This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/UPEC61344.2024.10892541, 2024 59th International Universities Power Engineering Conference (UPEC)

Machine Learning of Solar Energy Forecasting using ensemble LSTM method

Sahar Daebes
Department of Electronic and Electrical Engineering
College of Engineering, Design and Physical Sciences
Brunel University London, UK
Sahar.daebes@brunel.ac.uk

Dr Mohamed Darwish
Department of Electronic and Electrical Engineering
College of Engineering, Design and Physical Sciences
Brunel University London, UK
Mohamed.darwish@brunel.ac.uk

Dr Chun Sing Lai
Department of Electronic and Electrical Engineering
College of Engineering, Design and Physical Sciences
Brunel University London, UK
chunsing.lai@brunel.ac.uk

***Abstract*— The transition in energy systems aims for efficiency improvements at a higher level, targeting the reduction of climate change impacts. Investing in solar energy, endorsed by the global scientific community, is essential. One of the core obstacles hindering the seamless integration of photovoltaic (PV) systems into power grids is the associated uncertainty. This paper focuses on machine learning forecasting algorithms for solar power generation, specifically using the Long Short-Term Memory (LSTM) algorithm, a type of recurrent neural network (RNN). Accurate prediction models are crucial for maximising the efficiency and reliability of solar energy systems, especially with high PV penetration. The LSTM architecture's ability to capture temporal dependencies makes it well-suited for time series forecasting tasks such as solar irradiance prediction. Using ensemble LSTM improves output accuracy and reduces mean square error in solar energy plant production. Temporal variations in solar power production, analysed using SARIMA/ARIMA models, highlight the volatility of PV power generation, which causes issues such as frequency instability, dispatch difficulties, and surges in current/voltage on the grid.

Keywords— Machine Learning algorithms ML, Long Short-Term Memory LSTM, transition energy system, PV forecasting algorithms, SARIMA / ARIMA

I. INTRODUCTION

Higher grid system integration of PV power is now required due to the recent exponential growth in PV installations. Managing supply and demand, saving costs, maintaining grid stability, planning, and investing, among other tasks, becomes difficult without accurate solar energy generation forecasting[1].

Grid operators can effectively control the supply and demand for electricity in real-time by accurately forecasting solar energy production. This ensures efficient electricity distribution to customers and helps prevent blackouts [2]. Also, Utility companies can prevent the overproduction or underproduction of solar energy by using effective forecasts. This can reduce expenses by lowering fuel expenses (for conventional power generation) or eliminating the need for expensive energy storage systems [3]. Also, by predicting fluctuations in solar energy production, grid operators can anticipate and mitigate any likely impact on the grid's stability. Additionally, precise forecasting becomes more crucial the more solar energy is added to the electrical system. It allows grid operators to seamlessly integrate solar energy into the system and ensures energy reliability in the future [4]. When it comes to energy production, transmission, and distribution, solar energy forecasting can help energy companies with planning and investment, which is another essential component. It makes it possible to forecast revenue streams more precisely, which is beneficial for financial planning and investment decisions [5].

II. PREDICTING SOLAR POWER PRODUCTION USING MACHINE LEARNING ALGORITHMS

Machine learning algorithms forecast the amount of solar energy produced by finding trends in the data, collecting important statistical measures, and determining relationships between various factors [6]. The process involves the following steps:

- Machine learning algorithms forecast the amount of solar energy produced by finding trends in the data, collecting important statistical measures, and determining relationships between various factors. .
- Feature Engineering: Extracting relevant features from the data that can help the model make accurate predictions, such as weather conditions, solar irradiance, and historical power production [5].
- Model Training: The process of teaching a machine learning model to anticipate the production of solar power using past data. Various machine learning algorithms such as " LSTM, Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree (DT), Random Forest (RF), and other ensemble methods are commonly used for this purpose" [6].
- Evaluation and Selection of the Model: To guarantee the model's correctness and generalizability, it is assessed using suitable metrics after training [5].

Accurate solar power generation forecast is made possible by machine learning algorithms, which take into account several variables like solar panel efficiency, weather, and sunshine intensity as shown in Fig. 1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/UPEC61344.2024.10892541, 2024 59th International Universities Power Engineering Conference (UPEC)

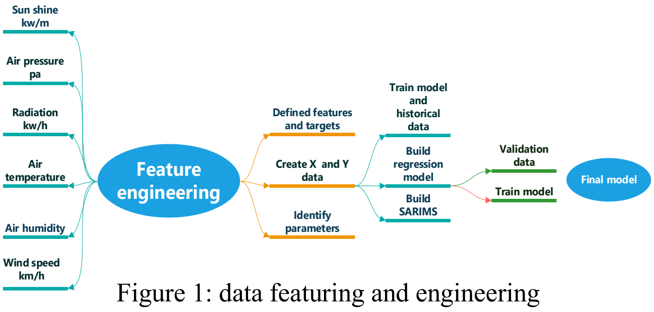

Figure 1: data featuring and engineering

III. Solar Power Prediction with Ensemble 2 LSTM Algorithms

The development of a forecasting model for solar power generation is the main goal of this model using the "Long Short-Term Memory (LSTM) algorithm, a type of recurrent neural network (RNN)". Time series forecasting problems like solar irradiance prediction are a good fit for the LSTM architecture because of their capacity to capture temporal dependencies. By producing solar energy plants, ensemble LSTM can increase output accuracy and decrease mean square error. [7]

To create the dataset, information on historical sun irradiance, meteorological parameters, and environmental factors must first be gathered. Time-stamped observations of solar irradiance, temperature, humidity, cloud cover, and other pertinent variables will be included in the data. Pre-processing will be done on the feature-engineered data to manage missing values, normalise the features, and provide time-related features. [8] Feature engineering techniques will be employed to extract pertinent patterns and correlations from the data to enhance the data's input for the LSTM model. [9]The LSTM architectural model development will be utilised to create a predictive model for solar power generation. Using historical data, the model will be trained to take advantage of the LSTM's capacity to identify patterns and long-term dependencies in sequential data. Forecasting horizons, forecasting methodologies, and input characteristics or algorithms are the three categories into which the flowing figure categorises the LSTM ensemble model to concentrate on the most accurate PVP output forecasting models based on the GHI global horizontal irritation. To concentrate on machine learning techniques, PV power prediction is used the stacked LSTM model is an extension of the LSTM model. It has the benefit of having several stacking hidden layers, which allows the model to learn about the raw temporal input at each time step. Additionally, because the parameters of such models are dispersed throughout the space, the stacked LSTM model can improve the non-linear operations of raw data and accelerate convergence. deep learning algorithms. [7]

Amongst RNNs, Long Short-Term Memory (LSTM) networks have exposed the best presentation and The use of LSTM networks for PV power prediction has gained popularity in recent years. [5] Quing and Niu [4] have created two ANN method types—one based on DNN and the other on LSTM—for the prediction of PV power output using data from a South Korean PV operator situated in Gumi City. [10]Two seasonal parameters (month of the year and day of the month) and four meteorological elements (temperature, humidity, cloudiness, and radiation) were taken into account in the first LSTM model, however, the second LSTM model (LSTM2) simply employed the four meteorological factors. Three hidden layers and an output layer for the hourly PV power output were present in both approaches. [9] The findings demonstrated that when compared to the suggested ANN, ANN2, DNN, and LSTM2 models as well as the traditional techniques of Arima and S-Arima, the suggested LSTM model produces the best results. [9]

Wan et al [11]have suggested using LSTM to precisely predict PV systems' power output one hour in advance. The model uses two PV datasets from different locations respectively. Five LSTM The authors created and evaluated models using a variety of architectures: Utilizing an LSTM network for regression (model 1) Regression using the window approach with LSTM (model 2), Regression using time steps and LSTM (model 3), layered LSTM with memory between batches and LSTM with memory between batches (model 4) (model 5). The input for each model was solely endogenous, historical PV data. Three forecasting models were tested with the suggested model 3, which performed much better than the other models: multiple linear regression (MLR), bagged regression trees (BRT), and neural networks (NN). Once more, for both datasets, the suggested technique had the lowest RMSE error.

Greco et al. [8] offered a stacked LSTM–RNN model in their work to anticipate the PV plants' monthly power output at new locations. For almost five years, the suggested approach made use of historical data from 164 dispersed PV facilities. Eight input variables are used: the month of operation, the estimated solar irradiance, the mean monthly temperature, the relative humidity, the wind speed, the quantity of precipitation, the amount of clouds, and the length of sunshine. Through cross-validation, the actual power output values of the test plants were compared with the anticipated power output values. The findings demonstrated that the suggested approach accurately assesses the potential for power production at each new location while also capturing the temporal patterns in monthly data.

In this work, the LSTM network—a substantial component of the RNN—was also taken into consideration. Long-term dependency issues can be resolved with LSTM neural networks. LSTM features a unique neuron structure termed a memory cell that may store information over an arbitrary period, in contrast to RNN networks that employ temporal information of the input data. A typical LSTM unit's architecture consists of a memory cell coupled by a series of gates. The forget gate determines which data from the memory cell can be safely discarded. The input gate decides which values from the input to update the memory state, and finally, the output gate generates the output for the current time step based on the input and the memory cell.

The stacked LSTM model is an extension of the LSTM model. It has the benefit of having several stacking hidden layers, which allows the model to learn about the raw temporal input at each time step. Additionally, because the parameters of such models are dispersed throughout the data, the stacked LSTM model can improve the non-linear operations of raw data and accelerate convergence. An expansion of the LSTM model is the stacked LSTM model. The model can learn about the raw temporal input at each time step thanks to its multiple stacking hidden layers.[7] Additionally, the stacked LSTM model can enhance the non-linear processes of raw data and speed up convergence because the parameters of such models are distributed throughout the space.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/UPEC61344.2024.10892541, 2024 59th International Universities Power Engineering Conference (UPEC)

IV. The Concept of Machine Learning ML Forecasting in Energy Transition

The term "ML forecasting" refers to the more general idea of machines being able to do tasks that are typically necessary for solar power prediction to increase efficiency and mitigate climate change. [3] ML forecasting is a collection of methods, mathematical models, and algorithms that may be used to conclude massive datasets, spot trends, and estimate the likelihood of various events in intricate, multivariate scenarios. It is not a single technology or product. [1] Machine learning algorithms for solar power have the potential to accelerate the energy transition. This section highlights the most promising machine learning applications for the energy transition in four important areas: renewable power generation and demand forecasting; grid operation and optimization; energy demand management; and materials discovery and innovation. [3] Photovoltaic solar energy is at the forefront of the global movement to cut carbon emissions from the production and transmission of electricity. Communities using solar energy will have the power they need now and the security that comes with it to create a resilient future. using predicting algorithms as a special method to provide the highest level of accuracy in the production of solar power. The energy transition to renewable energy is a new method of working from forecasting and planning to offer environmentally friendly power. [12] Algorithms for predicting solar power are combined with creativity and teamwork to produce workable solutions for Iraq's energy industry. implementing a net-zero future by integrating new energy sources into the grid, facilitating the transfer of current assets, and connecting a new sustainable energy system. Previous power measurements and meteorological forecasts of solar irradiance, relative humidity, and temperature at the location of the photovoltaic power system are required as inputs for the machine-learning approach. [1] The validity of the forecasting method is demonstrated by applying it to the power production of a real PV power plant. The time horizon of the prediction, the location, the data that is available, and the required level of accuracy all influence the forecasting method selection. It is common practice to use an ensemble or hybrid technique that combines several methodologies to improve overall forecasting performance [1] as shown in Fig. 2.

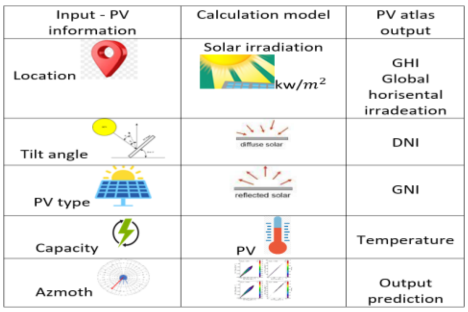

Figure 2: Input location and, PV system information to get the output of solar prediction.

The most often used forecasting techniques to estimate the generation of solar power include ensemble approaches, both conventional statistical techniques and machine learning algorithms. Certain methods and algorithms, like machine learning algorithms, are employed in the forecasting of solar power: Solar power generation is predicted using a variety of machine learning algorithms, including Random Forest (RF), Decision Tree (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM), and other ensemble techniques. [3] [10]For the most precise solar power production, machine learning and statistical methodologies are used.

By considering variables including weather patterns, environmental circumstances, and solar power plant efficiency, these algorithms are used to increase the accuracy of predictions for solar power generation. [9], [13]

Solar power forecasting can be done in several ways: long-term for a season, short-term for a week or longer, intra-day, and intra-hour. Because ANN can produce precise forecasts, it is a useful technique for optimising the integration of solar energy into the power grid in solar power forecasting. Based on a variety of input data, including meteorological conditions, it is used to forecast solar energy generation. ANN has been used for this purpose in several studies, showing that it is an effective tool for forecasting solar power generation. In addition, the effectiveness of ANN in predicting solar energy has been assessed by contrasting it with alternative algorithms such as Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). [4], [8], [9], [12]

V. Analysing Data in Python through Visual Representation and Additional Investigation

Understanding a solar power plant's performance through an examination of the data generated by its photovoltaic modules is the goal of the model analysis. The model is composed of Examining the association between all features and reviewing the exploratory Data Analysis (EDA) of photovoltaic systems (PV systems). The picture demonstrates A photovoltaic system, which uses photovoltaic cells to convert light energy into electricity directly, is used in PV systems to harvest solar power. These cells provide the basis of PV Modules, which are enclosed collections of cells connected in series and/or parallel and controlled by the Photovoltaic Effect. [5]. The aggregation of modules forms a PV Panel, and multiple panels constitute a complete PV Array as shown in Fig. 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/UPEC61344.2024.10892541, 2024 59th International Universities Power Engineering Conference (UPEC)

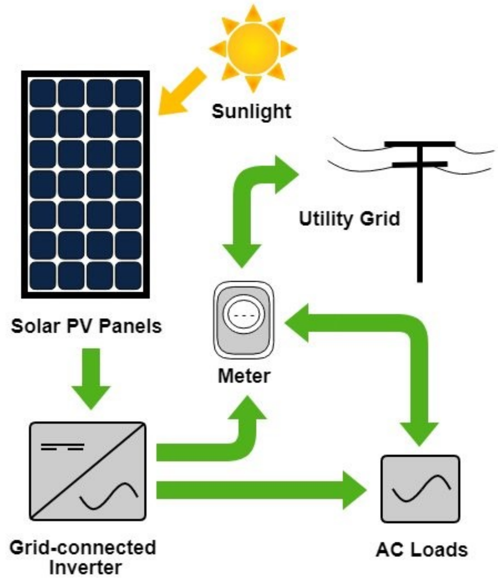

Figure 3: grid-connected PV system [5]

The essential PV inverter transforms DC power from solar arrays or batteries into AC power for common utility-powered products. It serves as the central component of PV systems, enabling the conversion of DC to AC electricity.

The weather has an impact on PV systems; in favorable weather, their yield is maximised, while in unfavorable weather, it is decreased. It is essential to comprehend how weather affects solar power plant yield [11].

Temperature and sun irradiation are essential elements of models used to anticipate solar power. The quantity of sunshine available for electricity generation is directly correlated with solar irradiance, whereas temperature affects solar panel efficiency. Energy operators and utilities can more effectively plan and manage their resources by predicting solar power generation with more accuracy when these characteristics are combined with other pertinent features in forecasting models [14].

VI. Introducing Weather Sensor Data

Vital weather sensor data can be used to gain insights into the solar plant analysis. The general climatic conditions indicated by ambient temperature affect the efficiency of solar systems. The heat that solar panels absorb is reflected in the temperature of the module, which affects energy production. Measurement of solar radiation is essential for determining the degree of sun exposure. Through the analysis of these metrics, the model looks for trends to maximize daily yield, improving productivity and comprehending how the system behaves in a variety of weather scenarios [14].

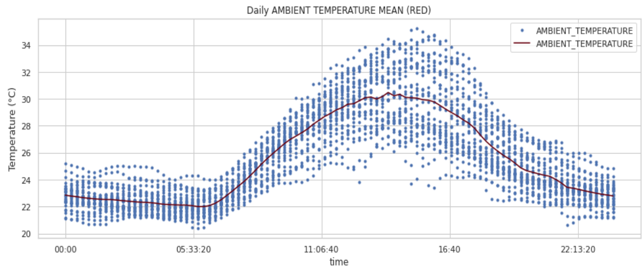

Figure 4: Daily Ambient temperature actual and predicted

The term "seasonal component" refers to recurring patterns or changes in ambient temperature that are associated with a particular season. Seasonal temperature variations can be observed, for example, in the SARIMA temporal variation of solar power output (season auto-regaration integration moving average). Summertime temperatures are higher while wintertime temperatures are lower. [9], [14] Secondly, the Trend Component: The trend in the ambient temperature data indicates the long-term direction or tendency as shown in Figs. 4&5. It records the overall movement and shows whether temperatures are generally rising, falling, or staying mostly constant over time. Finally, the residual component, sometimes referred to as the remainder or error term, considers any variations in the ambient temperature data that the seasonal and trend components are unable to explain. It includes noise, random variations, and unforeseen influences that affect temperature.

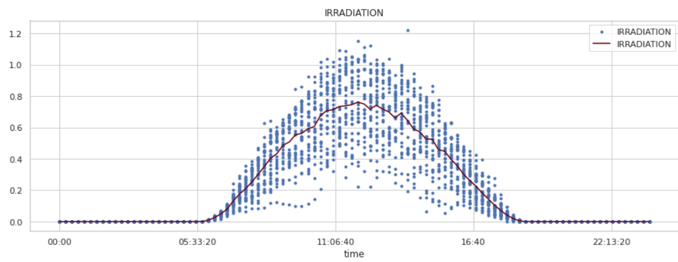

Figure 5: the amount of sunlight (irradiation) predicted and actual

VII. Model Generation

In this study, the Recurrent Neural Network (RNN), a class of artificial neural networks (ANN) that is frequently employed for time-series analysis, is constructed. Because of its internal memory state, the network can process any sequence of inputs by considering data from numerous earlier time steps. RNN has dynamic temporal behavior as a result. [11] Specifically, RNN functions similarly to short-term memory because it "remembers" and applies information from earlier observations as it progresses through time. [7] The LSTM model method uses Long Short-Term Memory (LSTM) layers to build and return a deep learning model. The model consists of two LSTM layers, each followed by an output-related dropout layer and a final dense layer. [8] The model is optimised for regression tasks by minimising the mean squared error through compilation using the Adam optimizer. The expected input shape for the model is sequences of length 1 with 6 features each. The figure below represents two LSTMs as ensemble method model-1 has hidden layers with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/UPEC61344.2024.10892541, 2024 59th International Universities Power Engineering Conference (UPEC)

50 units and model-2 has hidden layers with 80 units as illustrated in Fig. 6.

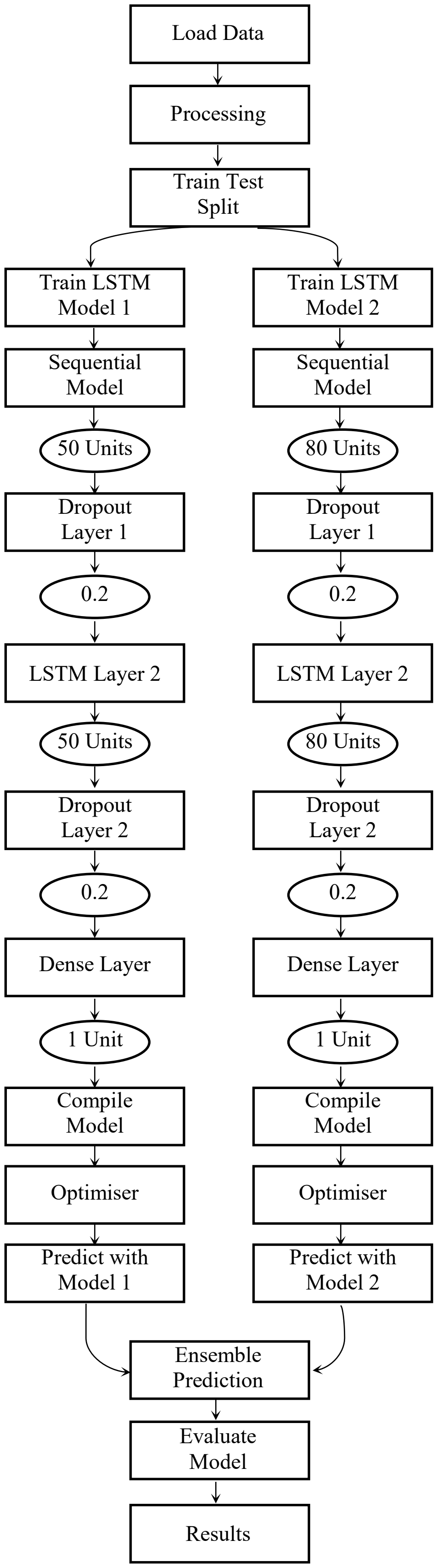

Figure 6: The 2 LSTM models created with hidden layers.

Dropout ensures that the model learns to generalise from the training data rather than memorise it by randomly turning off some neurons, causing the model to "forget" part of its training. Dense Layer: a completely connected layer, a dense layer's neurons are all connected to the neurons in the layer above it. A matrix-vector multiplication and bias offset are used to compute the output. It is positioned at the network's terminus to generate the required output dimensions.

Reshaping prepares the data to be processed by LSTM by clarifying sequence length and feature count [samples, timesteps, features]. The historical meteorological datasets used for model training of PV system power output readings. Data include six features in the figure: (wind speed, Irradiation, temperature, sunshine, air pressure, relative air humidity in the X definition, and system production in the Y definition as shown in Table 1.

Table 1: the definition of the X& Y parameters

X parameter	WindSpeed (km/h)	Sunshine (W/m²)	AirPressure (pa)	Radiation (W/m²)	Air temp. (°C)	Relative AirHumidity (%RH)
Y parameter	System Production (kWh)	To be predictid	To be predicted	To be predictid	To be pridicted	To be predicted

The data set characteristics X and Target y are provided, which can be divided into training and testing (Sequential) as (X _test, test, X _train, Y train, and test size typically (20 %). The model is a feedforward neural network (NN) with a linear stack of layers in the (Keras) deep learning toolkit. [15]

- The accuracy of output from two LSTM ensemble algorithms with a short-term forecast horizon was compared. [15]
- each followed by a dropout layer for regularisation and a final dense layer for output.
- The model is compiled with the Adam Optimizer and is designed to minimize the mean squared error, making it suitable for regression tasks.
- Adam is often recommended as the default optimizer to use for deep learning DL technique. Dense Layer: as a fully connected layer.
- The expected input shape for the model is sequences of length 1 with 6 features each.
- Dropout layer: to prevent overfitting. Used 20% dropped out of neurons. [14]

According to the results (Fig. 7), the seven-day forecasting model can predict well, since a visual examination of the results indicates that the predicted power output signal reacts to each fluctuation and follows the trend of the actual power output signal. Furthermore, the Root Mean Squared Error (RMSE) of our model when applied to the test data gives a value of: 0.11049111870458088 whereas when applying the k-fold cross-validation, the mean of the resulting Mean Absolute Error (MAE): 0.05512441745205155 with a standard deviation. R2 Score: 0.634048582919565.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/UPEC61344.2024.10892541, 2024 59th International Universities Power Engineering Conference (UPEC)

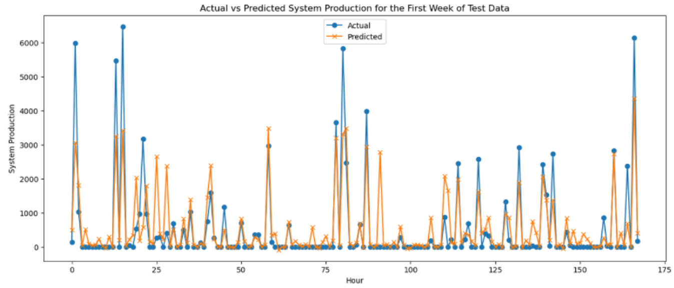

Figure 7: the actual and predicted energy production

CONCLUSIONS

One of the hardest things to forecast is solar photovoltaic energy generation, mostly because weather patterns are unpredictable. PV power output forecasting models can be used to raise the penetration level of those systems and enhance their planning, operation, and stability.

To lessen the effects of solar intermittency, this study addresses solar power forecasting, which is the process of obtaining and evaluating data to estimate solar power generation on different time horizons. The sun's path, atmospheric conditions, light scattering, and solar energy plant features are typically used to generate forecast information. There are four types of solar power forecasting techniques: intra-hour, medium-term, short-term, and long-term. Most nations typically rely on long-term solar power forecasts. But the other forecasting methods (intra-hour The case study in this paper aims to adapt and improve the work of the ANN algorithm based on the PV power forecasting methodology where most of the weather conditions have been considered like humidity, wind speed, temperature, and solar azimuth, zenith angel direct irradiation, defuse irradiation.

REFERENCES

[1] R. Bravo, "Modelling and Optimisation of Solar Power Plants with Energy Storage Systems," 2021.

[2] P. A. G. M. Amarasinghe and S. K. Abeygunawardane, "Application of machine learning algorithms for solar power forecasting in Sri Lanka".

[3] K. T. Selvi, S. M. Saranya, and R. Thamilselvan, "Solar Energy Forecasting for Devices in IoT Smart Grid," *Smart Grids and Internet of Things: An Energy Perspective*, pp. 365–394, Apr. 2023, doi: 10.1002/9781119812524.CH13.

[4] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," 2018, doi: 10.1016/j.energy.2018.01.177.

[5] R. J. Bessa, A. Trindade, C. S. P. Silva, and V. Miranda, "Probabilistic solar power forecasting in smart grids using distributed information," *International Journal of Electrical Power & Energy Systems*, vol. 72, pp. 16–23, Nov. 2015, doi: 10.1016/J.IJEPES.2015.02.006.

[6] P. Bouquet, I. Jackson, M. Nick, and A. Kaboli, "AI-based forecasting for optimised solar energy management and smart grid efficiency," *Int J Prod Res*, 2023, doi:

[7] "LSTM Response Models for Direct Marketing Analytics: Replacing Feature Engineering with Deep Learning." Accessed: Feb. 07, 2024. [Online]. Available: https://journals.sagepub.com/doi/epdf/10.1016/j.intmar.2020.07.002?src=getftr

[8] A. Greco *et al.*, "Neural network analysis of neutron and X-ray reflectivity data: automated analysis using mlreflect, experimental errors and feature engineering," *J Appl Crystallogr*, vol. 55, pp. 362–369, Apr. 2022, doi: 10.1107/S1600576722002230.

[9] Y. Wang, Y. Shen, S. Mao, G. Cao, and R. M. Nelms, "Adaptive Learning Hybrid Model for Solar Intensity Forecasting," *IEEE Trans Industr Inform*, vol. 14, no. 4, pp. 1635–1645, Apr. 2018, doi: 10.1109/TII.2017.2789289.

[10] S. Aslam, A. Khalid, and N. Javaid, "Electric Power Systems Research Towards efficient energy management in smart grids considering microgrids with day-ahead energy forecasting," 2020, doi: 10.1016/j.epsr.2020.106232.

[11] C. Wan *et al.*, "Photovoltaic and Solar Power Forecasting for Smart Grid Energy Management," *CSEE JOURNAL OF POWER AND ENERGY SYSTEMS*, vol. 1, no. 4, 2015, doi: 10.17775/CSEEJPES.2015.00046.

[12] A. Rai, A. Shrivastava, and K. C. Jana, "An Empirical Analysis of Machine Learning Algorithms for Solar Power Forecasting in a High Dimensional Uncertain Environment," *IETE Technical Review*, vol. 40, no. 4, pp. 558–573, 2023, doi: 10.1080/02564602.2022.2136270.

[13] P. Bouquet, I. Jackson, M. Nick, and A. Kaboli, "AI-based forecasting for optimized solar energy management and smart grid efficiency," *Int J Prod Res*, 2023, doi: 10.1080/00207543.2023.2269565.

[14] J. D. Osorio *et al.*, "Forecasting solar-thermal systems performance under transient operation using a data-driven machine learning approach based on the deep operator network architecture," *Energy Convers Manag*, vol. 252, p. 115063, 2022, doi: 10.1016/j.enconman.2021.115063.

[15] H. Verbois, Y. M. Saint-Drenan, A. Thiery, and P. Blanc, "Statistical learning for NWP post-processing: A benchmark for solar irradiance forecasting," *Solar Energy*, vol. 238, pp. 132–149, May 2022, doi: 10.1016/J.SOLENER.2022.03.017.