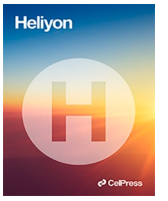
Heliyon 10 (2024) e27752

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A short- and medium-term forecasting model for roof PV systems with data pre-processing

Da-Sheng Lee, Chih-Wei Lai [1,*], Shih-Kai Fu

National Taipei University of Technology Energy and Refrigerating Air-conditioning Engineering, Room 610, College of Mechanical & Electrical Engineering, Integrated Technology Complex, No.1, Sec. 3, Zhongxiao E. Rd., Da'an Dist., Taipei City 10608, Taiwan

ARTICLE INFO

Keywords:
Long short-term memory (LSTM)
Multilayer perceptron (MLP)
Data pre-processing
Prediction of solar energy

ABSTRACT

This study worked with Chunghwa Telecom to collect data from 17 rooftop solar photovoltaic plants installed on top of office buildings, warehouses, and computer rooms in northern, central and southern Taiwan from January 2021 to June 2023. A data pre-processing method combining linear regression and K Nearest Neighbor (k-NN) was proposed to estimate missing values for weather and power generation data. Outliers were processed using historical data and parameters highly correlated with power generation volumes were used to train an artificial intelligence (AI) model. To verify the reliability of this data pre-processing method, this study developed multilayer perceptron (MLP) and long short-term memory (LSTM) models to make short-term and medium-term power generation forecasts for the 17 solar photovoltaic plants. Study results showed that the proposed data pre-processing method reduced normalized root mean square error (nRMSE) for short- and medium-term forecasts in the MLP model by 17.47% and 11.06%, respectively, and also reduced the nRMSE for short- and medium-term forecasts in the LSTM model by 20.20% and 8.03%, respectively.

1. Introduction

Energy transformations, reduced consumption of non-renewable energies, and increased use of renewable energies have become global trends in recent years due to energy shortages and climate change. Solar energy is currently the most widely used type of renewable energy as it holds advantages over other renewable energy sources in terms of availability, cost-effectiveness, accessibility, device capacity, and power generation efficiency [1]. A statistical report released by the International Renewable Energy Agency (IRENA) in 2023 revealed that the total installed capacity of renewable energy was 3,371,793 MW in 2022, with solar energy accounting for 31.2%. The installed capacity of renewable energy increased by 294,555 MW in 2022 alone, with solar energy accounting for 65% [2]. Solar photovoltaic technologies have been developed over many years and there are now mature and commercial applications available; the capacity of solar photovoltaic stations has now reached the GW-level [3] and construction costs have been greatly reduced [4].

Solar photovoltaic systems can be divided into three types according to installation location: rooftop type, ground type, and water

* Corresponding author. Room 610, College of Mechanical & Electrical Engineering, Integrated Technology Complex, No.1, Sec. 3, Zhongxiao E. Rd., Da'an Dist., Taipei City 10608, Taiwan.

E-mail addresses: f11167@ntut.edu.tw (D.-S. Lee), mac100450313@gmail.com (C.-W. Lai), com50302@gmail.com (S.-K. Fu).

[1] Present/permanent address: Room 610, College of Mechanical & Electrical Engineering, Integrated Technology Complex, No.1, Sec. 3, Zhongxiao E. Rd., Da'an Dist., Taipei City 10608, Taiwan (R.O.C.).

https://doi.org/10.1016/j.heliyon.2024.e27752
Received 10 October 2023; Received in revised form 15 February 2024; Accepted 6 March 2024
Available online 12 March 2024

type. Rooftop solar photovoltaic systems are installed on building rooftops to maximize utilization of solar energy [5] and are suitable for use in schools [6], residential buildings [7], commercial offices, and factories [8]. Ground-type solar photovoltaic systems, usually large-scale solar power stations, are mainly installed in open areas, so require large areas of land and abundant sunlight [9]. Finally, water-type solar photovoltaic systems are mainly installed on water surfaces [10] such as lakes, reservoirs, oceans, and other bodies of water to make full use of water areas and reduce land use. Both ground-type and water-type (also known as floating-type) solar photovoltaic systems require large installations sites on land or water, but Taiwan is a small and densely populated region that is more suited to rooftop-type solar photovoltaic systems.

Rooftop solar photovoltaic systems mainly absorb sunlight through solar panels, convert light energy into electrical energy to generate electricity, and convert direct current (DC) power into alternating current (AC) power through an inverter. The electricity generated by solar photovoltaic systems adds to building electrical loads; excess power can be integrated into power grids for deployment by power companies, while grid power can be used to supplement electricity usage if generated power is insufficient [11], thereby preventing issues from insufficient power generation due to weather factors or small rooftop areas.

Many other countries are also promoting installation of rooftop solar photovoltaic systems. Statistics for Europe released in 2021 [12] reveal that the Netherlands have 978 power generation systems with an installed capacity of 10.7 MW, Belgium has 4308 power generation systems with an installed capacity of 29.75 MW, Luxembourg has 86 power generation systems with an installed capacity of 1.56 MW, Germany has 24,204 power generation systems with an installed capacity of 325.73 MW, France has 474 power generation systems with an installed capacity of 5.15 MW, and Italy has 2694 power generation systems with an installed capacity of 24.23 MW. In Asia, the potential installed capacity of Hong Kong was estimated to be 5.97 GW in 2013 [13] and Turkey's rooftop solar photovoltaic system installed capacity in 2016 was 200 MW [14]. In Africa, Abu Dhabi (the capital city of the United Arab Emirates) had an installed capacity of 2.3 MW in 2011 [15].

There have been a number of studies on solar photovoltaic systems in recent years, particularly associated with use of crystals in solar panels, power generation efficiency, and power dispatching in smart grids. Following multiple years of technological development, solar panel manufacturing technologies have become more sophisticated and production costs have decreased. Therefore, construction costs for solar photovoltaic plants are decreasing year over year while construction scales are increasing year by year. Although there is an abundance of research on rooftop solar photovoltaic systems, the efficiency and stability of solar photovoltaic systems are still heavily affected by the amount of sunlight and weather conditions in practice [16]. This study worked with Chunghwa Telecom, the largest telecommunications company in Taiwan, to collect historical data from rooftop solar photovoltaic systems installed on Chunghwa Telecom's computer rooms and service centers for the purposes of training and building artificial intelligence (AI) models that could be used for predicting power generation performance of rooftop solar photovoltaic systems.

Solar power forecasting can be divided into short-term forecasting, medium-term forecasting, and long-term forecasting according to the length of the forecasting range. Short-term forecasting usually ranges from several hours or 1–7 days, and is commonly used for unit investment, scheduling, and to ensure safety of power grid operations. Medium-term forecasting ranges from 1 week to 1 month and is usually used to formulate power system and unit maintenance plans. Long-term forecasting ranges from 1 month to 1 year and is usually used when bidding on green energy trading platforms as well as when formulating plans for power generation, transmission, and distribution [17]. The sites selected for this study were located in northern, central, and southern Taiwan. As these were all recently constructed sites, there was a lack of historical data that could be used for training. Therefore, January 2021 to June 2023 was set as the data collection interval, and these 30 months of historical data were used to train hourly and daily power generation prediction models. The amount of data was considered to be sufficient for making hourly and daily predictions, although somewhat insufficient for monthly or annual predictions. Additionally, hourly power generation predictions can help enterprises clarify the amount of power generated by solar photovoltaic systems which can be used for temporary deployment or as a basis for judging equipment abnormalities, and daily power generation predictions can be used to make preliminary assessments of green electricity transaction needs. Based on these reasons, this study aimed to provide forecasts of hourly and daily power generation volumes. Hourly power generation predictions were defined as short-term power generation predictions, and daily power generation predictions were defined as medium-term power generation predictions.

An increasing number of prediction techniques have been applied to solar photovoltaic forecasting in recent years, including persistence forecasting, physical models, and statistical techniques. In particular, statistical techniques encompass time series based forecasting techniques and machine learning forecasting techniques. Machine learning forecasting technologies, including artificial neural networks (ANNs), multilayer perceptron (MLP) neural networks, recurrent neural networks (RNNs), feedforward neural networks (FFNNs) and feedback neural networks (FBNNs), are being increasingly used for pattern recognition, data mining, classification problems, filtering, and prediction [18].

MLP neural networks, which are widely used in solar energy forecasting applications [19], are derived from ANNs and can be adapted for different input and output predictions and frameworks. Adel Mellit et al. used MLP to predict solar irradiance in Trieste, Italy over 24 h and developed a fairly accurate model with a correlation coefficient (R^2) of 98–99% for sunny days and a correlation coefficient (R^2) of 94–96% for cloudy days. This model was used to predict the power generated by the grid-connected photovoltaic (GCPV) system on the rooftop of the Trieste municipal government building. The mean absolute error (MAE) of the model was less than 5% and the correlation coefficient (R^2) was between 90% and 92%. These results indicate that MLP is suitable for predicting power volumes generated by GCPV plants [20].

MLP can further reduce root mean square error (RMSE) errors by adjusting the number of neurons. For example, Fermín Rodríguez conducted a simulation on 5 to 15 different neurons and found that a model based on 15 neurons minimized the RMSE of training and verification data [21]. In terms of prediction accuracy for different timeframes, MLP displays better accuracy for very short-term power generation (7.5 min) compared to short-term power generation (15 min or 30 min) [22].

Table 1
Site locations and power generation volumes.

City	Site	Power generation volume (kW)	Data sampling interval	Training data	Verification data
Northern region					
Hsinchu City	N1	99.84	2021/01–2023/06	70%	30%
Hsinchu City	N2	19.84	2021/01–2023/06	70%	30%
Hsinchu City	N3	19.84	2021/01–2023/06	70%	30%
Hsinchu City	N4	19.84	2021/01–2023/06	70%	30%
Central region					
Taichung City	M1	1160	2021/01–2023/06	70%	30%
Taichung City	M2	67.2	2021/01–2023/06	70%	30%
Changhua County	M3	78.72	2021/01–2023/06	70%	30%
Nantou County	M4	39.68	2021/01–2023/06	70%	30%
Nantou County	M5	40.92	2021/01–2023/06	70%	30%
Southern region					
Tainan City	S1	43.68	2021/01–2023/06	70%	30%
Tainan City	S2	65.52	2021/01–2023/06	70%	30%
Tainan City	S3	97.645	2021/01–2023/06	70%	30%
Tainan City	S4	81.9	2021/01–2023/06	70%	30%
Kaohsiung City	S5	52.48	2021/01–2023/06	70%	30%
Kaohsiung City	S6	296	2021/01–2023/06	70%	30%
Pingtung County	S7	91.84	2021/01–2023/06	70%	30%
Pingtung County	S8	132.8	2021/01–2023/06	70%	30%

MLP can also be combined with other techniques for improved accuracy. For example, Qiang Liu et al. combined MLP with a knowledge-based neural network (KBNN), using MLP in instances with sufficient data and KBNN when there was insufficient data. Use of KBNN improved the prediction accuracy of the MLP model by 65.4% [23]. Di Huang combined MLP and LSTM neural network models to predict and analyze electricity consumption data in three regions. Even though MLP displayed higher RMSE compared to LSTM, both methods yielded good time series forecasting performance and were able to effectively and reliably predict photovoltaic power generation [24].

Model inputs can encompass a wide range of data types including time, weather, and power generation data obtained from sensors installed in power stations or open source databases on the Internet. For example, Jose Manuel Barrera et al. used open source data collected by PVOutput and the Photovoltaic Geographical Information System (PVGIS) as input parameters for an ANN model and reduced the mean square error (MSE) to 0.04, lower than the 0.05 seen in other studies [25]. As there are many different types of data, it is inevitable that model results will be affected by data scales and types. In other words, the quality of a prediction model not only depends on the selected technique, but also other aspects such as data pre-processing, feature engineering, and post-processing [26]. Most studies choose to preprocess weather data, and a previous study showed that the normalized root mean square error (nRMSE) of an ANN-based power prediction model could be reduced by 5∼6% when a clearness index or clear sky index was used to pre-process weather data [27]. The average error of a solar radiation prediction model which used wavelet analysis for pre-processing was found to be one-fourth of a model without wavelet analysis [28]. Jinxia Zhang developed an LSTM solar photovoltaic prediction model which used principal component analysis (PCA) to process data and reduce network complexity, then compared the prediction performance of this proposed PCA-LSTM model with an LSTM model and a support vector machine (SVM) model. The prediction results of the PCA-LSTM and LSTM models were closer to actual values compared with the SVM model, and the PCA-LSTM and LSTM models yielded similar results, but the training time of the PCA-LSTM model was shorter than the LSTM model [29]. Changsong Chen et al. recommended using a self-organized map (SOM) to classify possible weather conditions for the next 24 h by analyzing solar irradiance, relative humidity, temperature, power generation, weather forecasts, and other data collected from solar photovoltaic sites, thereby increasing weather forecasting accuracy and improving the forecasting accuracy of power generation prediction models. This method was determined to be suitable for predicting power generation volumes on sunny and cloudy days (R2 values fell between 96% and 99%), and could prove highly useful for operational planning of electricity market transactions [30].

2. Methods

2.1. Data sources

The training data used in this study were provided by Chunghwa Telecom, a leading telecommunications provider in Taiwan which has actively promoted corporate environmental, social, and governance (ESG) actions in recent years. Apart from purchasing green power, Chunghwa Telecom has also installed a number of rooftop solar photovoltaic systems on its service centers and computer room buildings located all over Taiwan. Taiwan is mainly a subtropical region. The north of Taiwan tends to be humid and rainy while the south has abundant sunshine and a climate close to tropical regions. According to a climate monitoring report from the Central Weather Bureau (CWB) [31], the average temperature in Taiwan is 23.6 °C. The lowest temperatures usually occur from late January to early February, when average temperatures are around 18 °C, and the highest temperatures usually occur in July, when average temperatures are around 33 °C. The average annual rainfall volume in Taiwan is 2207 mm. December to January of the following year tends to be the driest period, and the highest rainfall volumes are seen during the rainy season from May to June and the typhoon

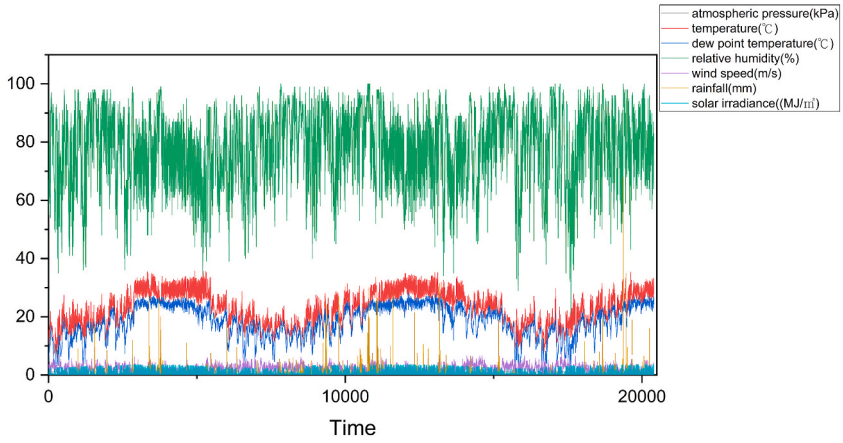


Fig. 1. Hsinchu City hourly weather data (2021/01–2023/06).

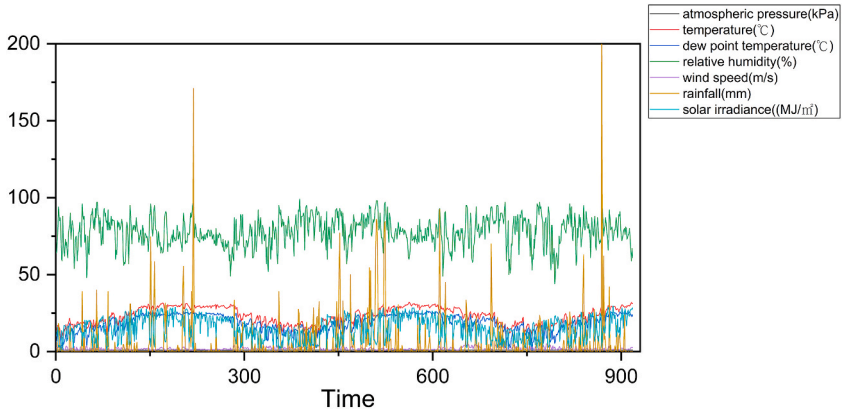


Fig. 2. Hsinchu City daily weather information (2021/01–2023/06).

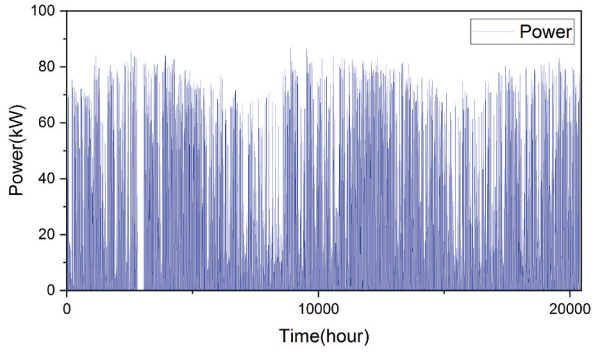


Fig. 3. Hourly power generation at N1 (2021/01–2023/06).

season from July to September.

This study selected sites located in the northern, central, and southern regions of Taiwan and collected information for training and simulation. Site information is shown in Table 1. The sites contained a number of equipment such as inverters, electricity meters, thermometers, pyranometers, and other measuring instruments, which were connected via RS485 or TCP. The data were collected by gateways and transmitted to a cloud platform every 3 min.

Hourly and daily data were compiled from these sites. Figs. 1 and 2 show hourly and daily weather data for Hsinchu City. Figs. 3 and 4 show the hourly power generation volumes at the N1 site and the daily power generation volumes at the N2 site, respectively. Power generation, temperature, and solar irradiance volumes were used as training data. In addition to data collected from study sites,

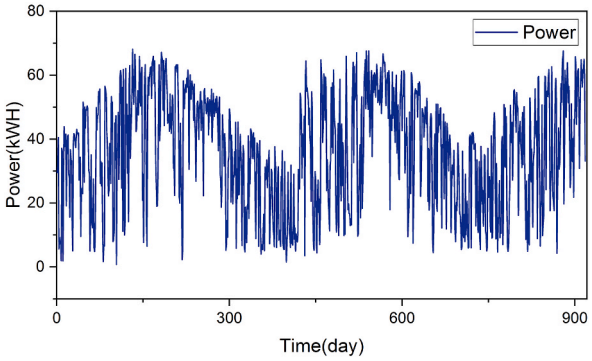


Fig. 4. Daily power generation at N2 (2021/01–2023/06).

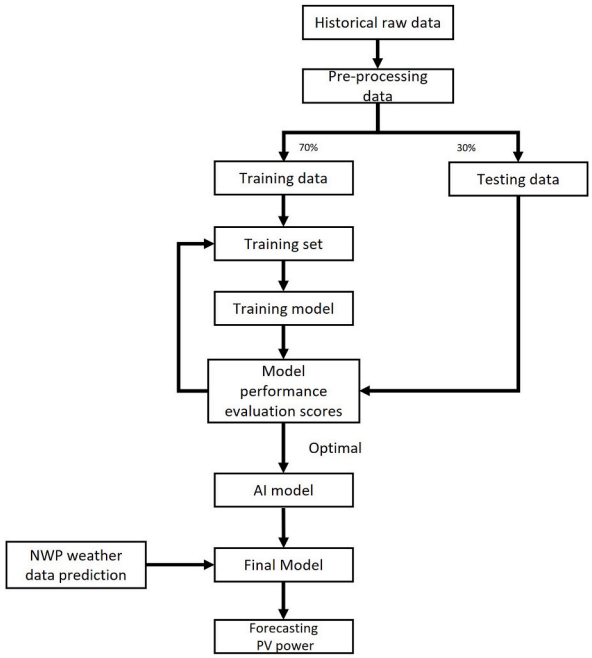


Fig. 5. Flowchart of power generation forecasting model.

weather data was collected from the CWB, including atmospheric pressure, temperature, dew point temperature, relative humidity, wind speed, rainfall, and solar irradiance volumes. This study analyzed and conducted feature extraction on data collected from study sites and the CWB to select model training parameters. Random sampling was used to select 70% of the data which was used as training data, and 30% of the data was used as verification data for the prediction model.

2.2. Forecasting process

This study developed a data pre-processing method which can improve the accuracy of existing AI solar energy prediction models. Fig. 5 shows the flowchart for the prediction model used in this study. The AI model was trained on pre-processed historical data taken from study sites and the CWB. During the training process, the grid search technique was used to adjust model hyperparameters and obtain the best evaluation indicators. The power generation prediction model was considered to be complete once the best evaluation parameters had been selected. Numerical Weather Prediction (NWP) data was subsequently used as model input data to forecast the power generation performance of the rooftop solar photovoltaic sites.

Fig. 6 shows the data pre-processing procedures proposed by this study. Historical data was separated into weather data and power generation data, then linear regression and K Nearest Neighbor (k-NN) were respectively used to estimate missing data, following which outliers were identified and deleted, and the values were estimated again using linear regression or k-NN.

Once all the data had been processed, the Pearson correlation coefficient was used to extract feature values and identify model parameter types before the data was normalized using min-max scaling.

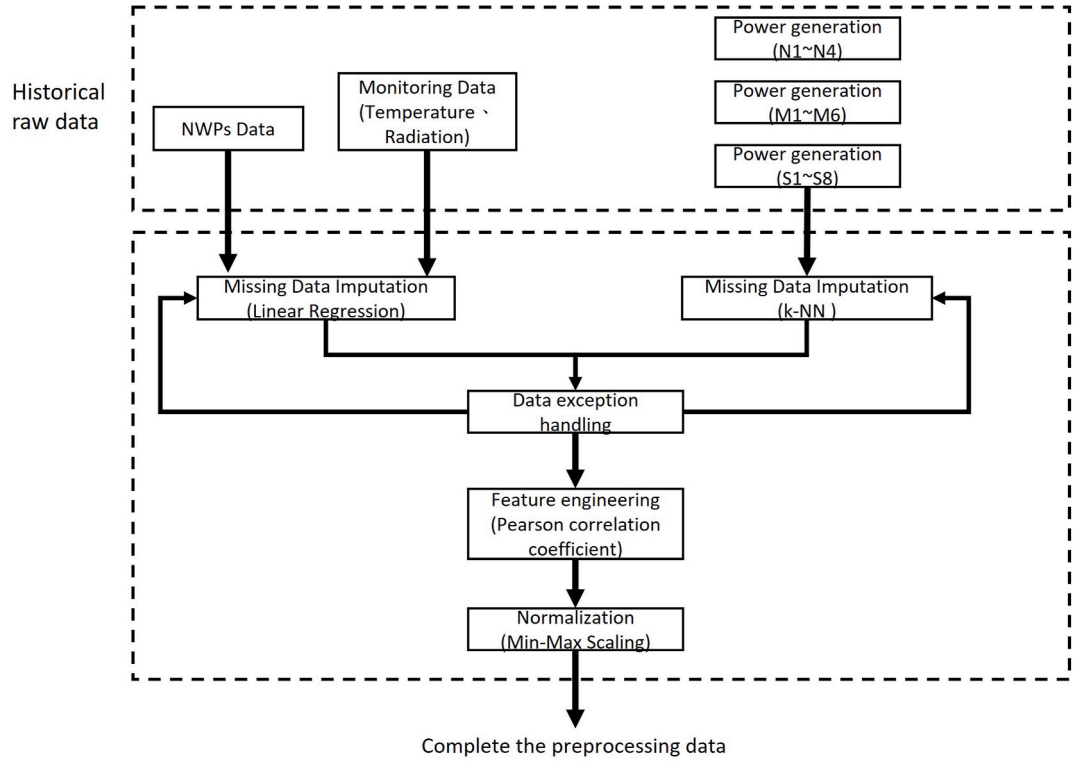


Fig. 6. Data pre-processing flowchart.

2.3. Data pre-processing

2.3.1. Estimation of missing values in weather data—linear regression

The weather data was mainly collected from two sources. The first source was publicly available data taken from CWB weather stations located across Taiwan. This data was easily obtainable, but a disadvantage was that the weather information could be inaccurate for solar photovoltaic sites located further away from CWB weather stations. Data was also collected from weather stations set up at each solar photovoltaic site, which provided timely and accurate information, but required high costs to build and maintain. Both types of weather data mentioned above were measured using sensors, so it was inevitable that some data would be lost due to network or sensor problems. Only around 1% of weather data was missing from the data retrieved from CWB, but around 5–7% of data was missing from the data retrieved from the weather stations set up at the rooftop solar photovoltaic sites. Therefore, the missing data posed a significant problem.

In 2017, Doreswamy et al. tested different measurement models to estimate missing values in weather data and compared the performance of different prediction methods such as linear regression, k nearest neighbor imputation (KNNI), imputation using a prediction model, random forest, SVM, and kernel ridge regression. The results of their study showed that linear regression and random forest yielded better RMSE, mean square error (MSE), variance of error (VARE), and mean absolute error (MAE) compared to the other three methods. Additionally, linear regression yielded a better R^2 value compared to random forest. Based on these results, this study chose to use linear regression as the method for processing weather data [32] using the following formula:

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

y_i^*: The interpolated predicted value of the missing value for the i-th observation.

β_0, β_1: The regression coefficients of the linear regression model, representing the intercept and the slope of the independent variable x_i, respectively.

x_i: An observed dependent variable used to predict missing values.

ε_i: The error coefficient representing the prediction error for the i-th observation.

2.3.2. Estimation of missing values in solar photovoltaic data—k-NN

In addition to weather information, historical power generation data is also paramount when establishing solar photovoltaic prediction models. Similar to weather data, power generation data may have missing values due to network outages or equipment abnormalities. The amount of missing data accounted for around 3–5% of historical power generation data.

The power generation data used in this study were provided by Chunghwa Telecom. As the largest telecommunications company in

Table 2
Training data attributes and names.

Feature	Meaning
Power	Site power generation (kW)
Temp	Site atmospheric temperature (°C)
Rad	Site solar irradiance (W/m^2)
CWB_temp	CWB atmospheric temperature (°C)
CWB_rad	CWB solar irradiance (MJ/m^2)
CWB_press	CWB atmospheric pressure (Pa)
CWB_ws	CWB wind speed (m/s)
CWB_dewtemp	CWB dew point temperature (°C)
CWB_hum	CWB relative humidity (%)
CWB_rain	CWB rainfall (mm)

Taiwan, Chunghwa Telecom has established a large number of small solar photovoltaic stations on rooftops and open spaces of computer rooms and office sites in various locations. Therefore, the company was able to provide information from multiple solar photovoltaic power stations in different regions located in the same county or city.

This data was combined with other related parameters such as time and weather variables, then organized into a dataset. Sites with missing values were set as target sites, and associated power generation data from each target site were used as test samples while data from surrounding sites were used as training samples. To estimate the missing values in target site data, this study calculated the Euclidean distance between each missing value and known values from other sites using the longitude and latitude values of each site to form two-dimensional coordinates. Assuming that the coordinates of point A are (x_1, y_1) and the coordinates of point B are (x_2, y_2), the formula for calculating Euclidean distance (D) was as follows:

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2}$$

Weights were calculated based on the distance between target sites and surrounding sites, with closer sites given higher weight. The formula for calculating the weight (W_i) of surrounding sites was as follows:

$$W_i = \frac{1}{D_i} \tag{3}$$

W_i represents the weight of the i-th site and D_i represents the Euclidean distance of the i-th site. As power generation scales differ for each site, scale correction factors (F_i) for target and surrounding sites were calculated using the following formula:

$$F_i = \frac{S_i}{S_0} \tag{4}$$

S_i represents the power generation scale of the i-th site and S_0 represents the power generation scale of the target site.

Power generation data from the k nearest neighbors were combined with the weight of surrounding sites (W_i) and power generation scale correction factors (F_i), following which the weighted average method was used to calculate the missing value of the target site. Assuming that the missing value of the target site at a certain point in time is Na and the site is surrounded by k nearest neighbors, P_1, P_2 … P_k refer to the power generated by each nearby site at the same time; W_1, W_2 … W_k are the weights of each nearby site; and F_1, F_2 … F_k are the power generation scale correction factors for each nearby site. The weighted average method was used to calculate the missing value of the target site using the following formula:

$$Data_{loss} = \frac{W_1{*}F_1{*}P_1 + W_2{*}F_2{*}P_2 + \ldots + W_k{*}F_k{*}P_k}{W_1{*}F_1 + W_2{*}F_2 + \ldots + W_k{*}F_k} \tag{5}$$

2.3.3. Handling data abnormalities

Abnormal values may appear in solar photovoltaic data due to equipment abnormalities and abnormal values may also be present in weather data taken from the CWB or self-built weather stations due to sensor or equipment abnormalities. Negative values in power generation data were replaced with zero [33]. In accordance with [34], values which exceeded three standard deviations were set as critical values for weather data anomalies. Data exceeding critical values were regarded as outliers and removed, then the missing values were subsequently estimated using linear regression.

2.3.4. Feature extraction

Data from study sites included temperature, illuminance, and power generation, while CWB weather data included atmospheric pressure, temperature, dew point temperature, relative humidity, wind speed, rainfall volume, and solar irradiance, as shown in Table 2. The correlation between input variables and the target variable (power generation) was measured using the Pearson correlation coefficient [35], an analysis method suitable for continuous time series input and target variables which measures correlations using a coefficient ranging between 1 and -1. The Pearson correlation coefficient is mainly used to calculate the standard deviations between input and target variables, and the strength of the correlation can be determined by the coefficient. If a coefficient falls between 0 and 1, the target variable increases alongside the input variable; if a coefficient falls between 0 and -1, the target variable

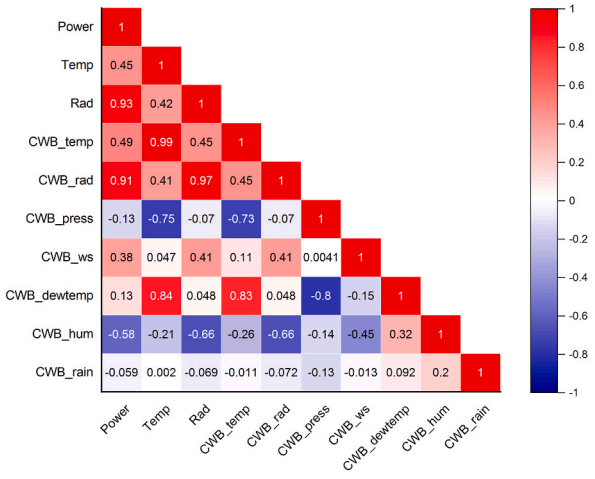


Fig. 7. Pearson correlation coefficients for solar photovoltaic prediction target variable (Power) and the other nine input variables.

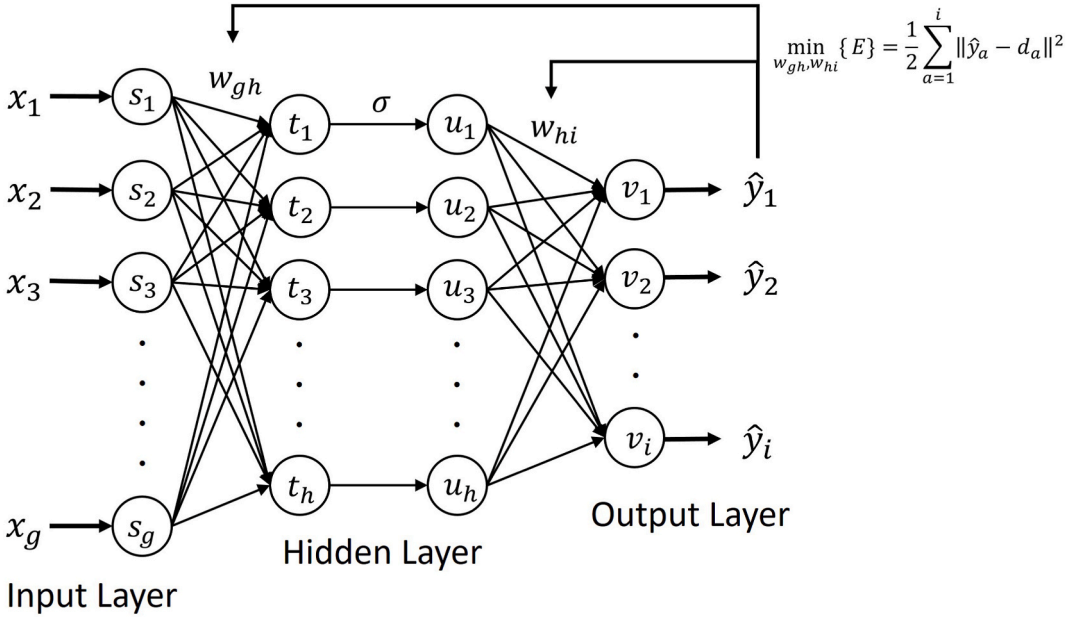


Fig. 8. MLP architecture diagram.

decreases as the input variable increases; and if the coefficient is 0, this means that there is no linear relationship between the input variable and the target variable. The Pearson correlation (P_{xy}) formula is as follows [36]:

$$P_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) - (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

x_i is the input variable, $\bar{x}$ is the mean value of the input variable, y_i is the target variable, and $\bar{y}$ is the mean value of the target variable

As shown in Fig. 7, the target variable Power has low correlation with the three input variables CWB_press, CWB_hum, and CWB_rain, so these three input variables were removed from the model. Only Temp, Rad, CWB_temp, CWB_rad, CWB_ws, and CWB_dewtemp were used as input variables.

2.3.5. Normalization

The prediction model included different types of training data including temperature, humidity, rainfall, solar irradiance, and power generation. As these data adhered to different scales, the model would likely yield poor results if the data were used directly without additional processing. Therefore, normalization was an indispensable step. According to Ref. [37], the min-max scaling

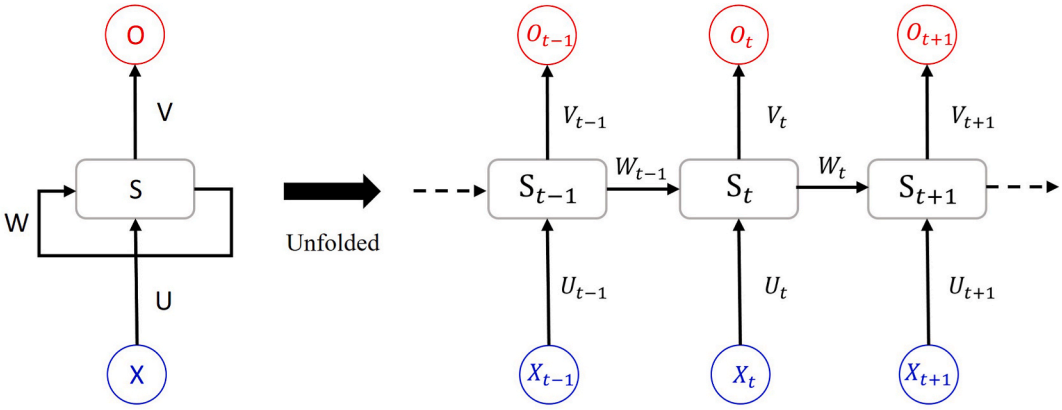


Fig. 9. RNN architecture diagram.

method is commonly used for normalization in solar and wind power prediction model studies: this method scales collected data based on maximum and minimum values so that data ranges fall between 0 and 1, making different data types comparable and consistent. The min-max scaling formula is as follows [38]:

$$X_{scaled} = \frac{X_{rawdata} - \min\left(X_{rawdata}\right)}{\max\left(X_{rawdata}\right) - \min\left(X_{rawdata}\right)} \tag{7}$$

$X_{rawdata}$ represents the original data, and max and min represent the maximum and minimum values

2.4. Multilayer perceptron (MLP)

ANNs are currently the most widely used AI prediction technique for solar photovoltaic forecasting [39]. ANN is a prediction technique that imitates the human nervous system, where each unit is regarded as a neuron, and units are connected together to enable information transmission and data processing. MLP, a prediction method derived from ANN, is commonly applied to classification and regression problems. MLP models are mainly composed of multiple neurons arranged in layers, where the neurons of each layer are connected to the neurons of the previous and next layers.The architecture of MLP is shown in Fig. 8.

The first layer of the entire model is known as the input layer, the last layer is known as the output layer, and the remaining layers are known as hidden layers. MLP models can contain one or more hidden layers, and the neurons in the same layer are not connected to each other [40]. A neuron network built using multiple interconnected neurons can be used to solve complex problems such as classification, pattern recognition, and time series prediction [41]. Each neuron in an MLP model receives inputs from a neuron in the previous layer which it is connected to, and each neuron has a weight parameter (w_{gh}, w_{hi}) and an activation function (σ) which acts on received inputs and generated outputs [42]. The MLP architecture is shown in Fig. 6. Output data is set as $\widehat{y}_i$, x_i is the input data, s_g is the input layer, v_i is the output layer, t_n, u_b are the hidden layers, and e_n represents the weight deviation. The associated formulas are shown below [43].

$$\widehat{y}_i = v_i = w_{hi} \sum_{b=1}^{h} u_b \tag{8}$$

$$u_b = \sigma(t_n + e_n) \tag{9}$$

$$t_n = w_{gh} \sum_{c=1}^{g} s_g = w_{gh} \sum_{c=1}^{g} x_g \tag{10}$$

During the training process for MLP models, an expected value (d_a) is set and errors are backpropagated layer by layer using the backward propagation method. Model weight parameters are updated after each iteration to minimize the difference between predicted and expected outputs using the following formula [44]:

$$\min_{w_{gh}, w_{hi}} \left\{ E \right\} = \frac{1}{2} \sum_{a=1}^{i} \left\| \widehat{y}_a - d_a \right\|^2 \tag{11}$$

2.5. Long short-term memory (LSTM)

Traditional ANNs generated predictions from historical data, but did not consider time correlations in data sequences. Therefore, ANN models cannot capture temporal relationships, which limit their usefulness as a time series forecasting method [45].

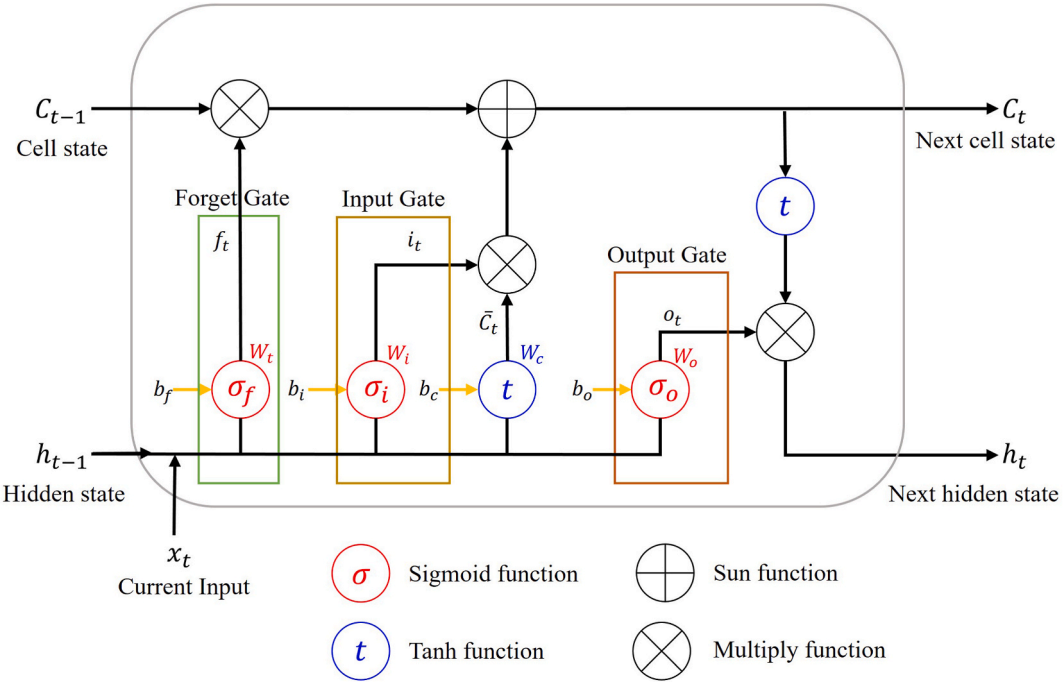


Fig. 10. LSTM unit architecture diagram.

RNNs are a type of AI model developed for processing time series data. Compared with traditional ANNs, which use independent input vectors, RNNs have a feedback function, so output values depend on current and previous input values, output values, or hidden states in neural networks. RNNs have a fundamental component known as a recurrent neuron, which maintain hidden states or remember and utilize previous inputs to capture the time dependence of data [46]. The RNN architecture is shown in Fig. 9 and associated formulas are shown as follows. X is the input layer; O is the output layer; S is the hidden layer; U, V, and W represent weight parameters; and t represents the timepoints.

$$S_t = S_{t-1} \times W_{t-1} + X_t \times U_t \tag{12}$$

$$O_t = S_t \times V_t \tag{13}$$

LSTMs were derived from RNNs to solve difficulties in the training process associated with vanishing or exploding gradients resulting from backpropagation [47] as weights and learning on hidden layers do not change when gradients vanish, but weights increase when gradients explode [48]. LSTMs are suitable for complex functions such as solar photovoltaic prediction, which requires processing of time series data and identification of linear relationships between various data types [49].

LSTM models are usually composed of three layers: the input layer, output layer, and hidden layer. The hidden layer is made up of multiple storage units which each contain an input gate, forget gate, and output gate [50]. The architecture of a unit is shown in Fig. 10. At time point t, x_t is the input value, and h_{t-1}, C_{t-1} refer to the output and cell state left over from the previous time point t-1. At the forget gate, x_t and h_{t-1} are used as input data, and f_t uses the function σ_f to determine what information needs to be deleted. At the input gate, x_t and h_{t-1} generate i_t as new input information based on function σ_i. At the output gate, x_t and h_{t-1} are entered into function σ_o to generate output o_t at time point t. In addition to the training processes of the forget, input, and output gates described above, the new information generated by the model i_t is also combined with new candidate vector values $\overline{C}_t, f_t$, and the cell state at time point t-1 C_{t-1} to generate the cell state C_t at time point t. C_t is then entered into the tanh function and combined with o_t to generate the output left over at the time point t h_t to be used for training at time point t+1. Calculation formulas for each stage in the training process are as follows [51]:

$$\sigma(t) = \frac{1}{1+e^{-t}}$$

$$\tanh(t) = \frac{e^{t} - e^{-t}}{e^{t} + e^{-t}} \tag{15}$$

$$i_t = \sigma_i(W_i[x_t, h_{t-1}] + b_i) \tag{16}$$

$$o_t = \sigma_o(W_o[x_t, h_{t-1}] + b_o) \tag{17}$$

Table 3
MLP hyperparameter settings.

Hyperparameters	Predetermined range
Number of hidden layers	1, 2
Number of neurons	10, 20, 30, 40
Weight optimization	"lbfgs", "adam", "sgd"
Complexity	0.001, 0.01, 0.1
Initial learning rate	0.001, 0.01, 0.1
Maximum number of iterations	100, 200, 300, 400, 500

Table 4
LSTM hyperparameter settings.

Hyperparameters	Predetermined range
Optimizer	"adam", "rmsprop"
Number of units (LSTM layer)	200, 400, 600, 800, 1000
Dropout rate (Dropout layer)	0.1, 0.2, 0.3, 0.4, 0.5, 0.6
Iterations	50, 100, 150, 200
Batch size	72

$$\overline{C}_t = \tanh(W_c[x_t, h_{t-1}] + b_c) \tag{18}$$

$$C_t = (i_t \otimes \overline{C}_t) \oplus f_t \otimes C_{t-1} \tag{19}$$

$$h_t = o_t \otimes \tanh(C_t) \tag{20}$$

where.

σ	Sigmoid function
tanh	Hyperbolic tangent function
x_t	Input value at time t
h_t	Output value at time t
h_{t-1}	Output value at time t-1
C_t	Cell state at time t
C_{t-1}	Cell state at time t-1
f_t	Forget gate output
W_f	Forget gate weight
σ_f	Forget gate sigmoid function
b_f	Forget gate bias
i_t	Input gate output
W_i	Input gate weight
σ_i	Input gate sigmoid function
b_i	Input gate bias
o_t	Output gate output
W_o	Output gate weight
σ_o	Output gate sigmoid function
b_o	Output gate bias
$\overline{C}_t$	Vector of new candidate values for time step t
W_c	tanh layer weight, used to calculate $\overline{C}_t$
b_c	tanh layer bias, used to calculate $\overline{C}_t$

2.6. Hyperparameter adjustment

In machine learning, various hyperparameters are set before training and can affect the performance of the resulting AI model. This study used the grid search technique to adjust model parameters and identify the best hyperparameter combination. Grid search is an efficient method for hyperparameter tuning which avoids blind trial-and-error by defining the ranges for optimal hyperparameters based on MSE, MAE, and RMSE values.

The hyperparameter ranges for the two models used in this study, MLP and LSTM, are shown in Tables 3 and 4.

In MLP, the activation function mainly determines the nonlinear conversion method used to generate neuron outputs, the weight optimization reduces loss or objective functions to find the best parameter combination in the training model, the complexity prevents overfitting (higher values represent stronger regularization and lower model complexity), and the initial learning rate determines the step size when updating weights. Smaller learning rates require more iterations until convergence, but large learning rates may lead to

unstable models during training.

In LSTM, the optimizer is mainly used to adjust weights, reduce loss functions, and lower dropout rates. The dropout rate of the dropout layer affects model robustness and learning ability. The batch size is used to specify the number of training samples.

2.7. Model evaluation criteria

A number of criteria are available to compare the accuracy and performance of different AI prediction models. The forecasting performance of different prediction models can easily be affected by factors such as timeline scales, model parameters, and local climate conditions. Therefore, it is necessary to compare the performance of different prediction models through common performance evaluation indicators, namely, errors between predicted and actual values [39]. A study released by Robert Blaga et al., in 2019 [52] stated that most solar photovoltaic forecasting studies usually use mean bias error (MBE), MAE, and root mean square error (RMSE) as accuracy and performance evaluation indicators, but some other studies [53–55] used the coefficient of determination (R^2) as an indicator of model accuracy. This study therefore used all four of these indicators to evaluate model performance. Indicator definitions are as follows:

Mean Bias Error (MBE): Calculated by dividing the sum of the differences between predicted and actual values by the number of samples to determine the overall deviations between predicted values and actual values. Positive MBE values indicate overly high predicted values while negative MBE values indicate overly low predicted values, and MBE values close to zero indicate low deviation between predicted and actual values. The calculation formula is as follows [56]:

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{21}$$

n is the sample size, $\hat{y}_i$ is the predicted value, and y_i is the actual value

Mean Absolute Error (MAE): Calculated by taking the average value of the absolute error between model predicted values and actual values. MAE is used to determine the average deviation between predicted and actual values. Smaller MAE values mean that model predictions are more accurate. The calculation formula is as follows [56]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{22}$$

Root mean square error (RMSE): Calculated by taking the square root of the difference between predicted and actual values. The main purpose of doing this is to amplify the difference between predicted and actual values, and avoid offsetting differences in positive and negative values. Smaller RMSE values indicate smaller deviations between predicted and actual values, meaning that model predictions are more accurate. Compared with MAE, RMSE gives higher weight to predicted values with large extreme differences, so can better highlight model uncertainties in some situations. The calculation formula is as follows [56]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{23}$$

As this study collected solar photovoltaic data from sites all over Taiwan with different power generation scales, the data were normalized using the following formula to enable comparisons [52]:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{24}$$

The revised formulas for MBE, MAE, and RMSE were as follows:

$$nMBE = \frac{MBE}{\mu} \tag{25}$$

$$nMAE = \frac{MAE}{\mu} \tag{26}$$

$$nRMSE = \frac{RMSE}{\mu} \tag{27}$$

Coefficient of determination (R^2): A common statistical indicator used to evaluate prediction models which mainly measures the degree to which the prediction model explains the variability of the target variable, that is, the degree of similarity between the predicted value and the actual value. The range value of R^2 falls between 0 and 1. Values closer to 1 indicate that predicted values adhere closely to actual values. The calculation formula is as follows [57]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2} \tag{28}$$

$\bar{y}_i$ is the average value of $\hat{y}_i$.

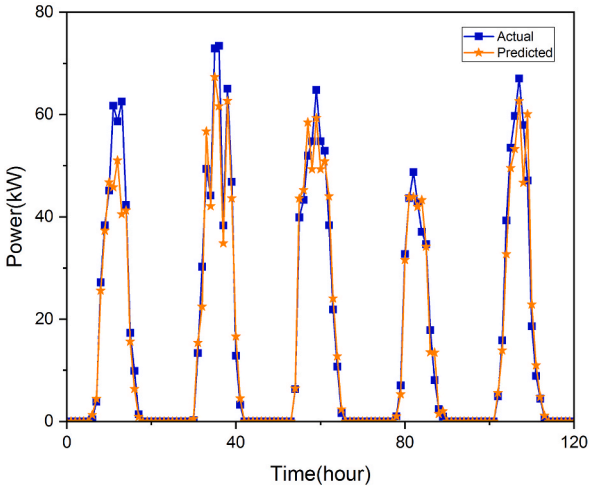


Fig. 11. Actual and predicted values of the MLP short-term (hourly) power generation prediction model without data pre-processing.

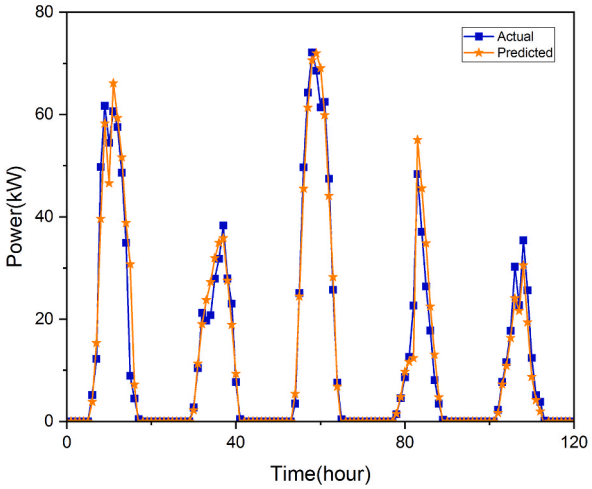


Fig. 12. Actual and predicted values of the LSTM short-term (hourly) power generation prediction model without data pre-processing.

3. Results

This study collected short-term (hourly) and medium-term (daily) rooftop solar photovoltaic system data from 17 Chunghwa Telecom sites distributed in the northern, central, and southern regions of Taiwan. MLP and LSTM were the AI methods used for building forecasting models. This study aimed to compare differences in model prediction performance with and without data pre-processing based on MBE, MAE, RMSE, and R^2 values. nRMSE values were used to evaluate model effectiveness in different areas of Taiwan to determine the best method for each region.

Figs. 11 and 12 show actual and predicted values at the N1 site when MLP and LSTM were used to build short-term (hourly) power generation prediction models without data pre-processing. Figs. 13 and 14 show actual and predicted values at the N1 site when MLP and LSTM were used to build short-term (hourly) power generation prediction models with data pre-processing. Table 5 shows that, without pre-processing, the R^2 value of the LSTM model is 90.61% and the R^2 value of the MLP model is 80.01%, indicating that the LSTM model yields better prediction performance. Data pre-processing reduced MBE, MAE, and RMSE values and increased R^2 values in both models, but the MLP model showed a higher level of improvement compared to the LSTM model.

Figs. 15 and 16 show actual and predicted values at the N2 site when MLP and LSTM were used to build medium-term (daily) power generation prediction models without data pre-processing. Figs. 17 and 18 show actual and predicted values at the N2 site when MLP and LSTM were used to build medium-term (daily) power generation prediction models with data pre-processing. Table 6 shows that, without pre-processing, the R^2 value of the LSTM model (89.02%) was higher than the R^2 value of the MLP model (80.11%). Data pre-processing reduced MBE, MAE, and RMSE values and increased R^2 values in both models, indicating that data pre-processing positively impacted AI model performance.

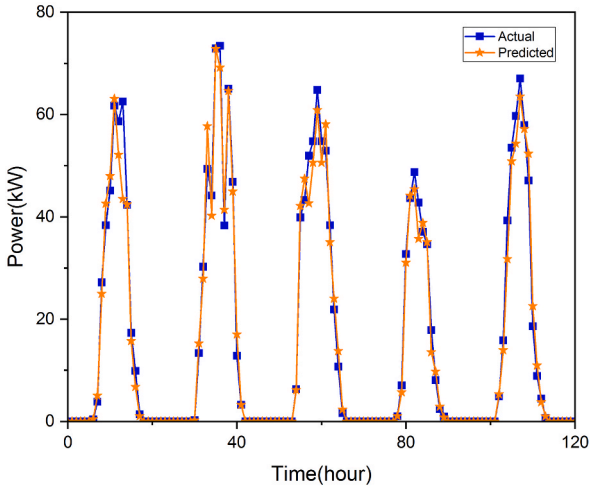


Fig. 13. Actual and predicted values of the MLP short-term (hourly) power generation prediction model with data pre-processing.

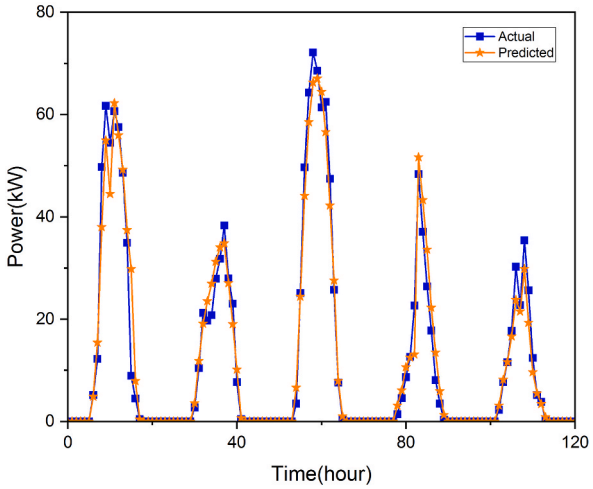


Fig. 14. Actual and predicted values of the LSTM short-term (hourly) power generation prediction model with data pre-processing.

Table 5
Evaluation parameters of MLP and LSTM short-term forecasting models at N1 site.

Method	MBE	MAE	RMSE	R^2 (%)
MLP	−0.27	4.28	6.71	80.01%
MLP + pre-processing	−0.25	2.14	4.13	93.32%
LSTM	0.69	3.39	4.62	90.61%
LSTM + pre-processing	0.56	2.34	3.89	93.35%

To verify the quality and repeatability of this study, the 17 Chunghwa Telecom sites were separated into three regions (northern, central, and southern regions) and the differences in AI prediction performance were compared using normalized MBE, MAE, and RMSE values (nMBE, nMAE, and nRMSE), which calculated the ratios of error values to observed value ranges (the difference between maximum and minimum values). Tables 7–9 show the calculation results for the northern, central, and southern regions, respectively. Irrespective of the region, data pre-processing positively improved the error values of the prediction models. A number of studies have used nRMSE to compare regression analysis results [58], and it can be seen that the nRMSE values for the LSTM model in this study were all lower than the MLP model, indicating that LSTM yields better prediction performance.

Without data pre-processing, LSTM yielded an average nRMSE of 5.00% in the northern region, 5.77% in the central region, and 4.73% in the southern region for short-term power generation forecasting, and an average nRMSE of 8.82% in the northern region, 8.35% in the central region, and 9.28% in the southern region for medium-term power generation forecasting. MLP yielded an average

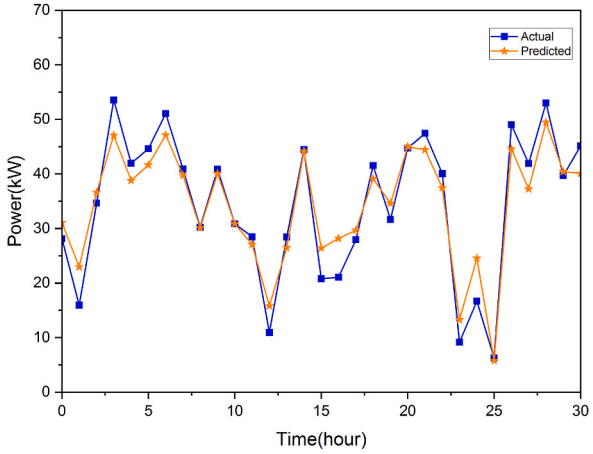


Fig. 15. Actual and predicted values of the MLP medium-term (daily) power generation prediction model without data pre-processing.

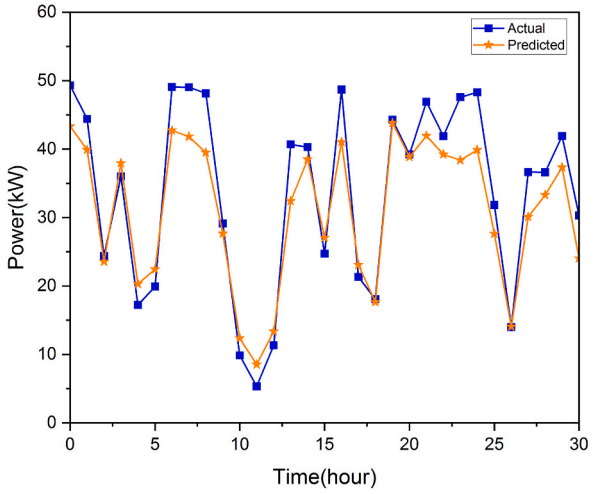


Fig. 16. Actual and predicted values of the LSTM medium-term (daily) power generation prediction model without data pre-processing.

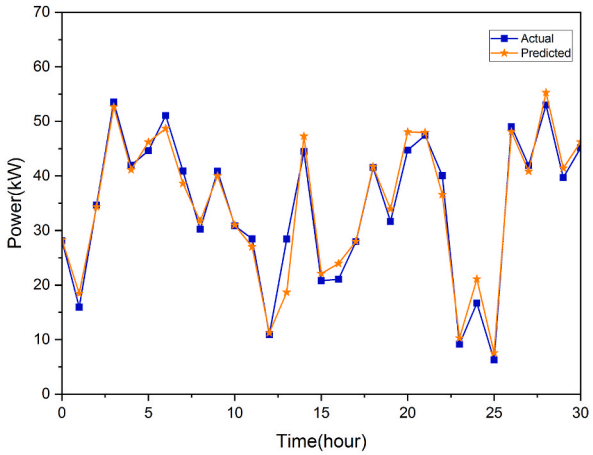


Fig. 17. Actual and predicted values of the MLP medium-term (daily) power generation prediction model with data pre-processing.

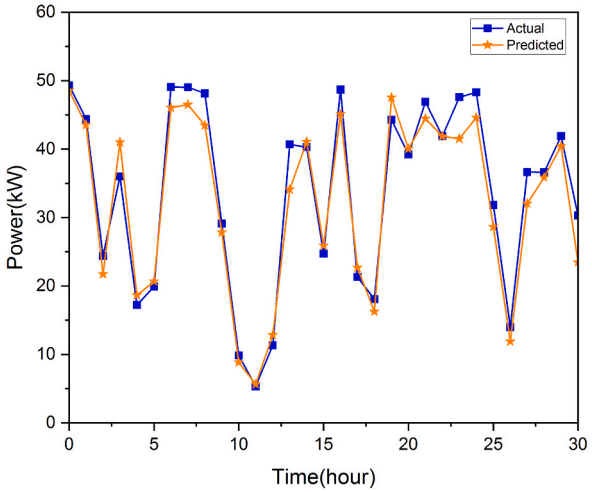


Fig. 18. Actual and predicted values of the LSTM medium-term (daily) power generation prediction model with data pre-processing.

Table 6
Evaluation parameters of MLP and LSTM medium-term forecasting models at N2 site.

Method	MBE	MAE	RMSE	R^2 (%)
MLP	−0.57	4.15	6.69	80.11%
MLP + pre-processing	−0.45	2.39	4.37	92.53%
LSTM	−1.38	3.67	5.00	89.02%
LSTM + pre-processing	0.05	2.36	3.86	93.44%

Table 7
Short- and medium-term power generation forecasting performance in the northern region.

Site	Time	Model	nMBE	nMAE	nRMSE
N1	1h	MLP	0.67%	12.88%	30.57%
		MLP + pre-processing	0.56%	10.18%	23.59%
		LSTM	−2.34%	12.65%	3.65%
		LSTM + pre-processing	−1.91%	11.02%	3.41%
	1day	MLP	−0.36%	6.47%	8.35%
		MLP + pre-processing	−0.15%	5.72%	7.57%
		LSTM	4.94%	7.43%	6.27%
		LSTM + pre-processing	2.59%	6.68%	5.92%
N2	1h	MLP	−0.66%	21.10%	45.79%
		MLP + pre-processing	−0.36%	12.33%	29.14%
		LSTM	−15.21%	20.75%	6.06%
		LSTM + pre-processing	5.28%	16.02%	4.09%
	1day	MLP	−1.70%	12.48%	20.10%
		MLP + pre-processing	−1.33%	7.10%	12.97%
		LSTM	−4.39%	11.65%	8.66%
		LSTM + pre-processing	0.15%	7.51%	6.70%
N3	1h	MLP	−0.23%	13.15%	27.83%
		MLP + pre-processing	−0.16%	13.06%	27.82%
		LSTM	−0.05%	0.14%	0.04%
		LSTM + pre-processing	0.04%	0.12%	0.04%
	1day	MLP	−0.06%	6.04%	8.06%
		MLP + pre-processing	−0.05%	5.77%	7.86%
		LSTM	4.49%	7.00%	6.17%
		LSTM + pre-processing	1.07%	6.11%	5.65%
N4	1h	MLP	−2.39%	21.99%	27.29%
		MLP + pre-processing	−2.24%	20.78%	26.70%
		LSTM	−19.19%	22.98%	14.59%
		LSTM + pre-processing	−14.93%	20.01%	12.79%
	1day	MLP	−2.39%	21.99%	27.29%
		MLP + pre-processing	−2.24%	20.78%	26.70%
		LSTM	−19.19%	22.98%	14.59%
		LSTM + pre-processing	−14.93%	20.01%	12.79%

Table 8
Short- and medium-term power generation forecasting performance in the central region.

Site	Time	Model	nMBE	nMAE	nRMSE
M1	1h	MLP	0.98%	11.83%	16.51%
		MLP + pre-processing	−0.40%	9.33%	14.56%
		LSTM	−10.33%	15.00%	9.41%
		LSTM + pre-processing	−8.89%	14.39%	9.15%
	1day	MLP	−1.17%	28.79%	46.19%
		MLP + pre-processing	−0.69%	20.36%	32.91%
		LSTM	−43.39%	50.03%	17.67%
		LSTM + pre-processing	−39.02%	48.43%	16.84%
M2	1h	MLP	−0.66%	5.06%	7.28%
		MLP + pre-processing	−0.53%	4.05%	6.23%
		LSTM	2.86%	6.56%	6.19%
		LSTM + pre-processing	0.69%	6.41%	6.14%
	1day	MLP	−0.14%	3.71%	8.04%
		MLP + pre-processing	−0.08%	3.36%	6.51%
		LSTM	0.44%	4.54%	2.17%
		LSTM + pre-processing	−0.77%	4.27%	1.55%
M3	1h	MLP	−0.84%	5.57%	7.76%
		MLP + pre-processing	−0.72%	4.94%	7.57%
		LSTM	8.50%	9.78%	4.36%
		LSTM + pre-processing	3.72%	7.37%	4.01%
	1day	MLP	0.26%	11.62%	21.11%
		MLP + pre-processing	0.19%	10.92%	20.11%
		LSTM	0.44%	8.82%	7.88%
		LSTM + pre-processing	−0.46%	7.65%	9.53%
M4	1h	MLP	−0.46%	7.65%	9.53%
		MLP + pre-processing	0.20%	7.21%	9.09%
		LSTM	4.91%	8.80%	8.01%
		LSTM + pre-processing	1.28%	8.15%	7.41%
	1day	MLP	−0.46%	7.65%	9.53%
		MLP + pre-processing	0.20%	7.21%	9.09%
		LSTM	4.91%	8.80%	8.01%
		LSTM + pre-processing	1.28%	8.15%	7.41%
M5	1h	MLP	−0.13%	8.06%	10.70%
		MLP + pre-processing	0.11%	7.47%	10.26%
		LSTM	9.99%	12.43%	10.00%
		LSTM + pre-processing	6.85%	11.50%	9.28%
	1day	MLP	0.34%	11.01%	20.34%
		MLP + pre-processing	0.11%	8.12%	16.09%
		LSTM	8.64%	12.13%	4.59%
		LSTM + pre-processing	7.04%	11.82%	4.18%

nRMSE of 41.64% in the northern region, 22.93% in the central region, and 17.68% in the southern region for short-term power generation forecasting, and an average nRMSE of 15.95% in the northern region, 10.6% in the central region, and 14.04% in the southern region for medium-term power generation forecasting. These results show that LSTM yields better prediction performance compared to MLP in all regions.

Table 10 shows the prediction results of the two models with data pre-processing. nRMSE values were reduced for both models. For LSTM, the nRMSE for short-term power generation forecasting was reduced by 20.02% and the nRMSE for medium-term power generation forecasting was reduced by 8.03%. For MLP, the nRMSE for short-term power generation forecasting was reduced by 17.47% and the nRMSE for medium-term power generation forecasting was reduced by 11.06%. These results show that the data pre-processing method proposed by this study significantly improved the short- and medium-term solar photovoltaic prediction of the LSTM and MLP models.

4. Conclusion

The solar photovoltaic industry has advanced rapidly in recent years, and the costs of solar panels, inverters, and other related components have decreased, making it possible for more solar photovoltaic devices and sites to be installed.

Traditional solar photovoltaic stations require large pieces of land, but Taiwan, unlike other countries, is a small and densely populated country. This, coupled with the Taiwanese government's 2050 net-zero emissions policy, has encouraged large-scale companies to actively participate in the renewable energy industry, and rooftop solar photovoltaic systems are receiving increasing attention as a result. Rooftop solar photovoltaic systems are limited by roof area sizes but are easy to install, so can be installed on top of offices, warehouses, and computer rooms to raise total installed capacity. Rooftop solar photovoltaic systems can also be integrated with power grids, so it is very important for power companies to predict future power generation volumes. Many previous studies have proposed a variety of methods for predicting power generation in traditional solar power plants using AI, but there have been few studies on solar photovoltaic systems. The results of this study can therefore help to bridge the knowledge gap in this field.

Table 9
Short- and medium-term power generation forecasting performance in the southern region.

Site	Time	Model	nMBE	nMAE	nRMSE
S1	1h	MLP	−0.12%	6.20%	14.30%
		MLP + pre-processing	0.05%	3.96%	8.22%
		LSTM	−1.34%	10.11%	5.62%
		LSTM + pre-processing	−0.89%	4.99%	1.84%
	1day	MLP	0.90%	8.88%	16.18%
		MLP + pre-processing	0.53%	7.99%	13.90%
		LSTM	−1.81%	8.64%	9.57%
		LSTM + pre-processing	−0.46%	7.76%	9.16%
S2	1h	MLP	0.06%	5.26%	12.43%
		MLP + pre-processing	0.03%	4.27%	8.51%
		LSTM	5.47%	7.52%	3.22%
		LSTM + pre-processing	1.15%	5.94%	2.46%
	1day	MLP	−0.52%	4.57%	9.56%
		MLP + pre-processing	−0.46%	4.43%	9.14%
		LSTM	6.69%	7.90%	8.54%
		LSTM + pre-processing	0.64%	5.09%	7.36%
S3	1h	MLP	−0.36%	13.72%	30.08%
		MLP + pre-processing	−0.40%	8.86%	16.55%
		LSTM	−2.64%	18.57%	5.55%
		LSTM + pre-processing	−1.51%	12.25%	3.35%
	1day	MLP	0.56%	14.38%	19.51%
		MLP + pre-processing	0.18%	12.99%	18.19%
		LSTM	1.76%	14.27%	12.29%
		LSTM + pre-processing	0.54%	13.54%	12.25%
S4	1h	MLP	0.42%	9.99%	30.31%
		MLP + pre-processing	0.08%	9.39%	25.56%
		LSTM	9.11%	19.25%	9.79%
		LSTM + pre-processing	5.44%	9.66%	5.11%
	1day	MLP	−1.96%	10.08%	17.69%
		MLP + pre-processing	−1.83%	9.47%	16.74%
		LSTM	−14.80%	16.60%	10.73%
		LSTM + pre-processing	−12.40%	14.31%	9.56%
S5	1h	MLP	−0.29%	5.43%	12.21%
		MLP + pre-processing	−0.15%	4.24%	11.68%
		LSTM	6.39%	8.12%	3.78%
		LSTM + pre-processing	4.59%	7.32%	3.59%
	1day	MLP	−0.32%	3.86%	4.84%
		MLP + pre-processing	−0.32%	3.68%	4.62%
		LSTM	4.68%	6.52%	7.58%
		LSTM + pre-processing	0.02%	5.44%	6.52%
S6	1h	MLP	0.08%	5.14%	11.26%
		MLP + pre-processing	0.01%	4.92%	11.08%
		LSTM	15.49%	15.61%	4.80%
		LSTM + pre-processing	7.00%	9.12%	3.60%
	1day	MLP	−1.16%	11.76%	17.42%
		MLP + pre-processing	−0.78%	4.91%	10.18%
		LSTM	11.64%	16.97%	9.86%
		LSTM + pre-processing	5.80%	9.55%	7.57%
S7	1h	MLP	−0.24%	7.68%	17.55%
		MLP + pre-processing	−0.18%	6.72%	16.87%
		LSTM	6.08%	10.89%	3.05%
		LSTM + pre-processing	1.71%	7.08%	2.89%
	1day	MLP	0.36%	7.72%	15.72%
		MLP + pre-processing	0.14%	7.20%	14.08%
		LSTM	−1.90%	9.05%	8.57%
		LSTM + pre-processing	1.43%	8.75%	8.51%
S8	1h	MLP	−0.19%	5.57%	13.30%
		MLP + pre-processing	−0.16%	4.98%	11.77%
		LSTM	−6.37%	9.62%	2.02%
		LSTM + pre-processing	−2.54%	6.34%	1.57%
	1day	MLP	0.32%	5.64%	11.40%
		MLP + pre-processing	−0.27%	5.36%	9.82%
		LSTM	4.13%	5.89%	7.12%
		LSTM + pre-processing	0.16%	5.03%	6.64%

In addition to choosing suitable prediction techniques, data pre-processing before training can also significantly affect model performance. This study proposed a data pre-processing method to improve the accuracy and reliability of AI prediction models. Generally, data pre-processing uses a single method (for example, regression analysis [59] or wavelet analysis [29]) to pre-process

Table 10
nRMSE improvement ratios at each site after data pre-processing.

Site	Short term		Medium term	
	MLP	LSTM	MLP	LSTM
N1	22.84%	6.61%	9.25%	5.55%
N2	36.36%	32.51%	2.44%	0.11%
N3	0.05%	4.14%	35.45%	22.71%
N4	1.66%	6.23%	2.15%	12.36%
M1	28.74%	4.71%	11.84%	2.68%
M2	19.09%	28.70%	14.50%	0.73%
M3	4.76%	8.10%	2.54%	3.58%
M4	6.62%	7.30%	4.61%	7.56%
M5	20.87%	9.01%	4.11%	7.16%
S1	42.50%	67.31%	14.11%	4.28%
S2	31.55%	23.69%	4.30%	13.83%
S3	45.00%	39.59%	6.77%	0.36%
S4	15.69%	47.86%	5.42%	10.90%
S5	4.36%	5.11%	4.60%	13.89%
S6	1.54%	25.13%	41.59%	23.21%
S7	3.88%	5.27%	10.43%	0.72%
S8	11.50%	22.16%	13.86%	6.83%
Average	17.47%	20.20%	11.06%	8.03%

different types of data. However, in consideration of the different attributes for the weather and power generation data collected in this study, two different techniques (linear regression and k-NN) were used to pre-process these two types of data for enhanced accuracy.

This study used data provided by Chunghwa Telecom, a large-scale national telecommunications company which has installed rooftop solar photovoltaic sites on offices, warehouses, and computer rooms all over Taiwan. Due to the high density of these sites, the k-NN method can be used to estimate missing values in a weighted manner based on the values of neighboring sites to increase the accuracy of missing data.

Study results indicated that LSTM performed better than MLP when data pre-processing was not applied: short-term power generation predictions for N1 produced by these two models yielded a difference of 10.6% in R^2 values and medium-term power generation predictions for N2 produced by these two models yielded a difference of 8.91% in R^2 values. After data pre-processing, R^2 values for short-term power generation predictions at N1 increased by 13.31% and 2.74% for MLP and LSTM, respectively, and R^2 values for medium-term power generation predictions at N2 increased by 12.42% and 4.42% for MLP and LSTM, respectively. When prediction scope was expanded to 17 rooftop solar photovoltaic sites across Taiwan, data pre-processing reduced the nRMSE of the LSTM model by 20.20% for short-term power generation forecasting and by 8.03% for medium-term power generation forecasting. Data pre-processing also reduced the nRMSE of the MLP model by 17.47% for short-term power generation forecasting and by 11.06% for medium-term power generation forecasting.

CRediT authorship contribution statement

Da-Sheng Lee: Conceptualization. **Chih-Wei Lai:** Project administration, Methodology, Investigation, Formal analysis. **Shih-Kai Fu:** Investigation, Formal analysis, Data curation.

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.

Acknowledgements

We would like to express our heartfelt gratitude to Chunghwa Telecom for providing historical data from 17 rooftop solar photovoltaic stations for our research. Their selfless assistance allowed us to conduct our study and complete our work. We also want to thank our colleagues at National Taipei University of Technology (Taiwan) for their helpful feedback and support. We are also grateful to JB Consulting & Co. for editing and proofreading this manuscript. Finally, we would like to thank our family and friends for their love and support throughout the research process. Without their encouragement and support, we would not have been able to complete this research.

References

[1] N. Kannan, D. Vakeesan, Solar energy for future world:-A review, Renew. Sustain. Energy Rev. 62 (2016) 1092–1105 (Solar energy for future world:-A review).
[2] IRENA, Renewable Energy Statistics 2020, International renewable energy agency, Abu Dhabi, 2020.

[3] K. Obaideen, A.G. Olabi, Y. Al Swailmeen, N. Shehata, M.A. Abdelkareem, A.H. Alami, C. Rodriguez, E.T. Sayed, Solar energy: applications, trends analysis, bibliometric analysis and research contribution to sustainable development goals (SDGs), Sustainability 15 (2) (2023) 1418.
[4] S. Comello, S. Reichelstein, A. Sahoo, The road ahead for solar PV power, Renew. Sustain. Energy Rev. 92 (2018) 744–756.
[5] R. Singh, A.Y. Nam, J.J. Park, Y.I. Kim, Analysis of in situ performance of rooftop PV system in Seoul, South Korea, International Journal of Air-Conditioning and Refrigeration 31 (1) (2023) 10.
[6] U. Agarwal, N.S. Rathore, N. Jain, P. Sharma, R.C. Bansal, M. Chouhan, M. Kumawat, Adaptable pathway to net zero carbon: a case study for Techno-Economic & Environmental assessment of Rooftop Solar PV System, Energy Rep. 9 (2023) 3482–3492.
[7] M.M. Akrofi, M. Okitasari, Beyond costs: how urban form could limit the uptake of residential solar PV systems in low-income neighborhoods in Ghana, Energy for Sustainable Development 74 (2023) 20–33.
[8] Y. Jing, L. Zhu, B. Yin, F. Li, Evaluating the PV system expansion potential of existing integrated energy parks: a case study in North China, Appl. Energy 330 (2023) 120310.
[9] P.K.S. Rathore, S. Rathore, R.P. Singh, S. Agnihotri, Solar power utility sector in India: challenges and opportunities, Renew. Sustain. Energy Rev. 81 (2018) 2703–2713.
[10] A.M. Pringle, R.M. Handler, J.M. Pearce, Aquavoltaics: synergies for dual use of water area for solar photovoltaic electricity generation and aquaculture, Renew. Sustain. Energy Rev. 80 (2017) 572–584.
[11] H.B. Tambunan, A.P. Purnomoadi, P.A. Pramana, B.B.S. Harsono, A.S. Surya, A.S. Habibie, Performance of ground mounted PV system affected by near shadings losses, in: 2020 2nd International Conference on Industrial Electrical and Electronics (ICIEE), IEEE, 2020, pp. 46–51.
[12] J. Schardt, H. te Heesen, Performance of roof-top PV systems in selected European countries from 2012 to 2019, Sol. Energy 217 (2021) 235–244.
[13] J. Peng, L. Lu, Investigation on the development potential of rooftop PV system in Hong Kong and its environmental benefits, Renew. Sustain. Energy Rev. 27 (2013) 149–162.
[14] A.C. Duman, Ö. Güler, Economic analysis of grid-connected residential rooftop PV systems in Turkey, Renew. Energy 148 (2020) 697–711.
[15] M. Emziane, M. Al Ali, Performance assessment of rooftop PV systems in Abu Dhabi, Energy Build. 108 (2015) 101–105.
[16] A. Afzal, A. Buradi, R. Jilte, S. Shaik, A.R. Kaladgi, M. Arıcı, C.T. Lee, S. Nižetić, Optimizing the thermal performance of solar energy devices using meta-heuristic algorithms: a critical review, Renew. Sustain. Energy Rev. 173 (2023) 112903.
[17] U.K. Das, K.S. Tey, M. Seyedmahmoudian, S. Mekhilef, M.Y.I. Idris, W.V. Deventer, B. Horan, A. Stojcevski, Forecasting of photovoltaic power generation and model optimization: a review, Renew. Sustain. Energy Rev. 81 (2018) 912–928.
[18] R. Ahmed, V. Sreeram, Y. Mishra, M.D. Arif, A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization, Renew. Sustain. Energy Rev. 124 (2020) 109792.
[19] A.H. Elsheikh, S.W. Sharshir, M. Abd Elaziz, A.E. Kabeel, W. Guilan, Z. Haiou, Modeling of solar energy systems using artificial neural network: a comprehensive review, Sol. Energy 180 (2019) 622–639.
[20] A. Mellit, A.M. Pavan, A 24-h forecast of solar irradiance using artificial neural network: application for performance prediction of a grid-connected PV plant at Trieste, Italy, Sol. Energy 84 (5) (2010) 807–821.
[21] F. Rodríguez, M. Genn, L. Fontán, A. Galarza, Very short-term temperature forecaster using MLP and N-nearest stations for calculating key control parameters in solar photovoltaic generation, Sustain. Energy Technol. Assessments 45 (2021) 101085.
[22] H. Zhou, Q. Liu, K. Yan, Y. Du, Deep learning enhanced solar energy forecasting with AI-driven IoT, Wireless Commun. Mobile Comput. 2021 (2021) 1–11.
[23] Q. Liu, Q.J. Zhang, Accuracy improvement of energy prediction for solar-energy-powered embedded systems, IEEE Trans. Very Large Scale Integr. Syst. 24 (6) (2015) 2062–2074.
[24] D. Huang, C. Zhang, Q. Li, H. Han, D. Huang, T. Li, C. Wang, Prediction of solar photovoltaic power generation based on MLP and LSTM neural networks, in: 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2), IEEE, 2020.
[25] J.M. Barrera, A. Reina, A. Maté, J.C. Trujillo, Solar energy prediction model based on artificial neural networks and open data, Sustainability 12 (17) (2020) 6915.
[26] Q.T. Phan, Y.K. Wu, Q.D. Phan, H.Y. Lo, A novel forecasting model for solar power generation by a deep learning framework with data preprocessing and postprocessing, IEEE Trans. Ind. Appl. 59 (1) (2022) 220–231.
[27] C. Paoli, C. Voyant, M. Muselli, M.L. Nivet, Forecasting of preprocessed daily solar radiation time series using neural networks, Sol. Energy 84 (12) (2010) 2146–2160.
[28] J.C. Cao, S.H. Cao, Study of forecasting solar irradiance using neural networks with preprocessing sample data by wavelet analysis, Energy 31 (15) (2006) 3435–3445.
[29] J. Zhang, Y. hi, L. Xiao, Solar power generation forecast based on LSTM, in: 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 869–872. IEEE.
[30] C. Chen, S. Duan, T. Cai, B. Liu, Online 24-h solar power forecasting based on weather type classification using artificial neural network, Sol. Energy 85 (11) (2011) 2856–2870.
[31] Central weather Bureau, MOTC, climate questions, URL:https://www.cwb.gov.tw/V8/C/K/Encyclopedia/climate/index.html(2023/8/1, , 2023.
[32] I. Gad, B.R. Manjunatha, Performance evaluation of predictive models for missing data imputation in weather data, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2017, pp. 1327–1334.
[33] W. Lee, K. Kim, J. Park, J. Kim, Y. Kim, Forecasting solar power using long-short term memory and convolutional neural networks, IEEE Access 6 (2018) 73068–73080.
[34] H. Zang, L. Cheng, T. Ding, K.W. Cheung, M. Wang, Z. Wei, G. Sun, Application of functional deep belief network for estimating daily global solar radiation: a case study in China, Energy 191 (2020) 116502.
[35] D. Chakraborty, J. Mondal, H.B. Barua, A. Bhattacharjee, Computational solar energy–Ensemble learning methods for prediction of solar power generation based on meteorological parameters in Eastern India, Renewable Energy Focus 44 (2023) 277–294.
[36] H. Zhou, Z. Deng, Y. Xia, M. Fu, A new sampling method in particle filter based on Pearson correlation coefficient, Neurocomputing 216 (2016) 208–215.
[37] G. Alkhayat, R. Mehmood, A review and taxonomy of wind and solar energy forecasting methods based on deep learning, Energy and AI 4 (2021) 100060.
[38] S. Sinsomboonthong, Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification, Int. J. Math. Math. Sci. 2022 (2022).
[39] C. Voyant, G. Notton, S. Kalogirou, M.L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: a review, Renew. Energy 105 (2017) 569–582.
[40] O. El Alani, M. Abraim, H. Ghennioui, A. Ghennioui, I. Ikenbi, F.E. Dahr, Short term solar irradiance forecasting using sky images based on a hybrid CNN–MLP model, Energy Rep. 7 (2021) 888–900.
[41] A. Ahmad, T.N. Anderson, T.T. Lie, Hourly global solar irradiation forecasting for New Zealand, Sol. Energy 122 (2015) 1398–1408.
[42] M.M. Rahman, M. Shakeri, S.K. Tiong, F. Khatun, N. Amin, J. Pasupuleti, M.K. Hasan, Prospective methodologies in hybrid renewable energy systems for energy prediction using artificial neural networks, Sustainability 13 (4) (2021) 2393.
[43] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, P.M. Atkinson, A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification, ISPRS J. Photogrammetry Remote Sens. 140 (2018) 133–144.
[44] S. Osowski, K. Siwek, T. Markiewicz, MLP and SVM networks-a comparative study, in: Proceedings of the 6th Nordic Signal Processing Symposium, 2004. NORSIG 2004, IEEE, 2004.
[45] F. Wang, Z. Xuan, Z. Zhen, K. Li, T. Wang, M. Shi, A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework, Energy Convers. Manag. 212 (2020) 112766.
[46] I. Jamil, H. Lucheng, S. Iqbal, M. Aurangzaib, R. Jamil, H. Kotb, A. Alkuhayli, K.M. AboRas, Predictive evaluation of solar energy variables for a large-scale solar power plant based on triple deep learning forecast models, Alex. Eng. J. 76 (2023) 51–73.

[47] S.C. Lim, J.H. Huh, S.H. Hong, C.Y. Park, J.C. Kim, Solar power forecasting using CNN-LSTM hybrid model, Energies 15 (21) (2022) 8233.
[48] A. Rai, A. Shrivastava, K.C. Jana, Differential attention net: multi-directed differential attention based hybrid deep learning model for solar power forecasting, Energy 263 (2023) 125746.
[49] T.M.L. Al-Jaafreh, A. Al-Odienat, Y.A. Altaharwah, The solar energy forecasting using LSTM deep learning technique, in: 2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA), IEEE, 2022, pp. 1–6.
[50] Al-Ali, M. Elham, et al., Solar energy production forecasting based on a hybrid CNN-LSTM-Transformer model, Mathematics 11 (3) (2023) 676.
[51] L.D. Bui, N.Q. Nguyen, B.V. Doan, E.R. Sanseverino, Forecasting energy output of a solar power plant in curtailment condition based on LSTM using P/GHI coefficient and validation in training process, a case study in Vietnam, Elec. Power Syst. Res. 213 (2022) 108706.
[52] R. Blaga, A. Sabadus, N. Stefu, C. Dughir, M. Paulescu, V. Badescu, A current perspective on the accuracy of incoming solar energy forecasting, Prog. Energy Combust. Sci. 70 (2019) 119–144.
[53] M. Elsaraiti, A. Merabet, Solar power forecasting using deep learning techniques, IEEE Access 10 (2022) 31692–31698.
[54] H.Y. Su, T.Y. Liu, H.H. Hong, Adaptive residual compensation ensemble models for improving solar energy generation forecasting, IEEE Trans. Sustain. Energy 11 (2) (2019) 1103–1105.
[55] N. Rahimi, S. Park, W. Choi, B. Oh, S. Kim, Y. Cho, S. Ahn, C. Chong, D. Kim, C. Jin, D. Lee, A comprehensive review on ensemble solar power forecasting algorithms, Journal of Electrical Engineering & Technology 18 (2) (2023) 719–733.
[56] Y. Natarajan, S. Kannan, C. Selvaraj, S.N. Mohanty, Forecasting energy generation in large photovoltaic plants using radial belief neural network, Sustainable Computing: Informatics and Systems 31 (2021) 100578.
[57] S.A. Haider, M. Sajid, S. Iqbal, Forecasting hydrogen production potential in islamabad from solar energy using water electrolysis, Int. J. Hydrogen Energy 46 (2) (2021) 1671–1681.
[58] I. Jebli, F.Z. Belouadha, M.I. Kabbaj, A. Tilioua, Prediction of solar energy guided by pearson correlation using machine learning, Energy 224 (2021) 120109.
[59] A. Bramm, S. Eroshenko, A. Khalyasmaa, Effect of data preprocessing on the forecasting accuracy of solar power plant, in: 2021 XVIII International Scientific Technical Conference Alternating Current Electric Drives (ACED), IEEE, 2021, pp. 1–5.