \leftarrow augmented data. 5.4 Finally, Train the GAN-CSVR in 100 epochs . Finish |

Figure 3. The pseudocode of GAN-CSVR.

Consequently, the GAN-CSVR model seeks to maintain representational characteristics without overfitting. The enhanced GAN-CSVR with minimal training and a simple network design can augment specimens which are fundamentally compatible with training data.

The process of tuning hyperparameters represents a challenge in this study, although the grid search is used to adjust them, which took 1958 h. However, after the appropriate settings were determined, the training and testing time was significantly decreased. Table 1 shows the optimal parameters of the GAN-CSVR model.

GAN is a type of deep learning model that can be computationally expensive, requiring significant computational resources for training. Nevertheless, gathering solar radiation data can be expensive and difficult because of erratic weather patterns and lost data from damaged sensors. This training process needs extensive computational resources to efficiently analyze massive quantities of data and execute computations for forward and backward propagation and data evaluation by the loss function to generate new augmented data. It is important to highlight that the computational processes and resource requirements of deep learning models are mostly determined during the training phase; once the model has been trained, generating data samples is often computationally easy

and quick. In addition, an early stopping criterion was utilized, which automatically stops training if the loss of validation data does not enhance after ten iterations. It was performed to avoid overfitting the training data and save time and computational resources.

Table 1. The optimal parameter used in the GAN-CSVR model.

| GAN-CSVR | Layer 1 CNN | Layer 2 CNN | Layer 3 CNN | Layer 4 SVR | Optimizer |
|---------------|---|--|--|---|---|
| Generator | Convolutional layer filters = 64 Kernel size = 3 Max pooling layer RELU | Convolutional layer filters = 128 Kernel size = 2 Max pooling layer RELU | Convolutional layer RELU filters = 64 Kernel size = 3 | Noise dimension = 100 Kernel type is RBF Regularization Parameter (C) = 100 epsilon = 0.01 Tolerance (tol) = 1×10^{-4} . | Adam |
| Discriminator | Convolutional layer filters = 64 Kernel size = 3 Max pooling layer RELU | Convolutional layer filters = 128 Kernel size = 2 Max pooling layer RELU | Convolutional layer RELU filters = 64 Kernel size = 3 | noise dimension = 100. Kernel type is RBF Regularization Parameter (C) = 100, epsilon = 0.01 Tolerance (tol) = 1×10^{-5} . | Adam Multi-Objective loss function MSE and BCE |

In this study, all tests were performed on a laptop computer (MacBook Pro) equipped with a 64-bit operating system, 16 GB of LPDDR3 memory, an Intel Core i7 quad-core processor, and Intel HD graphics 530 (1536 MB of graphics). The proposed code was implemented in Python 3.7 using the open-source Tensorflow, Keras, and sklearn libraries. The weights in the neural networks were initialized according to the Keras settings.

2.5.3. Step 3: Validation of the Augmented Data

In the third step, the validation data is used to evaluate the performance of the model during training based on the loss function values. The validation of the performance for the GAN-CSVR model by using the Multi-Objective loss function, which combines MSE and BCE loss functions to assess structural differences and determine the similarity between augmented and validation data, has been applied. The validation process primarily focuses on evaluating the generated samples by the GAN-CSVR algorithm and assessing how well the generator has learned to capture and replicate the characteristics of the actual data distribution. The threshold value for the training and validation loss function with a discriminator is highly dependent on the distribution of data and the desired quality of the augmented data, initially starting with a threshold value between 0.1 and 0.5 and then adjusted based on the evaluation of the augmented data. The standard deviation and cumulative distribution functions have been used to assess the quality and fidelity of the generated samples, in addition to visual inspection, as explained in Section 4.1 in detail.

The training and validation loss function values for the augmented and validation data have been compared. If the training loss function values are significantly lower than the validation loss function value it may indicate overfitting, where the model fits too closely to the augmented samples but does not generalize well to validation data. If the training loss function values closely match the validation loss values, augmented data are representative of the actual data distribution, and the Multi-Objective loss function is beneficial for improving generalization performance, and the augmented data will be accepted. The threshold between the training loss function values and the validation loss function values was determined as 0.2 based on the evaluation of augmented data, which achieved statistical similarity and pattern diversity for the augmented datasets. Finally, the GAN-CSVR model produces augmented samples that are fundamentally compatible with validation data and maintain representational characteristics without overfitting.

3. Experimental Design

3.1. Datasets

The National Renewable Energy Laboratory (NREL) offered standard datasets for three major locations in California: San Diego, San Francisco, and Los Angeles, which are used in the current study, which used solar radiation and other meteorological variables from January 2015 to December 2019 [48].

Various climatic factors, including pressure, temperature, wind speed, relative humidity, and others, significantly influence solar radiation availability. These factors also interact complicatedly with one another. Depending on the parameters utilized, these climatic variables affect the solar radiation forecasting model in a variety of ways [49]. Table 2 provides a description and analysis of solar radiation and meteorological data, which has been augmented, allowing for visualization of the measured values over time. The datasets utilized in this study to construct the solar radiation forecasting models are time-series data, including spatial and temporal features. The comparable or constant data over time as a result of seasonal and climatic impacts is a temporal feature in solar radiation time-series data [50]. The solar radiation data current value at the time of the forecast must be strongly linked to the solar radiation data value from the previous hours. Therefore, the data sampling time in this study is 30 min to obtain good matching between the modeled and measured data. When the latest 12 h of solar radiation data were fed into the forecasting models, the suggested models exhibited the best forecasting accuracy.

Table 2. The solar radiation data.

| Variables | Unit | Description | Example Value |
|--------------------------------|------------------|--|---------------|
| Date | Day | Data is five years, month, day, | 1 May 2023 |
| Time | Minute | Half an hour. | 00:30 |
| Globule solar Irradiance (GHI) | W/m ² | GHI refers to measurements of the solar radiation received from the Sun at a particular location on Earth. | 167 |
| Clear sky GHI | W/m ² | The amount of solar radiation that would be received on a horizontal surface. | 258 |
| Dew point | °C | The Dew point indicates the moisture content in the air. | 5 |
| Solar Zenith Angle | Degree | The Solar Zenith Angle depends on the latitude, time of day, and time of year. | 78.8 |
| Wind direction | Degree | Indicates the compass direction from which the wind is blowing, such as north, south, east, or west. | 3.7 |
| Wind speed | m/s | Wind speed represents the magnitude of wind flow. | 286.5 |
| Relative Humidity | % | It provides information about the moisture content in the atmosphere. | 62.23 |
| Temperature | °C | Refers to the ambient air temperature at a specific location and time. | 19 |
| Pressure | Bar | Pressure is the force exerted by the air above a specific location. | 1020 |

The performance of the GAN-CSVR model is affected by the size of the training data. When the training data size is reasonably large, it provides more diverse examples to learn from, which produces samples with diversity and distribution similar to the basic distribution of the training data, leading to better generalization. In contrast, the data's small size provides less training information, which results in similar samples that lead to overfitting and model collapse.

The data are collected at regular intervals every half an hour, serving as the original dataset for the case study. The sample sizes of the training set and evaluation set are 43,834, 78,515, and 52,688 for the three locations, respectively. In this study, the training samples

are supplemented and the sample size is doubled to improve the performance of deep learning models.

Pre-processing input data can significantly lower computing costs and improve the accuracy of the model. We used a few pre-processing methods in this study. The data are erased in the early morning and late at night, and missing data are added. The dataset contains variables with varying scales, which may affect the accuracy of the proposed model. As a result, the Min-Max scalar normalization technique is used to normalize the dataset variables.

One approach is to assume that the missing data for a particular day are similar to another day with the same data simultaneously. This assumption relies on the concept of temporal continuity, where adjacent days or periods may exhibit similar solar radiation patterns. Using this approach, we generate values instead of missing ones using our GAN-CSVR model. These values match the values of a similar nearby day with the same data at the same time in most, but there is a little difference overall. This approach leverages the capabilities of the GAN-CSVR model to retain the maximum pattern diversity from the original dataset samples. The model captures the temporal continuity and generalizes the patterns to generate plausible values for the missing data points. This ensures a complete dataset for training, evaluation, and maintaining the data time series for the forecasting model.

3.2. Performance Evaluation Metric

This study uses three performance metrics to measure the accuracy of forecasts to see how well the augmented data perform in forecasting compared to the original data.

- The Mean Absolute Error (MAE): It demonstrates the median of the absolute errors among the actual solar radiation values and the anticipated values, as shown in Equation (7):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^n |G_{a_i} - G_{F_i}| \quad (7)$$

- The Root Mean Square Error (RMSE): It is calculated by finding the quadratic root of the median of the quadratic variances that exist between the values measured and those forecasted for the solar irradiance. This calculation is shown in Equation (8):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^n (G_{a_i} - G_{F_i})^2} \quad (8)$$

- The Correlation Coefficient (R): It indicates the strength of the linear relation between the actual and forecast solar radiation and is computed as in Equation (9):

$$R = \frac{\sum_{i=1}^n (G_{F_i} - \overline{G_{F_i}})(G_{a_i} - \overline{G_{a_i}})}{\sqrt{\sum_{i=1}^n (G_{F_i} - \overline{G_{F_i}})^2 \sum_{i=1}^n (G_{a_i} - \overline{G_{a_i}})^2}} \quad (9)$$

where n represents the entire numeral of data points, G_{F_i} and G_{a_i} are the forecasted and actual values, respectively. $\overline{G_{a_i}}$ and $\overline{G_{F_i}}$ represent the mean of the actual and forecast values, respectively.

Evaluation metrics provide feedback on forecasting accuracy against the benchmark models, allowing models to be fine-tuned to achieve a target degree of precision. However, without metrics for comparison it is impossible to describe the performance of deep learning models. A few performance evaluation metrics, such as Mean Absolute Error (MAE), Correlation Coefficient R , and Root Mean Square Error (RMSE), are frequently used to assess the accuracy of forecasting models. These metrics can be effectively utilized in the majority of models, including those for rainfall, solar irradiance, wind power, etc. [51].

Given the debate regarding evaluation metrics' efficacy, R , MAE, and RMSE are adequate indicators for determining the optimum forecasting model [52,53]. It is also worth noting that R , MAE, and RMSE are different evaluation metrics reflecting various aspects

of model performance. R is a measure of the correlation, whereas MAE and RMSE are measures of forecasting error. The results have been evaluated using statistical metrics that are the most popular ones: MAE, R, and RMSE. This last one is still the most commonly used metric in forecasting models [54–57]. In addition, RMSE is considered in the literature as a good measure to evaluate and compare forecasting models; therefore, the RMSE metric has also been applied to provide a more significant representation of the results, which would help to make more reliable future studies.

4. Evaluation and Discussion

An evaluation of the quality of the generated data by the GAN-CSVR model is presented in Section 4.1. Furthermore, in Section 4.2, a thorough assessment of the performance of modeling tests for the proposed forecasting models utilizing augmented data vs. original data was conducted to demonstrate the effect of augmented data on the accuracy of different forecasting models.

4.1. Evaluation Index of the Quality of the Generated Data

This Section elaborates on an in-depth examination of the quality of the augmented data by the GAN-CSVR model. The GAN-CSVR aims to generate novel and distinct samples that capture the inherent properties of the original data; thus, evaluations using two indices, namely STD and CDF, have been performed. The first index evaluates the specimen pattern variety. The second index is used to determine the statistical similarity of newly augmented samples.

4.1.1. Standard Deviation (STD)

The STD, denoted by the Greek letter (σ), is a statistical metric that measures data value dispersion or difference [58]. A lower STD score means that the data points are close to the median (also known as the anticipated values) of the collection, whereas a higher STD score shows that data points are distributed throughout a broader scope of real values. Equation (10) describes the computation of STD.

$$\sigma_t = \sqrt{\frac{1}{N} \sum_{i=1}^{N_t} (y_{ti} - \mu_t)^2} \quad (t = 1, 2, \dots, 60) \quad (10)$$

where μ_t and σ_t are the anticipated value and the STD of a collection of solar radiation data y_{ti} at a given period t , and N_t is the overall specimen volume of that set.

This statistical measure examines in practice the statistical similarity between the original and augmented solar radiation data curves based on the distribution of the solar radiation data curve to assess the patterns and variety of produced specimens, such as vectors consisting of 60 medium values for solar radiation data that are identical at specific time points ($n = 60$), based on [1]. Regarding a collection of solar radiation data produced at a specific period, a higher STD could refer to the larger dispersion grade of the collection, which validates the solar radiation data variety at the same time point. However, a low STD score means that the generative model can only remember the training data and not provide diversity.

Figure 4 compares the produced and actual curves for all three locations (i.e., the dataset that was created depending on GAN-CSVR and the original solar radiation dataset). The blue curves match the orange curves closely in most, but there is a little difference overall. Moreover, the visualization curves of other meteorological data are illustrated in the Supplementary File. The samples generated by the GAN-CSVR model can sometimes cover a broader range of values than the actual samples. This fact shows that the GAN-CSVR model can retain the maximum pattern diversity for the original dataset samples. Thus, the GAN-CSVR model can generate new specimens not included in the original training data but that adhere to a similar statistical distribution as the authentic specimens.

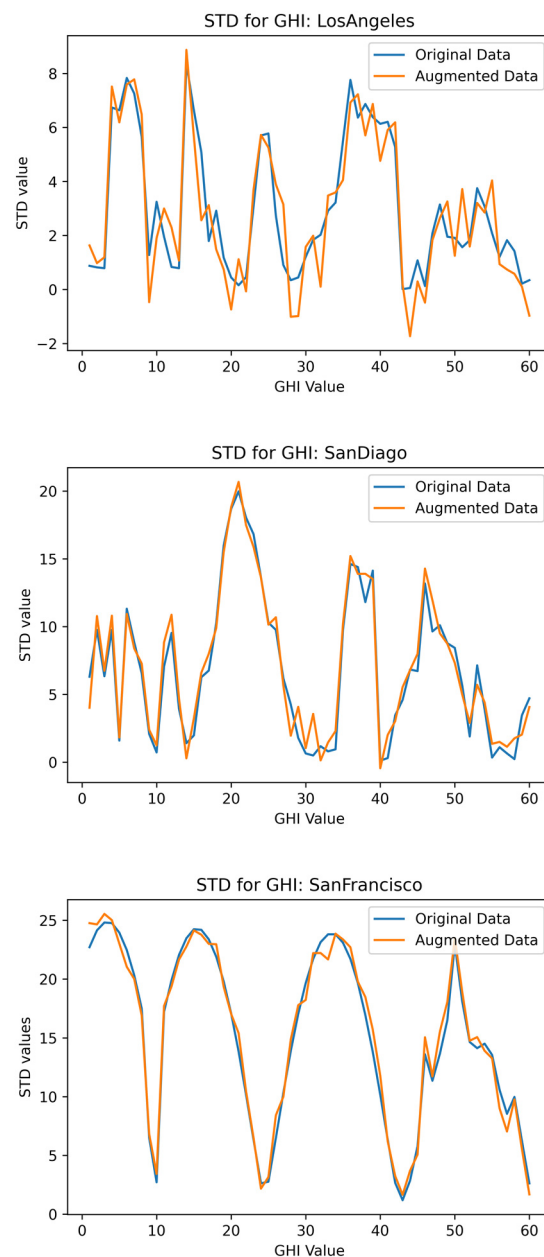


Figure 4. The STD curve of GHI data for the original and augmented dataset.

To summarize, the GAN-CSVR model aims to create distinctive new specimens that extract the inherent properties of the pristine data. The GAN-CSVR model offered high pattern diversity where the generated solar radiation data curves were similar to the original but not identical. This event further confirms the two critical properties of the GAN-CSVR model: statistical similarity and pattern diversity for the new augmented datasets.

The STD values of solar radiation data between the original and augmented datasets by the GAN-CSVR model are 0.0208, 0.1603, and 0.9393 for San Francisco, San Diego, and Los Angeles, respectively. The error value of the STD in Los Angeles is slightly higher than in San Francisco and San Diego. The varied geography of the area, in addition to the rapid climatic fluctuations and the periodic cloudiness cycle, may be responsible for the high value for error of the STD in Los Angeles.

This study utilizes standard datasets for three major zones of California, which are San Diego, San Francisco, and Los Angeles; these regions cover a variety of climates and experience a variety of weather conditions. Using three different datasets from distinct locations is a robust validation strategy and evaluates how well the GAN-CSVR model

generalizes across diverse data distributions. The normalization techniques help in the success of the GAN-CSVR model to adapt to different data types, making it versatile enough to generate realistic samples across various scenarios. The GAN-CSVR model performs well across different datasets, suggesting that it has a strong capacity for learning and generating samples representative of various real-world scenarios.

4.1.2. Cumulative Distribution Function (CDF)

The CDF, also known as the cumulative density function of a real-valued random variable Z , which is assessed at z , is the probability Z will have a value that is either less than or equal to z [1]. Under the scenario of continuous distribution, it provides the area under the probability density function ranging from minus infinity to z . The CDF of the random variable Z can be expressed as the integral of its probability density function (f_{XZ}), as will be shown in Equation (11):

$$F_z(z) = P(Z \leq z) = \int_{-\infty}^z f_z(t)dt \quad (11)$$

The CDF compares the statistical similarity between the original samples and samples made with the GAN-CSVR model by looking at the probability distribution of solar radiation data indirectly.

Regarding the two datasets, the CDF between the average original solar radiation data values and curves and the average generated solar radiation data values and curves by the GAN-CSVR model are compared. The average error value of the CDF of solar radiation data between the original and augmented datasets by the GAN-CSVR model is 7.443981, 4.968554, and 1.495882 for Los Angeles, San Diego, and San Francisco, respectively. The visualization curves for the CDF of solar radiation data between the original dataset and the augmented dataset by the GAN-CSVR model for San Francisco are quite near each other, followed by San Diego compared to Los Angeles. This difference in visualization curves is due to the diverse topography of Los Angeles, as well as quick weather variations and the periodic cloudiness cycle.

These findings support the statistical similarity features of new data generated by the GAN-CSVR model, depicted in Figure 5 (i.e., one curve is quite near another). It is important to note that there are very close curves in proximity between the two CDFs in all three locations. These demonstrate and validate that the GAN-CSVR model can produce new data with a similar distribution to the original dataset for the three locations. In conclusion, the GAN-CSVR model could better mimic the original data distribution and support the statistical similarity.

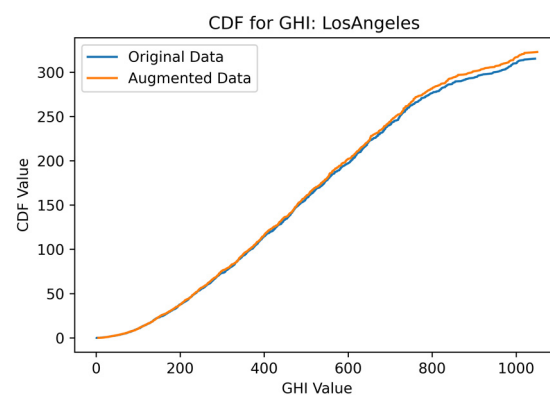


Figure 5. Cont.

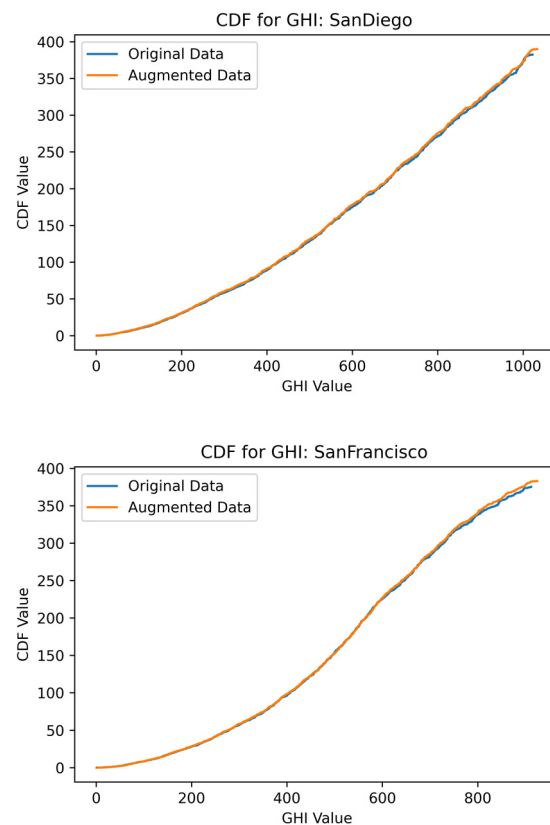


Figure 5. The CDF curve of GHI data for the original and augmented dataset.

In this study, the probability density functions (PDF) of the original and augmented data are continuous, which are represented by $P_g(t)$ and $P_{data}(t)$, respectively. As a result, the associated CDF is shown in Equations (12) and (13):

$$F_{data}(z) = P(Z \leq z) = \int_{-\infty}^z P_{data}(t) dt \quad (12)$$

$$F_z(z) = P(Z \leq z) = \int_{-\infty}^z P_g(t) dt \quad (13)$$

4.2. Performance of Forecasting Based on Augmented and Original Dataset

This Section provides a comprehensive examination of the performance of modeling tests carried out to predict global horizontal irradiance an hour in advance. The performance of six forecasting models has been compared, including SVM, ANN, LSTM, CNN, CNN-LSTM, and LSTM-CNN, to showcase the enhanced accuracy of the forecasting models utilizing augmented data over the models relying solely on original data. Table 3 illustrates that the GAN-CSVR model-based data augmentation exhibits significant potential in improving the accuracy of deep learning forecasting models compared to the original data. Unlike machine learning models, their efficiency remains constant as the amount of input data increases for three cities: San Diego, San Francisco, and Los Angeles. The evaluation metrics whose values are deemed to be the best across all models are given in bold for each site. The results indicate that the proposed forecasting models, when employed with augmented datasets, exhibit greater forecasting accuracy for the three locations compared to the same models applied with the original datasets. The following is an in-depth account of the study findings:

- The proposed models utilizing augmented datasets outperformed, with RMSE values of 68.56 Wm^2 , 60.39 Wm^2 , and 83.18 Wm^2 for SVM and 61.07 Wm^2 , 57.27 Wm^2 , and 74.59 Wm^2 for ANN in Los Angeles, San Diego, and San Francisco, respectively.

Furthermore, in Los Angeles, San Diego, and San Francisco, the LSTM model outperformed, with RMSE values of 36.91 Wm^2 , 33.28 Wm^2 , and 43.43 Wm^2 , while the CNN model outperformed with RMSE values of 48.84 Wm^2 , 44.23 Wm^2 , and 58.74 Wm^2 . However, the hybrid CNN-LSTM outperformed, with RMSE values of 29.68 Wm^2 , 23.64 Wm^2 , and 34.16 Wm^2 . Finally, LSTM-CNN outperformed, with RMSE values of 25.97 Wm^2 , 22.26 Wm^2 , and 29.14 Wm^2 in Los Angeles, San Diego, and San Francisco, respectively.

- The proposed forecasting models utilizing augmented data exhibited superior performance compared to their original data, yielding highly accurate projections for the specified three sites according to the Correlation Coefficient (R) metric. San Francisco has the highest accurate forecast ($R = 0.9313$), followed by San Diego ($R = 0.9589$) and Los Angeles ($R = 0.9356$) in the SVM model, and San Francisco ($R = 0.9361$), San Diego ($R = 0.9501$), and Los Angeles ($R = 0.9538$) in the ANN model. Furthermore, San Francisco ($R = 0.9678$), San Diego ($R = 0.9693$), and Los Angeles ($R = 0.9695$) in the LSTM model, and San Diego ($R = 0.9165$), San Diego ($R = 0.9408$) and Los Angeles ($R = 0.9392$) in the CNN model. Following that, San Francisco ($R = 0.9699$), followed by San Diego ($R = 0.9687$) and Los Angeles ($R = 0.9691$) in the CNN-LSTM model. Finally, in the LSTM-CNN model, San Francisco has the highest accurate forecast ($R = 0.9889$), followed by San Diego ($R = 0.9832$) and Los Angeles ($R = 0.9836$), as shown in Table 3.
- The performance of the proposed forecasting models utilizing augmented data vs. the original data is more accurate when comparing the RMSE and MAE values, as shown in Figure 6. For example, the enhancement of machine learning models utilizing augmented data over original data with respect to MAE is 1.55% to 1.74% for SVM and 4.09% to 4.78% for ANN for San Francisco, San Diego, and Los Angeles, respectively. According to our observations, the efficiency of machine learning models does not change as input data grow in quantity compared to deep learning models. The enhancement of the CNN model utilizing augmented data was 34.13%, 33.44%, and 32.64% for Los Angeles, San Diego, and San Francisco, respectively. Moreover, the enhancement performance of the LSTM model utilizing augmented data improved by 34.49% in Los Angeles, 34.71% in San Diego, and 36.13% in San Francisco, respectively. Furthermore, in San Francisco, San Diego, and Los Angeles, the hybrid model CNN-LSTM model improved by 44.17%, 42.54%, and 42.31%, respectively. Furthermore, the forecasting of augmented data improved the performance of the hybrid model LSTM-CNN in San Francisco, San Diego, and Los Angeles by 44.46%, 43.91%, and 43.12%, respectively. Figure 6 compares the percentage improvement of the proposed models based on augmented vs. original data in terms of RMSE and MAE. This enhancement demonstrates that providing sufficient training data for the forecasting model has an impact on how well the proposed models perform. Consequently, training data augmentation techniques can help overcome overfitting issues in deep learning models and improve forecasting accuracy.
- Augmenting training data to forecast solar radiation has profound scientific implications; it enables forecasting models to understand complex atmospheric processes better and improve decision making. More data allow for better training of deep learning models to capture patterns and relationships between meteorological features and solar radiation. In addition, missing solar radiation data are generated to provide temporal continuity of the time series data, resulting in more reliable predictions of solar radiation.
- Finally, this study demonstrated the superiority of standard models utilizing augmented data on the original data in all cases, based on the comparability of the datasets (climatology, geography of the study area, and dataset size). As a result, the current study's findings are congruent with the benchmark study reported by [59]. This improvement illustrates that augmented data affect how well the deep learning models function. Deep learning models are critical for estimating solar irradiance, especially when dealing with a complex problem with a large amount of data. Furthermore, a

hybrid model provides better accuracy than single deep learning models. However, it extracts temporal and spatial features from the data. In contrast to machine learning models, their efficiency remains constant as input data increase.

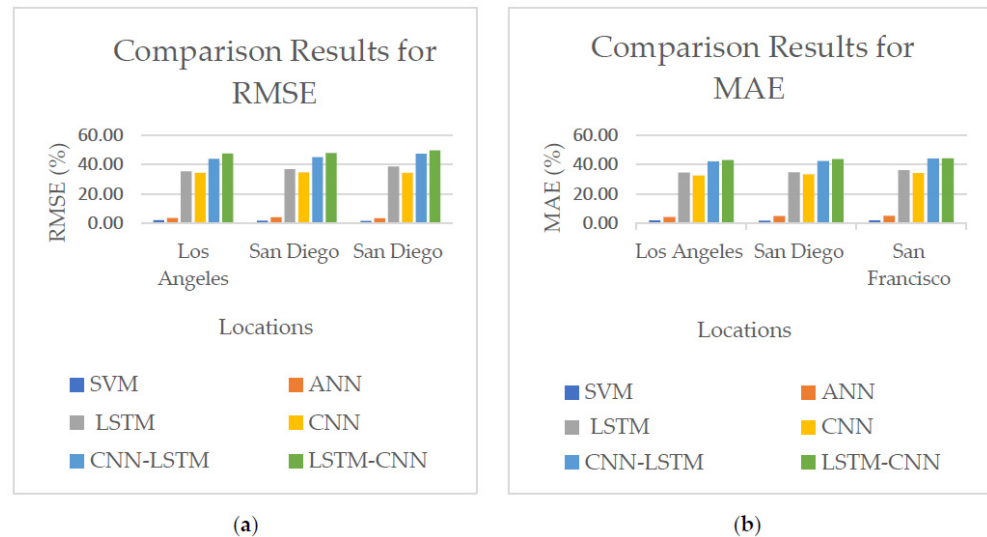


Figure 6. Comparison of improvement percentage for proposed models based on augmented and original data in terms of (a) RMSE (b) MAE.

Table 3. The forecasting performance of proposed models utilizing augmented data generated by the GAN-CSVR model vs. original data for the three target locations.

| Location | Model | Original Data | | | Augmented Data | | |
|---------------|----------|---------------|-------|--------|----------------|-------|--------|
| | | RMSE | MAE | R | RMSE | MAE | R |
| Los Angeles | SVM | 69.92 | 40.81 | 0.9348 | 68.56 | 40.12 | 0.9356 |
| | ANN | 63.18 | 36.88 | 0.9496 | 61.07 | 35.37 | 0.9538 |
| | LSTM | 57.16 | 36.82 | 0.9613 | 36.91 | 24.12 | 0.9695 |
| | CNN | 74.35 | 43.71 | 0.9221 | 48.84 | 29.52 | 0.9392 |
| | CNN-LSTM | 53.01 | 34.98 | 0.9672 | 29.68 | 17.73 | 0.9691 |
| | LSTM-CNN | 49.61 | 31.17 | 0.9701 | 25.97 | 17.31 | 0.9836 |
| San Diego | SVM | 61.33 | 36.03 | 0.9552 | 60.39 | 35.47 | 0.9589 |
| | ANN | 59.56 | 36.91 | 0.9491 | 57.27 | 35.19 | 0.9501 |
| | LSTM | 52.81 | 33.82 | 0.9675 | 33.28 | 22.08 | 0.9693 |
| | CNN | 67.89 | 39.33 | 0.9398 | 44.23 | 26.57 | 0.9408 |
| | CNN-LSTM | 44.98 | 29.96 | 0.9672 | 23.64 | 17.69 | 0.9687 |
| | LSTM-CNN | 42.89 | 27.38 | 0.9701 | 22.26 | 15.36 | 0.9832 |
| San Francisco | SVM | 84.38 | 51.64 | 0.9289 | 83.18 | 50.74 | 0.9313 |
| | ANN | 77.12 | 46.83 | 0.9319 | 74.59 | 44.59 | 0.9361 |
| | LSTM | 70.94 | 40.99 | 0.9587 | 43.23 | 26.18 | 0.9678 |
| | CNN | 89.63 | 50.98 | 0.9004 | 58.74 | 33.53 | 0.9165 |
| | CNN-LSTM | 61.08 | 37.91 | 0.9611 | 34.16 | 22.57 | 0.9699 |
| | LSTM-CNN | 58.12 | 37.02 | 0.9673 | 29.14 | 21.46 | 0.9889 |

Figure 7 shows a comparison of the accuracy of the proposed models based on augmented vs. original data in terms of RMSE for the target three locations. The RMSE improved by utilizing the augmented data for all three datasets compared to the original data for the deep learning model. A lower RMSE value means good model performance [23].

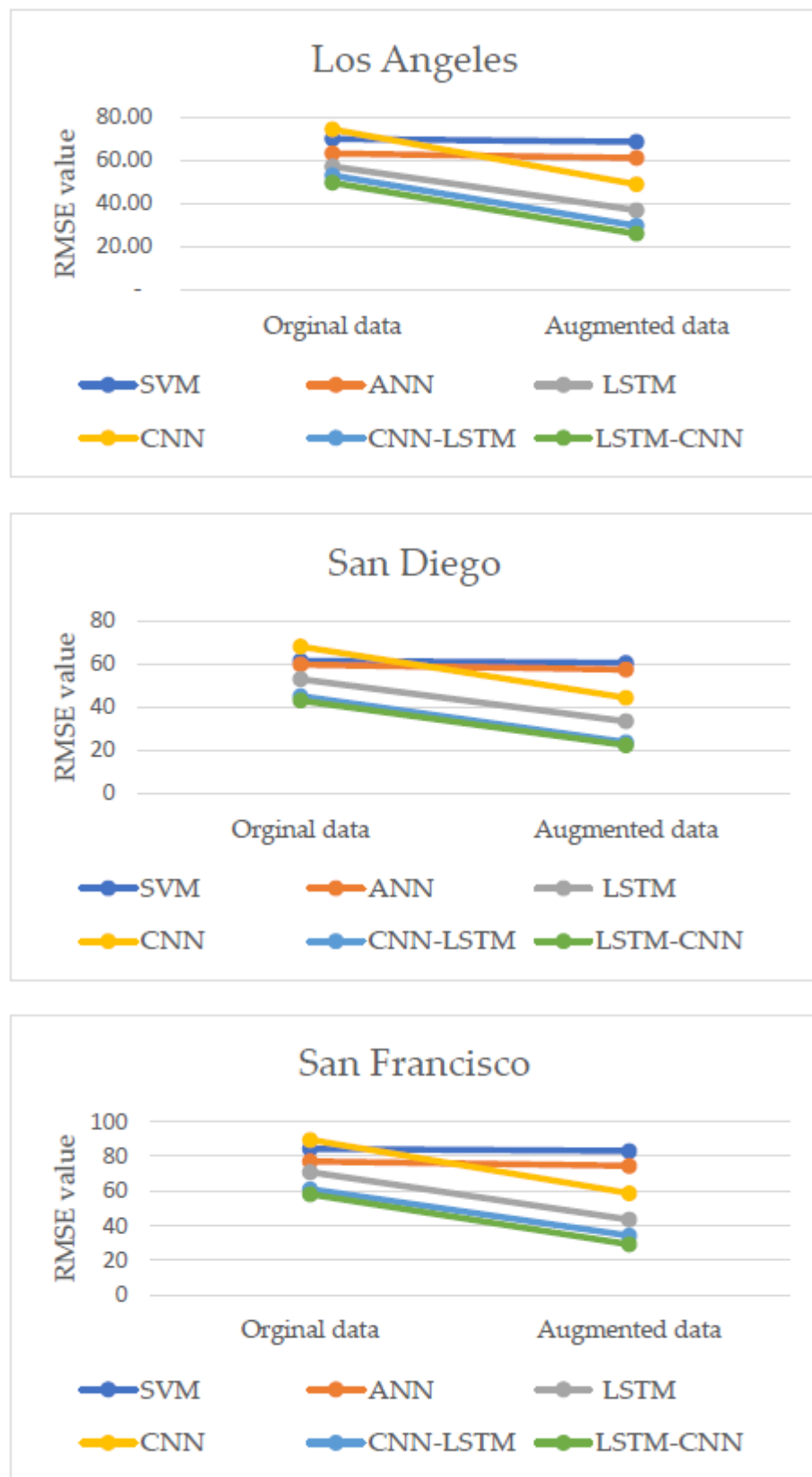


Figure 7. Compares the performance of proposed models utilizing augmented data over original data for three locations in terms of RMSE.

Conclusively, the GAN-CSVR model demonstrates the power of deep learning along with the hybrid model that has been developed to augment datasets, in which the GAN-CSVR model is able to efficiently capture the intrinsic features from the time series data and generate new data identical to the original data. Furthermore, the proposed GAN-CSVR model can be generalizable to different geographical locations or longer times, such as (medium- and long-term periods) for forecasting models. However, the proposed approach can be recommended for use as a viable generative model to extend data sets in a variety of domains, including solar power data, electricity consumption data, load forecasting data, wind power data, rainfall data, and other practical engineering applications that are according to time series data.

5. Conclusions

Although deep learning algorithms offer solid performance, particularly when dealing with spatial and temporal features, a lack of training data poses significant hurdles in forecasting models for solar energy. Therefore, the GAN-CSVR generative model has been proposed to expand datasets to improve predicting performance. The convolutional GAN has been developed by replacing the fully connected CNN layer with a simpler linear SVR layer to restrict the deviation of generated specimens. The proposed model was trained using a Multi-Objective loss function that combines MSE and BCE to ensure consistency between the input and generated samples. Two indexes are used to evaluate the quality of the data generated by the GAN-CSVR model (i.e., STD and CDF). According to the findings of simulations, GAN-CSVR can produce new and distinctive samples that accurately reflect the fundamental characteristics of the original data.

Furthermore, the results of deep learning forecasting utilizing augmented data have been enhanced. In contrast to machine learning models, their efficiency remains constant. In conclusion, the proposed model offers a feasible solution to the problem of a small sample size during the training of the model. Balancing the augmented data according to the distribution of weather patterns and seasons will be explored in the future.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app132312768/s1>.

Author Contributions: Conceptualization, A.M.A. and H.H.; Data curation, H.N.A.H.; Formal analysis, F.A.G., M.E.D. and T.A.E.E.; Funding acquisition, M.E.D. and T.A.E.E.; Investigation, F.A.G., A.M.A. and T.A.E.E.; Methodology, A.M.A. and H.H.; Resources, M.E.D. and T.A.E.E.; Software, H.N.A.H.; Supervision, H.H.; Validation, F.A.G., H.H., H.N.A.H. and M.E.D.; Writing—original draft, A.M.A.; Writing—review and editing, A.M.A. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deanship of Scientific Research at King Khalid University. This work is under grant number RGP2/52/44.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The National Renewable Energy Laboratory (NREL) provides hourly global horizontal irradiance, and solar radiation data are freely available at http://www.nrel.gov/midc/srrl_bms (accessed on 1 May 2021). Software License: Python 3.7 in environment Anaconda was used with the following libraries: Tensorflow, Keras, and sklearn libraries.

Acknowledgments: The authors extend their gratitude to the Deanship of Scientific Research at King Khalid University for funding this work through Grant Number RGP2/52/44, and the first author would like to thank the Iraqi Ministry of Electricity, General Company for North Electricity Distribution.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, F.; Zhang, Z.; Liu, C.; Yu, Y.; Pang, S.; Duić, N.; Shafie-Khah, M.; Catalao, J.P. Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting. *Energy Convers. Manag.* **2019**, *181*, 443–462. [[CrossRef](#)]
2. Wang, F.; Zhen, Z.; Wang, B.; Mi, Z. Comparative study on KNN and SVM based weather classification models for day ahead short term solar PV power forecasting. *Appl. Sci.* **2017**, *8*, 28. [[CrossRef](#)]
3. Douzas, G.; Bacao, F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Syst. Appl.* **2018**, *91*, 464–471. [[CrossRef](#)]
4. Wei, W.; Li, J.; Cao, L.; Ou, Y.; Chen, J. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web* **2013**, *16*, 449–475. [[CrossRef](#)]
5. Vickers, N.J. Animal communication: When I'm calling you, will you answer too? *Curr. Biol.* **2017**, *27*, R713–R715. [[CrossRef](#)]
6. Tran, D.-H.; Sareni, B.; Roboam, X.; Espanet, C. Integrated optimal design of a passive wind turbine system: An experimental validation. *IEEE Trans. Sustain. Energy* **2010**, *1*, 48–56. [[CrossRef](#)]
7. Engerer, N. Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia. *Sol. Energy* **2015**, *116*, 215–237. [[CrossRef](#)]
8. Wang, F.; Zhen, Z.; Liu, C.; Mi, Z.; Shafie-khah, M.; Catalão, J.P. Time-section fusion pattern classification based day-ahead solar irradiance ensemble forecasting model using mutual iterative optimization. *Energies* **2018**, *11*, 184. [[CrossRef](#)]
9. Liu, W.; Ren, C.; Xu, Y. Missing-Data Tolerant Hybrid Learning Method for Solar Power Forecasting. *IEEE Trans. Sustain. Energy* **2022**, *13*, 1843–1852. [[CrossRef](#)]
10. Solomon, C.; Breckon, T. *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
11. Chatterjee, S.; Simonoff, J.S. *Handbook of Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
12. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [[CrossRef](#)]
13. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
14. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
15. Chen, Z.; Jiang, C. Building occupancy modeling using generative adversarial network. *Energy Build.* **2018**, *174*, 372–379. [[CrossRef](#)]
16. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [[CrossRef](#)]
17. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [[CrossRef](#)]
18. Wang, G.; Kang, W.; Wu, Q.; Wang, Z.; Gao, J. Generative adversarial network (GAN) based data augmentation for palmprint recognition. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, Australia, 10–13 December 2018; pp. 1–7.
19. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
20. Lappalainen, K.; Wang, G.C.; Kleissl, J. Estimation of the largest expected photovoltaic power ramp rates. *Appl. Energy* **2020**, *278*, 115636. [[CrossRef](#)]
21. Doorga, J.R.S.; Dhurmea, K.R.; Rughooputh, S.; Boojhawon, R. Forecasting mesoscale distribution of surface solar irradiation using a proposed hybrid approach combining satellite remote sensing and time series models. *Renew. Sustain. Energy Rev.* **2019**, *104*, 69–85. [[CrossRef](#)]
22. Zhang, X.; Li, Y.; Lu, S.; Hamann, H.F.; Hodge, B.-M.; Lehman, B. A solar time based analog ensemble method for regional solar power forecasting. *IEEE Trans. Sustain. Energy* **2018**, *10*, 268–279. [[CrossRef](#)]
23. Jumin, E.; Basaruddin, F.B.; Yusoff, Y.B.M.; Latif, S.D.; Ahmed, A.N. Solar radiation prediction using boosted decision tree regression model: A case study in Malaysia. *Environ. Sci. Pollut. Res.* **2021**, *28*, 26571–26583. [[CrossRef](#)]
24. Essam, Y.; Ahmed, A.N.; Ramli, R.; Chau, K.-W.; Idris Ibrahim, M.S.; Sherif, M.; Sefelnasr, A.; El-Shafie, A. Investigating photovoltaic solar power output forecasting using machine learning algorithms. *Eng. Appl. Comput. Fluid Mech.* **2022**, *16*, 2002–2034. [[CrossRef](#)]
25. Ehteram, M.; Ahmed, A.N.; Fai, C.M.; Afan, H.A.; El-Shafie, A. Accuracy enhancement for zone mapping of a solar radiation forecasting based multi-objective model for better management of the generation of renewable energy. *Energies* **2019**, *12*, 2730. [[CrossRef](#)]
26. Zang, H.; Cheng, L.; Ding, T.; Cheung, K.W.; Wei, Z.; Sun, G. Day-ahead photovoltaic power forecasting approach based on deep convolutional neural networks and meta learning. *Int. J. Electr. Power Energy Syst.* **2020**, *118*, 105790. [[CrossRef](#)]
27. Assaf, A.M.; Haron, H.; Abdull Hamed, H.N.; Ghaleb, F.A.; Qasem, S.N.; Albarrak, A.M. A Review on Neural Network Based Models for Short Term Solar Irradiance Forecasting. *Appl. Sci.* **2023**, *13*, 8332. [[CrossRef](#)]
28. Ehteram, M.; Ahmed, A.N.; Khozani, Z.S.; El-Shafie, A. Graph convolutional network–Long short term memory neural network–multi layer perceptron–Gaussian process regression model: A new deep learning model for predicting ozone concentration. *Atmos. Pollut. Res.* **2023**, *14*, 101766. [[CrossRef](#)]

29. Ehteram, M.; Ahmed, A.N.; Sheikh Khozani, Z.; El-Shafie, A. Convolutional Neural Network-Support Vector Machine Model-Gaussian Process Regression: A New Machine Model for Predicting Monthly and Daily Rainfall. *Water Resour. Manag.* **2023**, *37*, 3631–3655. [[CrossRef](#)]
30. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
31. Wang, Y.; Li, H. A novel intelligent modeling framework integrating convolutional neural network with an adaptive time-series window and its application to industrial process operational optimization. *Chemom. Intell. Lab. Syst.* **2018**, *179*, 64–72. [[CrossRef](#)]
32. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
33. Cavalli, S.; Amoretti, M. CNN-based multivariate data analysis for bitcoin trend prediction. *Appl. Soft Comput.* **2021**, *101*, 107065. [[CrossRef](#)]
34. Ju, Y.; Sun, G.; Chen, Q.; Zhang, M.; Zhu, H.; Rehman, M.U. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting. *IEEE Access* **2019**, *7*, 28309–28318. [[CrossRef](#)]
35. Shreya, M.; Rai, R.; Shukla, S. Forest Fire Prediction Using Machine Learning and Deep Learning Techniques. In *Computer Networks and Inventive Communication Technologies: Proceedings of Fifth ICCNCT 2022, Coimbatore, India, 1–2 April 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 683–694.
36. Xie, G.; Shangguan, A.; Fei, R.; Ji, W.; Ma, W.; Hei, X. Motion trajectory prediction based on a CNN-LSTM sequential model. *Sci. China Inf. Sci.* **2020**, *63*, 212207. [[CrossRef](#)]
37. Salcedo-Sanz, S.; Rojo-Álvarez, J.L.; Martínez-Ramón, M.; Camps-Valls, G. Support vector machines in engineering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2014**, *4*, 234–267. [[CrossRef](#)]
38. Piri, J.; Shamsirband, S.; Petković, D.; Tong, C.W.; ur Rehman, M.H. Prediction of the solar radiation on the Earth using support vector regression technique. *Infrared Phys. Technol.* **2015**, *68*, 179–185. [[CrossRef](#)]
39. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [[CrossRef](#)]
40. Li, Z.; Huang, X.; Zhang, Z.; Liu, L.; Wang, F.; Li, S.; Gao, S.; Xia, J. Synthesis of magnetic resonance images from computed tomography data using convolutional neural network with contextual loss function. *Quant. Imaging Med. Surg.* **2022**, *12*, 3151. [[CrossRef](#)]
41. Chanchal, A.K.; Lal, S.; Kini, J. Deep structured residual encoder-decoder network with a novel loss function for nuclei segmentation of kidney and breast histopathology images. *Multimed. Tools Appl.* **2022**, *81*, 9201–9224. [[CrossRef](#)]
42. Arunkumar, K.; Kalaga, D.V.; Kumar, C.M.S.; Kawaji, M.; Brenza, T.M. Forecasting of COVID-19 using deep layer recurrent neural networks (RNNs) with gated recurrent units (GRUs) and long short-term memory (LSTM) cells. *Chaos Solitons Fractals* **2021**, *146*, 110861. [[CrossRef](#)]
43. Ghimire, S. Predictive Modelling of Global Solar Radiation with Artificial Intelligence Approaches Using MODIS Satellites and Atmospheric Reanalysis Data for Australia. Ph.D. Thesis, University of Southern Queensland, Toowoomba, Australia, 2019.
44. Deo, R.C.; Wen, X.; Qi, F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Appl. Energy* **2016**, *168*, 568–593. [[CrossRef](#)]
45. Al-Musaylh, M.S.; Deo, R.C.; Adamowski, J.F.; Li, Y. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. *Adv. Eng. Inform.* **2018**, *35*, 1–16. [[CrossRef](#)]
46. Al-Musaylh, M.S.; Deo, R.C.; Li, Y.; Adamowski, J.F. Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. *Appl. Energy* **2018**, *217*, 422–439. [[CrossRef](#)]
47. Ghimire, S.; Bhandari, B.; Casillas-Perez, D.; Deo, R.C.; Salcedo-Sanz, S. Hybrid deep CNN-SVR algorithm for solar radiation prediction problems in Queensland, Australia. *Eng. Appl. Artif. Intell.* **2022**, *112*, 104860. [[CrossRef](#)]
48. NREL. *MIDC/SRRL Baseline Measurement System*; NREL: Golden, CO, USA, 2021.
49. Meenal, R.; Selvakumar, A.I. Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. *Renew. Energy* **2018**, *121*, 324–343. [[CrossRef](#)]
50. Qing, X.; Niu, Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy* **2018**, *148*, 461–468. [[CrossRef](#)]
51. Osman, A.I.A.; Ahmed, A.N.; Huang, Y.F.; Kumar, P.; Birima, A.H.; Sherif, M.; Sefelnasr, A.; Ebraheem, A.A.; El-Shafie, A. Past, present and perspective methodology for groundwater modeling-based machine learning approaches. *Arch. Comput. Methods Eng.* **2022**, *29*, 3843–3859. [[CrossRef](#)]
52. Gronkowski, P. The outbursts of the comet 29P/Schwassmann-Wachmann 1: A new approach to the old problem. *Astron. Nachrichten* **2014**, *335*, 124–134. [[CrossRef](#)]
53. Jebli, I.; Belouadha, F.-Z.; Kabbaj, M.I. The forecasting of solar energy based on Machine Learning. In *Proceedings of the 2020 International Conference on Electrical and Information Technologies (ICEIT)*, Rabat, Morocco, 4–7 March 2020; pp. 1–8.
54. Ahmed Mohammed, A.; Aung, Z. Ensemble learning approach for probabilistic forecasting of solar power generation. *Energies* **2016**, *9*, 1017. [[CrossRef](#)]
55. Sobri, S.; Koochi-Kamali, S.; Rahim, N.A. Solar photovoltaic generation forecasting methods: A review. *Energy Convers. Manag.* **2018**, *156*, 459–497. [[CrossRef](#)]

56. Ding, K.; Feng, L.; Wang, X.; Qin, S.; Mao, J. Forecast of pv power generation based on residual correction of markov chain. In Proceedings of the 2015 International Conference on Control, Automation and Information Sciences (ICCAIS), Changshu, China, 29–31 October 2015; pp. 355–359.
57. Yadav, H.K.; Pal, Y.; Tripathi, M.M. Photovoltaic power forecasting methods in smart power grid. In Proceedings of the 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 17–20 December 2015; pp. 1–6.
58. Ahmad, M.W.; Mourshed, M.; Rezgui, Y. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy* **2018**, *164*, 465–474. [[CrossRef](#)]
59. Kumari, P.; Toshniwal, D. Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. *Appl. Energy* **2021**, *295*, 117061. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.