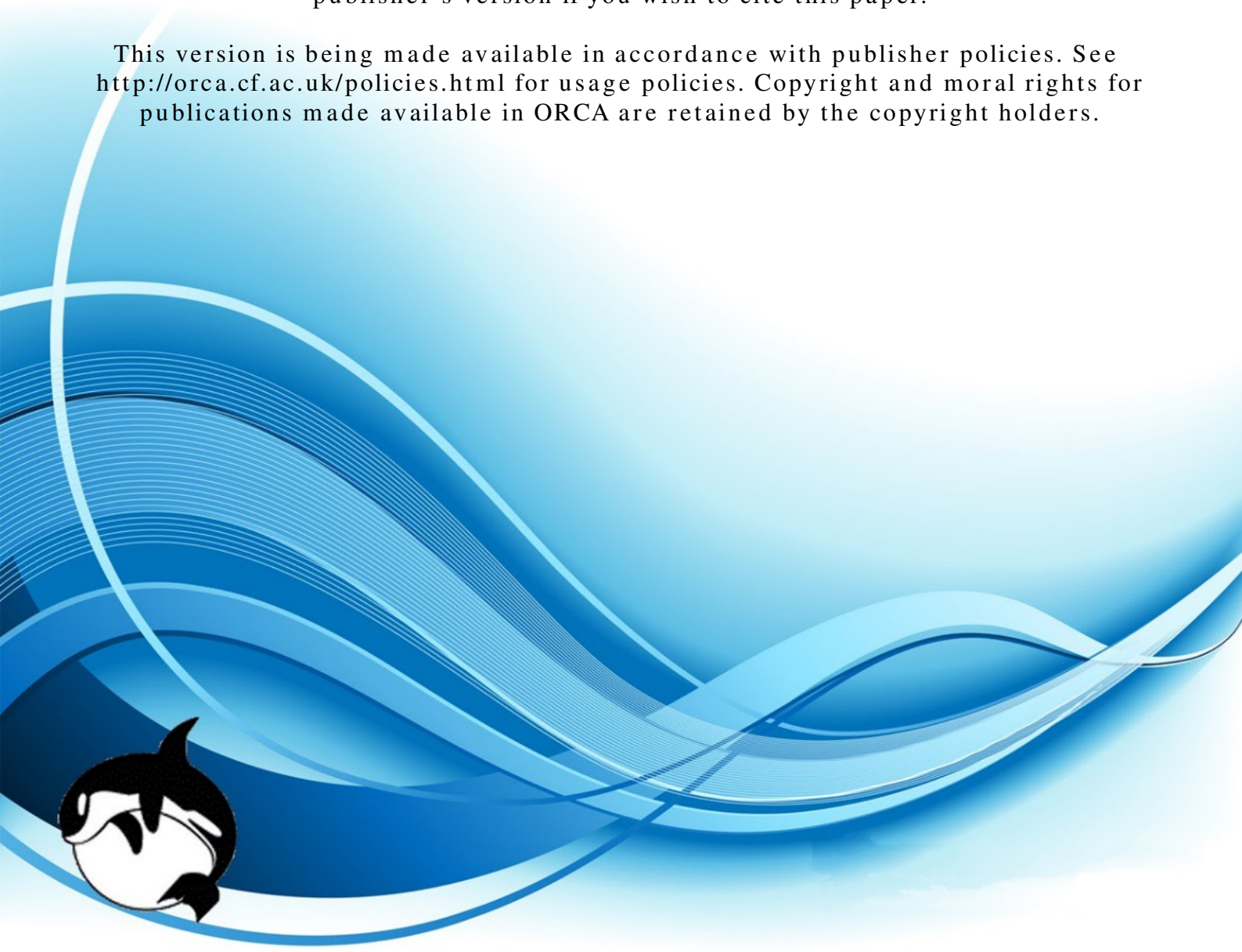


Live Power Generation Predictions via AI-Driven Resilient Systems in Smart Microgrids

Xueyi Wang, Shancang Li, and Muddesar Iqbal

Abstract—The 5G technology can significantly benefit smart consumer devices powered by microgrids in several ways, enhancing their efficiency, reliability, and overall performance, which play a pivotal role in advancing consumer electronics by providing a more reliable, efficient, and sustainable source of power for these devices. The growing environmental awareness and emergence of new technologies have made smart microgrids a good renewable and resilient power to serve consumer electronics. This work developed a secure AI-driven predictable and resilient power generation system for efficient microgrid energy use and management. Specifically, we first developed an intelligent power generation forecasting model based on a joint distribution of power generation and weather data; then, a resilient eXtreme Gradient Boosting (XGBoost) power generation forecast model was proposed that allows incorporating the weather intermittency in the joint distribution. The scheme has been validated using real-time power generation data together with weather data. The experimental results show that the proposed scheme can provide a more accurate and robust prediction of the microgrid against weather intermittency.

Index Terms—Smart Microgrid, Artificial Intelligence, Resilient and Sustainability, Power Prediction, XGBoost.

1 Introduction

The integration of cutting-edge technologies, including 5G and artificial intelligence (AI), empowers microgrids to significantly enhance consumer electronics in various ways. These advanced technologies contribute to a more dependable, efficient, and sustainable power supply for electronic devices, promoting their optimal performance [1]. Electricity plays a pivotal role in fostering economic development and advancing technological progress, particularly in the context of rapid consumer electronics [2], [3]. With the increasing demand for energy, electricity generation and distribution have been considered as the base of industrial and nation [4]. Among all electricity generation methods, renewable energy sources like wind, tide, and solar energy provide sufficient and clean energy. In particular, solar energy is one of the most stable and efficient renewable energy resources to generate electricity [5], [6]. The development of renewable energy reduces the dependency on traditional fossil and fuel resources, which are not only freely available all around the world, but also reduce carbon emissions whilst generating electricity [7].

According to the United Nations Development Programme (UNDP), solar energy resource has a worldwide potential of 1,600 to 49,800 exajoules (4.4×10^{14} to $1.4 \times 10^{16} kWh$) per year [8]. Considering the huge generating potential and environmental benefits of solar energy, solar photovoltaic (PV) panels have been widely installed [9]. The Chinese government continues to dominate both new and cumulative capacity, which added 106 GW in 2022 and reached 414.5 GW cumulative capacity. In 2022, at least 240 GW PV panels were installed worldwide, which made the total PV installed capacity reach at least 1185 GW [9].

As a part of smart production, the topics of smart grids, like energy prediction and cyber security have been meticulously studied recently [10], [11]. On the one hand, from the national grid perspective, with the Net Zero strategy by 2050, the UK's renewable energy took 42.1% of the total electricity generation in Quarter 2 2023 [12]. According to the UK Energy Trends Report 2023, 8.6% of total electricity generation is from solar PV generation and it was 6.3% in 2022, with the 1.1 GW solar PV capacity increasing [12]. On the other hand, from a microgrid perspective, many electricity storage systems like PV systems have been used to support electricity in individual houses and autonomous devices also design active generators [5], [13].

However, the fluctuating and uncertain output is still a significant drawback in PV energy systems and the concern of research [14], [15]. Facing the situation that the usage of PV panels sharply increasing and intermittency problems of PV generation in smart grid national-wide and microgrid individual use, an effective and resilient method estimating the electricity generation of the PV panel needs to be researched and developed eagerly [15]. The live predictive analytics can play a crucial role in addressing security concerns in smart microgrids, e.g., anomaly detection, early warning, incident response, behavioral analytics, threats intelligence integration, etc., enhancing the overall security and resilience of the smart microgrids. Depending on the predicting time span, PV generation prediction in smart microgrids can be divided into short-term prediction (under one day ahead), medium-term prediction (1 week to 1 month ahead), and long-term prediction (one month to one year ahead). Short-term predictions heavily rely on high-frequency factors such as changes in 24-hour weather conditions, cloud cover, and solar radiation. Medium-term predictions involve forecasting based on more generalized (1 week to 1 month) weather patterns and climate mod-

- *Wang and Li are with the School of Computer Science and Informatics at Cardiff University, Cardiff, CF24 4AG, UK.*
  Muddesar Iqbal is with College of Engineering, Prince Sultan University, Riyadh, Saudi Arabia.

els. Long-term predictions require consideration of seasonal variations and climate trends, which range from one month to a year. Overall, the various temporal input features decide different time durations of PV generation forecasting.

Short-term prediction can help in power dispatching, storage, and smoothing; medium-term prediction can help in power system management and scheduling; long-term prediction can help in grid device distribution and operation planning. For short-term prediction, statistical and numerical methods show advantages, however, for medium and long terms, physical models like numerical weather prediction, and sky/satellite image models have better performance [16]. Figure. 1 illustrates the varied time duration, highlighting the purposes and benefits of PV generation prediction.

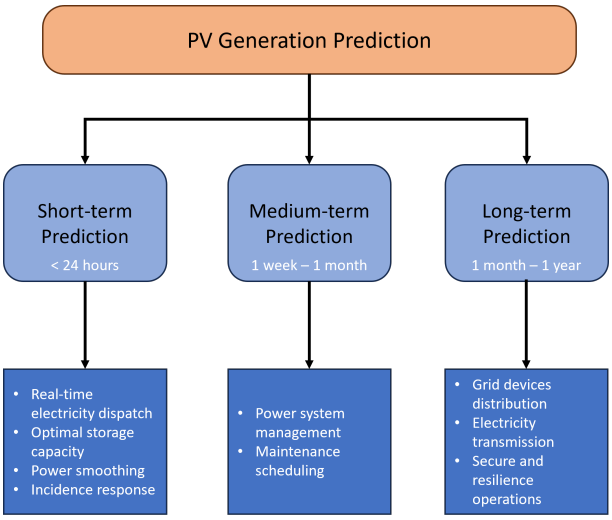


Fig. 1. Microgrid power generation prediction in different time duration

In this paper, we tried to improve the performance and resilience of short-term weather-based machine learning methods in PV electricity generation prediction. We used the locally collected PV generation and weather data to create the dataset and make the prediction based on that. The newly developed time-series concerned Machine Learning algorithms: Random Forest (RF) and XGBoost are compared with the traditional multi-variable Long Short-Term Memory model. The performances of the models are evaluated via R^2 and $nRMSE$ scores. The main contributions of this work can be summarised as follows:

1) Using machine learning models (RF, XGBoost, LSTM), integrated historical PV generation predictive models along with weather data were developed;

2) A short-term PV generation resilient model was developed by combining time-related data and historical PV generation data.

3) A use case using the MIDAS UK dataset and Sheffield PV live data is developed to demonstrate the proposed models.

2 Related Works

A host of reports [17], [18], [19] have delved into the intricate landscape of smart grids, particularly focusing on the application of deep learning algorithms. These algorithms have emerged as formidable tools, facilitating energy forecasting, bolstering security detection, and optimizing the management of smart grid operations. Moreover, they play a pivotal role in enhancing resiliency in the face of contingencies and addressing customer demands [17], [20].

The existing research and literature review the prevalence of three distinct methodological approaches. These approaches can be categorized as follows: 1) Statistical time series relies on historical PV generation data to construct predictive models for PV generation; 2) Physical models leverage Numerical Weather Prediction (NWP) data, sky images, or satellite images to craft models that predict PV generation patterns; 3) Machine learning approaches to harness the power of multi-variable weather data, utilizing machine learning algorithms to predict PV generation [21], [22].

Several reports have employed time-series algorithms to tackle the challenge of PV generation in short-term forecasting. Kardakos and his research team, for instance, leveraged the seasonal ARIMA time-series algorithm as a foundation for their predictions, augmenting its accuracy by incorporating solar radiation data derived from the Numerical Weather Prediction (NWP) model. This synergy between SARIMA and NWP enhanced the precision of their forecasts, which has an 11.12% annual basis in $nRMSE$ [23]. In a different approach, Maria Malvoni sought to forecast PV generation one day in advance, employing the Group Least Square Support Vector Machine (GLSSVM) time-series algorithm in conjunction with Least Square Support Vector Machines (LS-SVM) and the Group Method of Data Handling (GMDH) algorithm. These methods, designed to handle multiple weather variables, aimed to refine the prediction process [24]. On the other hand, a study by Hindawi concluded that, in the context of short-term PV generation prediction, ARIMA models outperformed Artificial Neural Networks (ANN) models [22]. In a similar vein, another study [21] conducted a comparative analysis of SARIMA, SARIMAX, modified SARIMA, and ANN algorithms in the context of short-term prediction of PV generation.

Some reports have delved into constructing purely physical models for PV generation prediction. In [25], the author mentioned that PV power fully depends on uncertain meteorological factors, like solar irradiance, temperature, wind direction, wind pressure, and humidity. In one such study [26], Sun proposed a method that captured instantaneous images around PV panels to monitor cloud movements, which could impact PV electricity output. This information was then analyzed using a Convolutional Neural Network (CNN) to predict PV electricity generation based on sky image analysis. Additionally, our previous work introduced the Plane of Array (POA) PV system model, developed by Sandia National Laboratories to predict PV generation [27]. Graditi *et al.* compared the physical PV prediction model (Schokley-Sandia) with the MLP model and regression model. Also, they mentioned the identification of minimum and representative training dataset selection methods.

For reports centered on predicting PV generation using multi-variable weather data through machine learning methods, Artificial Neural Network (ANN) models emerged as the preferred choice [22]. Stanley and his team, for instance, presented a short-term prediction strategy, fore-

casting 20 minutes ahead using the Multilayer Perceptron (MLP) model, which achieved an impressive PV generation prediction accuracy ranging from 82% to 95% [7]. Another study [28] explored four distinct models for short-term PV generation prediction: Multilayer Perceptron (MLP), Elman Recurrent Neural Network (ENN), Radial Basis Function neural network (RBF), and Time Delayed Neural Network (TDNN). Among these models, MLP exhibited the best performance, with an error rate of 0.62 in PV electricity generation prediction [28]. In [29], Empirical Mode Decomposition (EMD) and Support Vector Machine (SVM) methods were employed to analyze PV generation. SVM, a supervised machine learning model known for its prowess in generalized linear classification, featured prominently in the report. Notably, the report underscored that Artificial Neural Networks (ANN) and SVM were the two most frequently utilized prediction methods and emphasized the crucial role of daily temperature as a key weather factor influencing PV panel electricity output. Recently, Huawei Cloud published Pangu-Weather, which is an AI prediction model and has more competitive accuracy than traditional NWP methods. Pangu-weather utilizes a 3D transformer structure to capture spatial dependencies and many subresearches have been conducted based on the model [30] [31].

3 METHODOLOGY

3.1 Problem Formulation

Assume the forecast lead time is t, the power generation data $\mathbf{x}_{t:(t+i\cdot\Delta)}$ during Δ periods and the corresponding weather data is $\mathbf{w}_{t:(t+i\cdot\Delta)}$. The combined power generation $\mathbf{q}_{t:(t+i\cdot\Delta)}$ can be expressed with the conditional joint probability distribution

$$\mathbf{q}_{t:(t+i\cdot\Delta)} \sim p(\mathbf{x}_{t:(t+i\cdot\Delta)}|\mathbf{w}_{t:(t+i\cdot\Delta)}) \tag{1}$$

in which i is the number of samples. For sampling $i = N$ times, the power generation samples can be represented by

$$\mathbf{Q}_N = \{\mathbf{q}_{t:(t+N\cdot\Delta)}\} = \{(x_1, w_1), \ldots, (x_N, w_N)\} \tag{2}$$

Assume the combined power generation $\mathbf{Q}_N$ follow a prior distribution $p(\eta)$ and can be represented using a neural network as

$$\mathbf{q}_{t:(t+i\cdot\Delta)} = \mathcal{M}(\eta, \mathbf{x}_{t:(t+i\cdot\Delta)}, \theta) \sim p(\mathbf{x}_{t:(t+i\cdot\Delta)}|\mathbf{w}_{t:(t+i\cdot\Delta)}) \tag{3}$$

in which $\mathcal{M}$ is a generative models and θ is the learning parameter of the neural network. These parameters are learned during the training process to capture the underlying patterns and dependencies in the data.

We use $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ to denote $\mathbf{x}_{t:(t+i\cdot\Delta)}$ and $\mathbf{w}_{t:(t+i\cdot\Delta)}$, respectively. The power generation samples $\boldsymbol{x}_i \sim p(\boldsymbol{x}_i)$. Assuming the prior distribution $\boldsymbol{x}_i \sim \mathcal{N}(0, 1)$ is a multivariate Gaussian distribution. It represents a stochastic process for generating power generation samples over time. Then we have

$$\boldsymbol{x}_n = \sqrt{1-\beta_n}\boldsymbol{x}_{n-1} + \beta_n\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \tag{4}$$

in which $\beta \in (0, 1)$ is the scaling factor. The following equation specifies the conditional probability distribution of $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n-1}$ in the stochastic process can be described by

$$p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) = \mathcal{N}(\sqrt{1-\beta_n}\boldsymbol{x}_{n-1}, \beta_n\mathbf{1}) \tag{5}$$

The power generation samples can be described as

$$p(\boldsymbol{x}_n \mid \boldsymbol{x}_0) = \prod_n p(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}) \tag{6}$$

The weather distribution can be modeled using an auto-regressive (AR) integrated moving average (ARIMA) model, which includes three number (n_p, n_d, n_q), in which n_p denotes the number of auto-regressive terms, n_d denote the number of nonseasonal differences, and n_q denote the number of lagged forecast errors. Then the weather can be modeled as

$$w_t = \varphi_1 w_{t-1} + \cdots\cdots + \varphi_p w_{t-n_p} + e_t \tag{7}$$

in which w_t is the value at time t, φ_1 is AR coefficient, e_t is error value at time t.

The forecast error $\boldsymbol{e}$ is transformed into a posterior sample of $\boldsymbol{w}$. The forecast error $\boldsymbol{e}_n$ is obtained by subtracting the $\hat{\boldsymbol{e}}$ from the measured power $\boldsymbol{e}$. During the time Δ, the generated power can be modeled as

$$\boldsymbol{q} = \hat{\boldsymbol{x}} + \boldsymbol{e}_0 = f_d(z, \boldsymbol{x}, \theta) + f_c(w, \theta_e) \tag{8}$$

in which θ_d and θ_e are the learnable parameters of the network; $f_d(\cdot)$ and $f_e(\cdot)$ are the mapping functions of network, respectively.

As shown in Figure. 2, the proposed PV generation forecasting process consists following steps: 1) Data Alignment: PV generation data and local weather data are synchronized based on their date and time stamps; 2) Data Pre-processing: This phase begins with merging weather data and time-related data. Subsequently, the combined dataset undergoes pre-processing to eliminate any missing data and includes a filter to exclude hours of no electricity generation during nighttime. Additionally, historical PV generation data is incorporated into the dataset; 3) Prediction Models: Using Random Forest, XGBoost, and LSTM algorithms to predict PV generation in the short term. Furthermore, the study explores resilient models for PV generation prediction.

3.2 Random Forest

As an ensemble learning algorithm, the random forest (RF) is one of the powerful traditional machine learning algorithms that excels at making predictions for forecasting tasks. A random forest model is a meta estimator, that combines all the output predictions from multiple decision trees and calculates the average predicting values for the final output. All the classifying decision trees are trained and fitted on sub-samples from the training dataset. The max-sample parameter in the Random Forest model is to adjust the sub-sample size for training to prevent overfitting and improve model performance [32].

The core of the RF algorithm relies on its ability to combine the predictive strength of numerous decision trees. It achieves this through an ingenious process known as bootstrap aggregating. The RF algorithm starts by creating multiple random subsets from the original training dataset. These subsets are generated by randomly selecting data points with replacements. This process introduces diversity into the training data for each decision tree, making them distinct and robust.

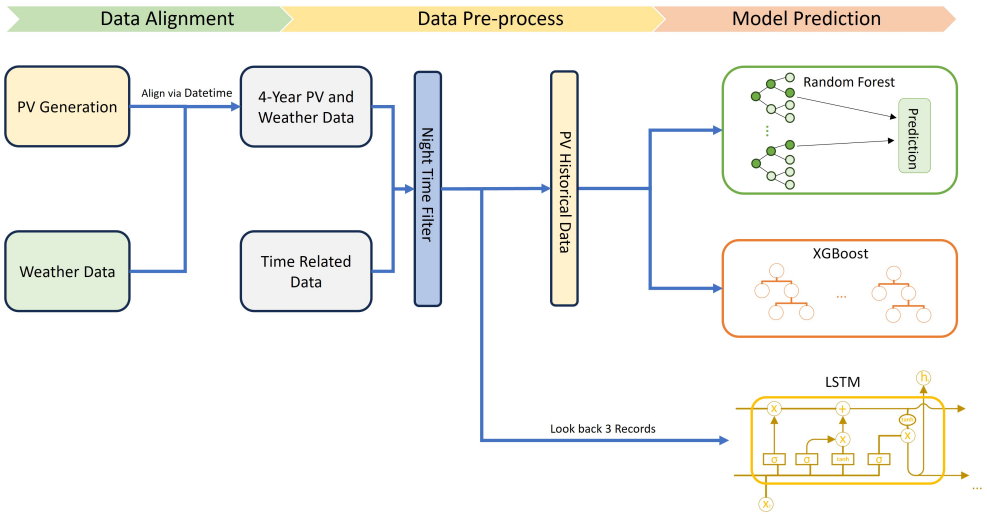


Fig. 2. Microgrid PV generation prediction framework

In RF, individual decision trees are trained on these bootstrapped samples. Each tree is a separate entity, oblivious to the existence of others. This isolation ensures that they make diverse and independent predictions. In RF, the calculation of coefficient Gini can be obtained by Eq. (9), which is commonly used as a measure of impurity.

$$Gini(S) = 1 - \sum_{i=1}^{m} P_{\mathrm{i}}^2 \tag{9}$$

in which P_i denotes the probability of category C_j in the sample set. To calculate the coefficient formula of the split node,

$$Gini_{split}(S) = \frac{|S_1|}{|S|} Gini\,(S_1) + \frac{|S_2|}{|S|} Gini\,(S_2) \tag{10}$$

in which $|S|$ is the number of samples in sample set S, S_1 and S_2 are both subsets of S. The Gini impurity is used to evaluate the quality of a split, e.g., how often a randomly chosen element would be incorrectly labeled.

The profound beauty of the Random Forest algorithm lies in its ability to remember the data splits and target variable values during training. This memory serves as a valuable reference when new data is introduced. With this knowledge, the algorithm can seamlessly compute the target variables for the new data by replicating the same data-splitting process utilized during training. This elegant approach ensures the model's adaptability and makes Random Forest a stalwart tool for regression tasks and predictive modeling [32].

3.3 XGBoost Model

The XGBoost algorithm is a powerful supervised machine learning technique, particularly in regression tasks. It leverages the principles of gradient boosting and ensemble learning, harnessing the collective strength of numerous CART (Classification and Regression Tree) trees. For samples $\mathcal{D} = \{(\mathbf{x_i}, \mathbf{w_i})\}, \mathbf{i} = \mathbf{1} \ldots \mathbf{n}$, we can use an ensemble learning model

$$\hat{\boldsymbol{q}}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k\,(\mathbf{x}_i)\,, f_k \in \mathcal{F} \tag{11}$$

in which $\mathcal{F} = \{f(\mathbf{x}) = \mathbf{w_{s(x)}}\}$ is CART tree. CART trees are non-parametric decision tree algorithms used for both classification and regression tasks. In regression, for this paper, they predict numerical values other than class labels. The fundamental operation of a CART tree involves recursively splitting the input dataset based on attributes, aiming to find the best threshold for maximizing homogeneity in each subset. This process continues until a pure subset is achieved or a predefined maximum node depth is reached, culminating in the creation of leaf nodes, which hold the final decisions.

XGBoost, on the other hand, takes the concept of CART trees to the next level. It is an ensemble model that integrates multiple CART trees into a single, robust predictive model. The magic behind XGBoost lies in its optimized objective function, which consists of a training loss and a regularization term. The training loss function $\mathcal{L}$ measures the model's predictive performance by comparing the predicted values ($\hat{\boldsymbol{q}}_i$) to the actual measurements ($\boldsymbol{q}_i$), as

$$\mathcal{L}\,(\phi) = \sum_{i} l\,(\hat{\boldsymbol{q}}_i, \boldsymbol{q}_i) + \sum_{k} \Omega\,(f_k) \tag{12}$$

in which the regularization term $\Omega\,(f) = \gamma T + \frac{1}{2}\lambda\,\|w\|^2$ controls model complexity, preventing overfitting by penalizing certain aspects of the tree structure [5]. λ and γ are both constants, which control the strength of regularization and the strength of penalizes, respectively [33].

XGBoost optimizes this objective function by iterative adding new trees that simulate the residuals (differences between predictions and actual values) from the previous iteration. The goal is to find the optimal f_p value that minimizes the objective function, effectively reducing the prediction error. For a dataset $\mathcal{D} = (\mathbf{x_i}, \mathbf{w_i})$ where $\mathbf{x_i}$ rep-

resents examples and $\mathbf{w_i}$ represents features, the objective function $L^{(p)}$ is represented as:

$$\mathcal{L}(\mathcal{D}, f_p) = \sum_{i=1}^{n} l(\mathbf{w_i}, \hat{\mathbf{w}}_\mathbf{i}^{(\mathbf{p-1})} + \mathbf{f_t}(\mathbf{x_i})) + \mathbf{\Omega}(\mathbf{f_p}) \quad (13)$$

in which $\hat{\mathbf{w}}_i^{(p)}$ represents the prediction value on the p-th iteration for the i-th instance.

It can be seen that the XGBoost enhances the predictive power of CART trees by strategically combining them and optimizing an objective function. This makes it a formidable choice for regression tasks, allowing it to capture complex relationships in the data while preventing overfitting through regularization.

3.4 Long Short-Term Memory Network (LSTM)

A LSTM is a variant of Recurrent Neural Networks (RNNs) that effectively addresses the challenges of exploding and vanishing gradients, particularly in tasks involving long-term dependencies. LSTMs employ recurrent neurons to process input data through a series of activations, allowing them to retain and utilize information over extended sequences of data [34]. LSTMs excel in solving sequence-related problems, a feat achieved by their ability to overcome the limitations of Simple Recurrent Networks [35]. To illustrate the architecture of a typical vanilla LSTM model [34], [36], one can envision it as a composition of multiple memory blocks, as depicted in Figure. 3 .

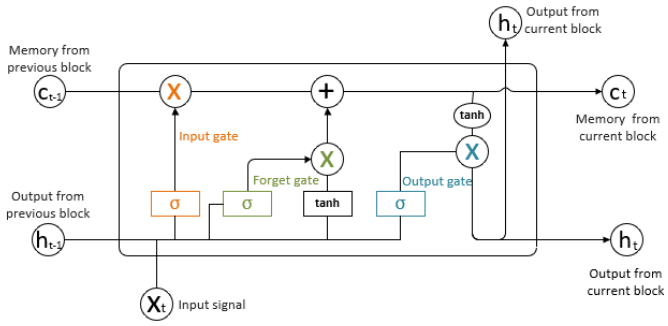


Fig. 3. LSTM memory block demonstration

The input gate receives the previous memory state (c_{t-1}), the previous output (h_{t-1}), and the current input signal (X_t) as its input. It processes this information to determine which new information should be added to the current memory state.

$$i_t = \sigma(W_{xi}X_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (14)$$

in which the forget gate decides what information should be discarded from the previous memory state (c_{t-1}) based on the current input signal (X_t) and the previous output (h_{t-1}). It ensures that irrelevant information is not carried forward.

$$f_t = \sigma(W_{xf}X_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (15)$$

in which the output gate generates the current output by combining the current input signal (X_t), the previous memory state (c_{t-1}), and the previous output (h_{t-1}). It controls the flow of information to the next layer in the network.

$$o_t = \sigma(W_{xo}X_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (16)$$

This iterative process continues as more input data is fed into the LSTM model. In multi-variables regression tasks, where multiple input features are used to predict a numerical output, LSTM networks shine. They capture complex dependencies between variables, making them well-suited for tasks such as time series forecasting, natural language processing, and many others.

The LSTM networks, with their intricate memory blocks and recurrent connections, excel in handling data sequences, making them a potent choice for multi-variable regression tasks where capturing temporal dependencies and intricate patterns is paramount.

4 Evaluation

4.1 Data Preparation

The quality of the training dataset may affect the training model's performance and accuracy [37]. In this work, for the Random Forest, XGBoost, and LSTM algorithms, 4 years of London area's PV electricity generation record from Sheffield open-source PV live data and weather data from MIDAS UK open weather data were used[1]. Both of the datasets (PV and weather) start from 2016/01/01 to 2019/12/31, recording every 60 minutes. Due to the synchronisation failure, and misoperations, there are some repeated data or vacant data. In the data pre-processing step, we removed the repeat data in both datasets according to the date and time. As a result, both the PV generation dataset and weather dataset have in total of 35065 records.

The PV generation data itself is recorded in Megawatts, and hourly generation records can reach up to thousands megawatts, sometimes even more than 5 thousand Megawatts. The weather data air temperature are stored in the degree $Celsius$; the global solar irradiance mount is stored in KJ/m^2; cloud cover is measured in *okta*, in the scale of [0,8]; precipitation amount measured in millimeters (mm); relative humidity expressed in percentage. Because there is no electricity generation during the night-time, this work removes the night-time 0 PV generation data by filtering the 0 solar irradiance records from the weather dataset.

4.2 Data Analyse

The utilisation of the Pearson correlation coefficient model serves as a valuable tool for exploring the interrelationships among the input features. This technique provides a normalised assessment of the covariance between any two feature combinations, shedding light on the strength and direction of their linear correlation [38].

In our analysis, we explored different weather-related features to find inner connections between time-related factors and PV generation. Figure. 4 visually illustrates these correlations. Among all the weather data, global radiance has high correlations with cloud cover, relative humidity (increase when precipitate), and temperature (related to season and weather conditions). Upon closer inspection of the feature correlations within the dataset, it becomes evident that pure temporal features such as hour, month, day, and

1. https://catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1

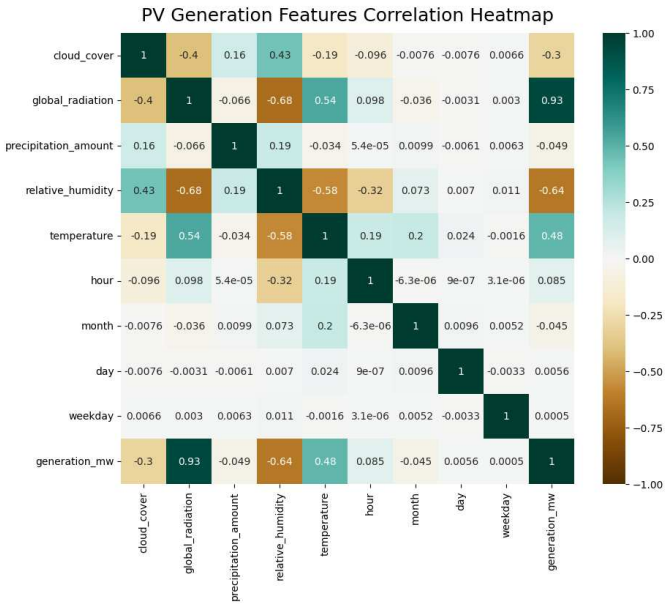


Fig. 4. Heatmap of PV Generation Features Correlation

weekday exhibit negligible direct correlation with hourly PV generation, with correlation values hovering around 0.

In comparison with the temporal features, antecedent records of global radiance (0.93), relative humidity (-0.64), temperature (0.48), and cloud cover (-0.3) reveal substantial correlations with PV generation for the following hour prediction, shown in the bottom row of Figure. 4. This suggests that these weather-related variables play a pivotal role in our predictive model. An intriguing revelation arises in the form of the correlation between humidity and PV generation: heightened humidity levels lead to the formation of a water vapor layer above the PV panels, resulting in energy loss through absorption and reflection [39].

Turning our focus to another dimension of the analysis, the relationship between the PV generation of the past 24 hours and the current PV generation has been studied as well. The heatmap of correlations between historical PV records spanning 24 hours and the contemporaneous PV generation values is presented in Figure. 5. Here, "P_24" signifies PV generation from 24 hours ago, while "P_0" represents PV generation in the forthcoming hour, which is the target of our prediction model.

Figure. 5 shows features P_1 to P_3 (previous 3 hours) and P_21 to P_24 (yesterday 3 hours after) PV generation both have high correlations (in red value above 0.6) to the future one-hour PV generation value. The high correlations of features P_1 to P_3 are attributed to the continuity of weather conditions and the temporal dependency of PV generation. Features P_21 to P_24 demonstrate strong correlations due to the predictability of mid-term solar irradiance patterns, indicating similar PV generation at corresponding hours across successive days. The future 1-hour PV generation has 0.95, 0.82, and 0.64 high statistic correlations to the previous 3-hour PV generation, respectively.

In fact, statistical and numerical solutions have been proven to perform well in short-term prediction, which only uses historical PV generation data to predict. Incorporating historical PV generation data into the input is crucial for accurate predictions. However, relying solely on statistical methods may have limitations when it comes to handling unexpected events like sudden weather changes. Including additional weather data as input can help compensate for these limitations by serving as an adjustment to the statistical predictions.

In this work, the weather data, including information like cloud cover, temperature, wind speed, and precipitation, can provide context and insights into how these external factors influence PV generation. By adding weather data as input features, the model can learn to associate specific weather conditions with deviations in PV generation. This allows the model to make more accurate predictions, especially when unusual weather events occur. Essentially, the weather data acts as a supplementary source of information that helps the model fine-tune its predictions and account for deviations from the statistical norms.

4.3 Random Forest

This work used a total of 35065 records (from 2016/01/01 0:00 to 2019/12/31 23:00) to train the random forest model. The original pure weather dataset is developed in 10 columns: hour, month, day, weekday, cloud cover, global radiation, precipitation amount, relative humidity, temperature, and capacity. In order to validate the model around the whole year time span, we randomly separated the training (70%) and testing dataset (30%) splits. The green line in Figure. 6, which only plots a part of the testing dataset, shows that the Random Forest model's performance is trained by pure 10-column weather data, which has an average of 0.9116 R^2 score and 0.0676 $nRMSE$ value.

After analysing the historical PV data, which shows that features P_1, P_2, and P_3 have very high correlations with the future PV generation value. Also, on the other hand, the historical PV generation data is not hard to acquire, if it is measured and stored securely. In Figure. 6, the blue line refers to the same Random Forest model as the green line but added the 3 historical PV records (P_1, P_2, and P_3) to improve the performance. The blue line is part of the prediction in the testing dataset and has an average of 0.9848 R^2 score and 0.028 $nRMSE$ value.

4.4 XGBoost

As a comparison group, we used XGBoost model on both training input with and without 3-hour historical PV records along with the weather data. The XGBoost model predicts future PV generation based on the previous 1-hour weather data and the time data without the data normalisation process to keep the original data features. As a result, PV generation prediction uses megawatts as its measurement unit. In Figure. 7, the green line is the training model without applying the 3-hour historical PV sequence, which has 0.9321 R^2 score and 0.0591 $nRMSE$ and the blue line is the training model with the 3-hour historical PV sequence, which has 0.9842 R^2 score and 0.0286 $nRMSE$ value. All the values are evaluated by calculating the average of 10 model fits. A small part of two PV generation prediction comparisons are illustrated in Figure. 7.

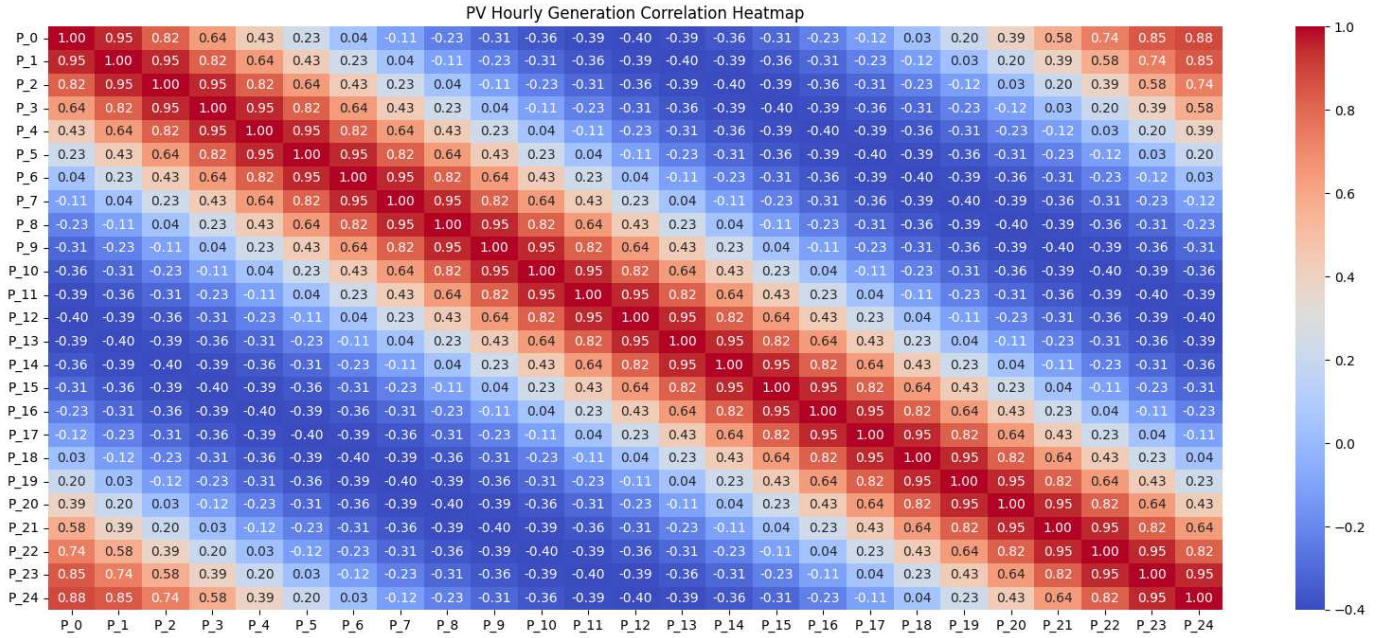


Fig. 5. Heatmap of PV Generation 24-hour Sequence Correlation

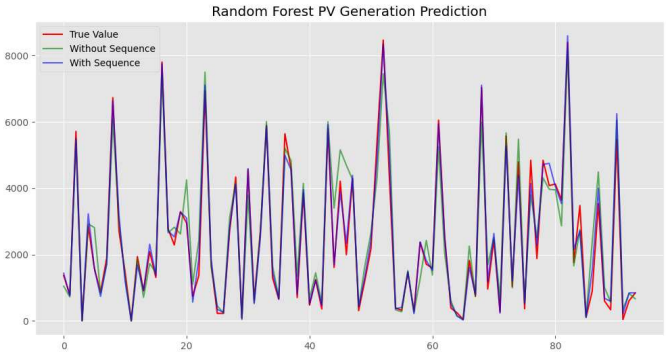


Fig. 6. Random Forest PV generation prediction

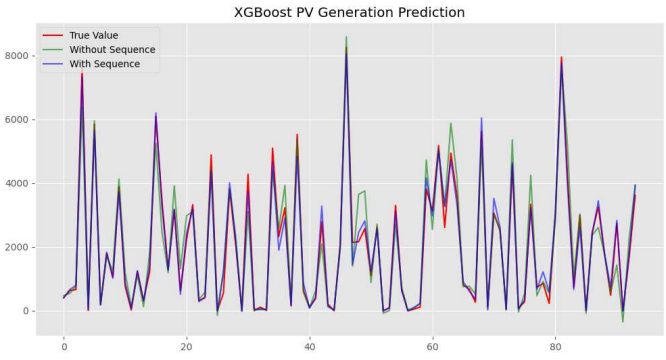


Fig. 7. XGBoost PV generation prediction

Figure. 8 shows the different feature importance between the pure weather data model (left) and the model combined weather and historical PV sequence (right). It is interesting to see that in the left figure, the pure weather data model, global radiation has the dominating position (above 70% importance) in the XGBoost PV generation prediction model. On the other hand, the other weather features do not contribute much to the model, even the relatively important weather features mentioned in other work like temperature and cloud cover.

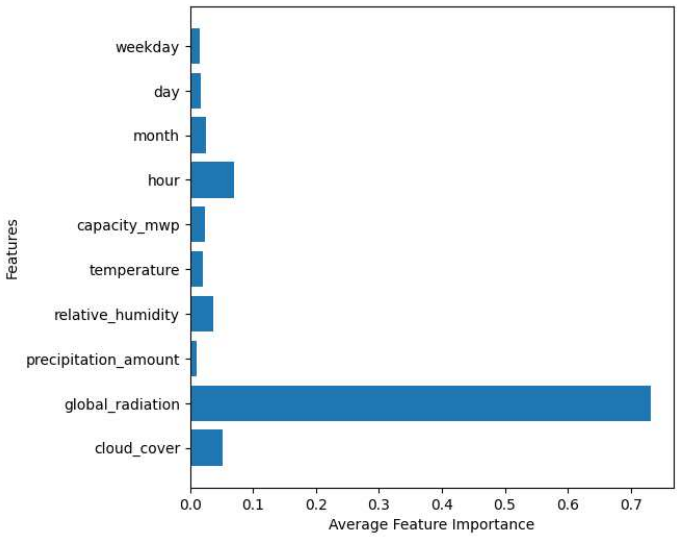
However, when the dataset added the 3-hour historical PV sequence data (lag 1, lag 2, and lag 3 in the right of Figure. 8), the dominant prediction feature switches from global radiation to lag 1, which is the previous 1 hour PV generation record. On the contrary, the weather feature – global radiation, has much less importance than in the pure weather dataset prediction (with an over 60% importance drop).

For the short-term PV generation prediction, this XGBoost model is very accurate and the feature importance contributes to the model makes sense as well, which mainly utilizes the historical PV sequence (the statical model) to predict and then associate weather conditions (global radiation, cloud cover, temperature, etc.) to the deviations in predicting PV generation.

4.5 LSTM

The LSTM model is used to conduct one-step ahead PV generation prediction as a comparison group with the Random Forest and XGBoost models. We set the look-back window as 3 (last 3-hour records) so that the LSTM algorithm will take previous 3-hour PV sequence records to predict the next hour PV generation value, which should work the same as the proposed crafted dataset.

The LSTM model contains 100 neurons; mean absolute error as loss function; and the adam optimizer. LSTM model's average score (R^2) is around 0.9739, average $nRMSE$ is around 0.037. Also, different from the Random Forest and XGBoost models, LSTM model requires data normalization to guarantee performance. LSTM short-term prediction using 3-hour historical PV sequence has a rela-



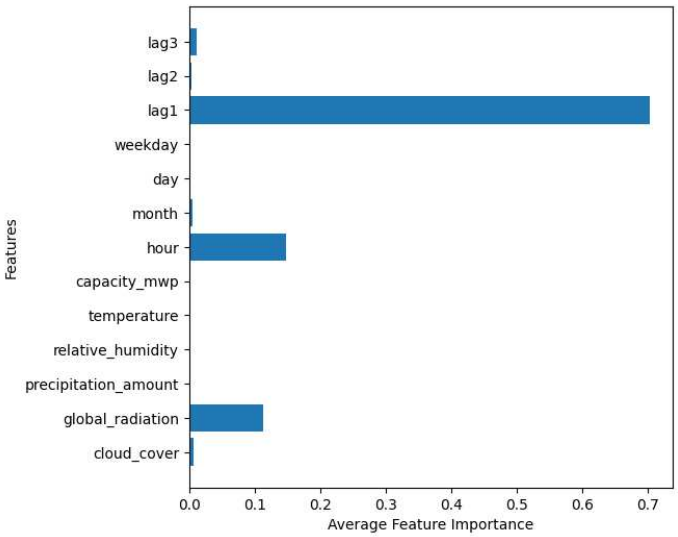


Fig. 8. XGBoost PV generation prediction feature importance

tively good performance. Part of the LSTM multi-variable weather-based model can be illustrated in Figure. 9:

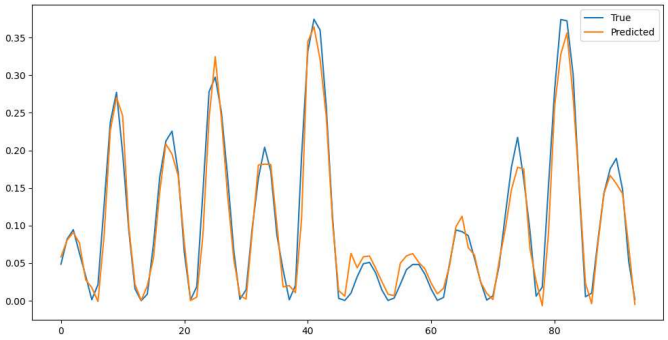


Fig. 9. LSTM PV generation prediction

4.6 Proposed Resilient PV Prediction Model

As discovered in the feature importance Figure. 8, *lag 1*, *global radiation*, and hour are the three main features to predict the PV generation in the historical PV generation included dataset. However, global radiation, hour, cloud cover, and relative humidity are the main features in pure weather data prediction. In this work, we research more into the short-term resilient PV generation prediction that aims at using minimum data to achieve high performance.

When the random forest/XGBoost models are purely trained by feature lag 1 (previous 1-hour PV generation), the models' R^2 scores are around 0.83 and $nRMSE$ is around 0.09.

However, using our proposed model building method, when just combining feature lag 1 with the date time features (hour, month, day, weekday), which are already known and do not require any measurement. The Random Forest/XGBoost models can perform up to 0.9804 R^2 score and 0.0318 $nRMSE$ value. It is a resilient and effective way to conduct short-term PV generation, which in fact

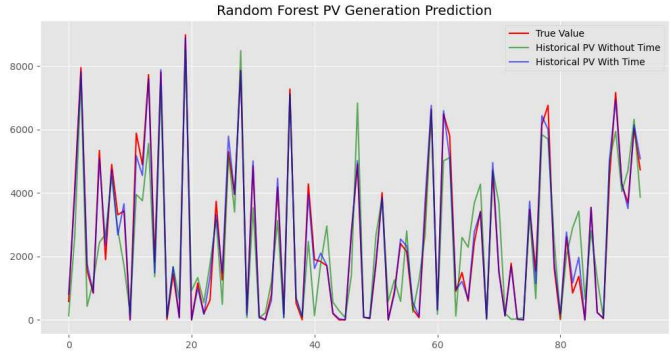


Fig. 10. Comparison between historical PV generation dataset with/without time-related data

only requires the measurement of 1 record of historical PV generation (the previous hour's PV generation).

The comparison of the performances between pure historical PV generation dataset and time-related data plus historical PV generation dataset can be illustrated in Figure. 10, in which the red line is the true value; the green line is the only 1-hour PV generation model; and the blue line only added the time-related features, respectively. It is surprising that just purely adding the time-related data (hour, month, day, weekday), the performance of the model sharply will increase by 15% in R^2 score.

4.7 Result Comparison

In this work, we compared the multi-variable time-series PV generation algorithm LSTM, Random Forest, and XGBoost, which use previous PV generation data combined with weather data to predict future PV generations. Two common evaluation methods R square score (R^2) and Normalised Root Mean Squared Error ($nRMSE$) were used.

R^2 score, which also means coefficient of determination, describes the accuracy of dependent variable changes according to the prediction of independent variables. In other words, it explains how well the data fit the model and where

TABLE 1
Comparison Between Random Forest, XGBoost, and LSTM Models

	LSTM	RF-S	XGBoost-S	RF-W	XGBoost-W	RF-L	XGBoost-L	RF-R	XGBoost-R
R^2	0.9739	0.8286	0.8462	0.9116	0.9321	0.9848	0.9842	0.9803	0.9831
$nRMSE$	0.037	0.0944	0.0898	0.0676	0.0591	0.028	0.0286	0.0319	0.0295

$R^2 = 1$ represents the perfect fit. Where y_i denotes the true value of the dataset, $\hat{y}_i$ denotes the predicted value, and $\overline{y}_i$ denotes the mean value.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2} \tag{17}$$

RMSE is the square root of the MSE value, which measures the standard deviation of residuals and can be calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2} \tag{18}$$

However, because different works use different datasets and various data-processing methods, RMSE is sensitive to the scale of the dataset, on the contrary, normalised RMSE ($nRMSE$) provides a measure of the error relative to the range of the data, making it easier to interpret between predicted and observed values, which can be calculated as follows:

$$nRMSE = \frac{RMSE}{max(y_i) - min(y_i)} \tag{19}$$

Table. 1 shows the evaluation results of three models trained with different datasets: '-S' denotes the dataset that only uses 3-hour historical PV records for training and prediction, which only have around 0.83 R^2 score; '-W' denotes the dataset that only includes weather-related features: cloud cover, global radiation, precipitation amount, relative humidity, and temperature, which requires a lot of accurate measurements to conduct the prediction.

The prediction results are relatively good but not very accurate (0.9321 R^2 score), in which '-L' refers to the proposed dataset that combines weather data with the 3-hour lagging PV generation data. The dataset including the historical PV generation has significant performance improvement in predicting PV generation, which has 7.32% improvement in the random forest model and 5.21% improvement in the XGBoost model. Also, our crafted historical PV generation prediction dataset has around 1% more accuracy on random forest and XGBoost compared to the multi-variable LSTM model; and '-R' in the table refers to the proposed resilient dataset that combines time-related data (hour, month, day, and weekday) and historical PV generation data, which does not require any additional weather data. With only 1-hour historical PV data and predicting time features, the confidence of the proposed model can reach up to 0.9831 R^2 with 0.0295 $nRMSE$.

5 Conclusion

Accurate power generation in smart microgrids can significantly enhance the performance and security of 5G communication for consumer devices. Aiming at improving the performance of predicting short-term PV panel electricity generation in microgrids, this work developed power generation resilient models using machine models. Specifically, models based on Random Forest, XGBoost, and LSTM were developed and evaluated. The experimental results demonstrate the efficacy of the proposed resilient short-term PV generation models, particularly when combining weather data with historical PV data and when incorporating time-related data alongside historical PV data. The robust performance is attributed to the accuracy and availability of correctly collected and stored historical PV sequence records. The resilient model can perform around 0.983 R^2 score and 0.029 $nRMSE$ value on average.

References

[1] H. Liao, Z. Jia, Z. Wang, Z. Zhou, X. Wang, S. Mumtaz, and M. Guizani, "Adaptive learning-based secure and energy-aware resource management for multi-mode low-carbon piot," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 5171–5176.

[2] A. Agga, A. Abbou, M. Labbadi, Y. E. Houm, and I. H. Ou Ali, "Cnn-lstm: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production," *Electric Power Systems Research*, vol. 208, p. 107908, 2022.

[3] M. Manohar, E. Koley, and S. Ghosh, "Stochastic weather modeling-based protection scheme for hybrid pv–wind system with immunity against solar irradiance and wind speed," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3430–3439, 2020.

[4] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, "Short-term residential load forecasting based on resident behaviour learning," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1087–1088, 2018.

[5] H. Kanchev, D. Lu, F. Colas, V. Lazarov, and B. Francois, "Energy management and operational planning of a microgrid with a pv-based active generator for smart grid applications," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 10, pp. 4583–4592, 2011.

[6] H. Li, S. Li, and G. Min, "Lightweight privacy-preserving predictive maintenance in 6g enabled iiot," *Journal of Industrial Information Integration*, vol. 39, p. 100548, 2024.

[7] S. K. Chow, E. W. Lee, and D. H. Li, "Short-term prediction of photovoltaic energy generation by intelligent approach," *Energy and Buildings*, vol. 55, pp. 660–667, 2012.

[8] U. N. D. Programme, *World Energy Assessment: Energy and the Challenge of Sustainability*, 1st ed., New York, NY 10017, 2000.

[9] A. Detollenaere and G. Masson, "Snapshot of global pv markets 2023 task 1 strategic pv analysis and outreach," pp. 8–10, 2022.

[10] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, and F. De Felice, "Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions," *Sustainability*, vol. 12, no. 2, 2020.

[11] S. Li, "Zero trust based internet of things," *EAI Endorsed Transactions on Internet of Things*, vol. 5, no. 20, pp. e1–e1, 2019.

[12] D. for Energy Security and N. Zero, "Energy trends september 2023," p. 12, 09 2023.

[13] A. D. Martin, J. M. Cano, J. F. A. Silva, and J. R. Vázquez, "Backstepping control of smart grid-connected distributed photovoltaic power supplies for telecom equipment," *IEEE Transactions on Energy Conversion*, vol. 30, no. 4, pp. 1496–1504, 2015.

[14] G. G. Kim, J. H. Choi, S. Y. Park, B. G. Bhang, W. J. Nam, H. L. Cha, N. Park, and H.-K. Ahn, "Prediction model for pv performance with correlation analysis of environmental variables," *IEEE Journal of Photovoltaics*, vol. 9, no. 3, pp. 832–841, 2019.

[15] S. Ferlito, G. Adinolfi, and G. Graditi, "Comparative analysis of data-driven methods online and offline trained to the forecasting of grid-connected photovoltaic plant production," *Applied Energy*, vol. 205, pp. 116–129, 2017.
[16] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. Mohamed Shah, "Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques," *IET Renewable Power Generation*, vol. 13, no. 7, pp. 1009–1023, 2019.
[17] M. Massaoudi, H. Abu-Rub, S. S. Refaat, I. Chihi, and F. S. Oueslati, "Deep learning in smart grid technology: A review of recent advancements and future prospects," *IEEE Access*, vol. 9, pp. 54 558–54 578, 2021.
[18] O. A. Omitaomu and H. Niu, "Artificial intelligence techniques in smart grid: A survey," *Smart Cities*, vol. 4, no. 2, 4 2021.
[19] M. Pérez-Ortiz, S. Jiménez-Fernández, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martínez, and S. Salcedo-Sanz, "A review of classification problems and algorithms in renewable energy applications," *Energies*, vol. 9, no. 8, 2016.
[20] S. Li, S. Zhao, G. Min, L. Qi, and G. Liu, "Lightweight privacy-preserving scheme using homomorphic encryption in industrial internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 16, pp. 14 542–14 550, 2021.
[21] S. I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis, "Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting," in *2016 IEEE International Energy Conference (ENERGYCON)*, 2016, pp. 1–6.
[22] A. Álvarez Gallegos, L. Fara, A. Diaconu, D. Craciunescu, and S. Fara, "Forecasting of energy production for photovoltaic systems based on arima and ann advanced models," in *International Journal of Photoenergy*, 2021.
[23] E. G. Kardakos, M. C. Alexiadis, S. I. Vagropoulos, C. K. Simoglou, P. N. Biskas, and A. G. Bakirtzis, "Application of time series and artificial neural network models in short-term forecasting of pv power generation," in *2013 48th International Universities' Power Engineering Conference (UPEC)*, 2013, pp. 1–6.
[24] M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Forecasting of pv power generation using weather input data-preprocessing techniques," *Energy Procedia*, vol. 126, pp. 651–658, 2017.
[25] U. K. Das, K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. Van Deventer, B. Horan, and A. Stojcevski, "Forecasting of photovoltaic power generation and model optimization: A review," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 912–928, 2018.
[26] Y. Sun, V. Venugopal, and A. R. Brandt, "Convolutional neural network for short-term solar panel output prediction," in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC and 34th EU PVSEC)*, 2018, pp. 2357–2361.
[27] Sandia National Laboratories, "Plane of array (poa) irradiance." [Online]. Available: https://pvpmc.sandia.gov/modeling-steps/1-weather-design-inputs/plane-of-array-poa-irradiance/
[28] L. A. Fernandez-Jimenez, A. Muñoz-Jimenez, A. Falces, M. Mendoza-Villena, E. Garcia-Garrido, P. M. Lara-Santillan, E. Zorzano-Alba, and P. J. Zorzano-Santamaria, "Short-term power forecasting system for photovoltaic plants," *Renewable Energy*, vol. 44, pp. 311–317, 2012.
[29] M. Mao, W. Gong, and L. Chang, "Short-term photovoltaic output forecasting model for economic dispatch of power system incorporating large-scale photovoltaic plant," *2013 IEEE Energy Conversion Congress and Exposition, ECCE 2013*, pp. 4540–4545, 09 2013.
[30] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, pp. 912–928, 2023.
[31] W. Cheng, Y. Yan, J. Xia, Q. Liu, C. Qu, and Z. Wang, "The compatibility between the pangu weather forecasting model and meteorological operational data," 2023.
[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, p. 5–32, 2001.
[33] R. Johnson and T. Zhang, "Learning nonlinear functions using regularized greedy forest," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 942–954, 2014.
[34] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 1573–7462, 2020.
[35] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.
[36] M. A. Saleem, X. Li, K. Mahmood, S. Shamshad, M. F. Ayub, A. K. Bashir, and M. Omar, "Provably secure conditional-privacy access control protocol for intelligent customers-centric communication in vanet," *IEEE Transactions on Consumer Electronics*, 2023.
[37] R. Ahmad, Pratyush, and R. Kumar, "Very short-term photovoltaic (pv) power forecasting using deep learning (lstms)," in *2021 International Conference on Intelligent Technologies (CONIT)*, 2021, pp. 1–6.
[38] J. Benesty, J. Chen, and Y. Huang, "On the importance of the pearson correlation coefficient in noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 757–765, 2008.
[39] R. Abbasi, A. K. Bashir, A. O. Almagrabi, M. B. B. Heyat, and G. Yuan, "Efficient lossless based secure communication in 6g internet-of-things environments," *Sustainable Energy Technologies and Assessments*, vol. 57, p. 103218, 2023.

Xueyi Wang received the B.S. degree in Software Engineering from Changzhou Institute of Technology in 2020, the B.S. degree in Digital Technology (Networks & Communications) from the University of Hertfordshire in 2020, and the M.S. degree in Cybersecurity from Cardiff University in 2021. Currently, he is pursuing the Ph.D. degree at Cardiff University. His research interests include Secure Microgrids, Resilient Microgrids, Cybersecurity, Reliability, and XAI.