School of Computing

National College of Ireland

Renewable power generation and weather conditions

MSc Research Project

Name : Shaik. Saida Hussain

Student ID : x23174323

Programme Name : MSc Data Analytics

Supervisor : Shubham Subhnil

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Shaik. Saida Hussain

Student ID: X23174323

Programme: MSc Data Analytics **Year:** 2024-2025
Module: MSc Research Project

Supervisor: Shubham Subhnil
Submission Due Date: 12th December 2024

Project Title: Renewable power generation and weather conditions

Word Count: 7898 **Page Count:** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: shaik.saida hussain

Date: 12th December 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Renewable power generation and weather conditions

Shaik. Saida Hussain

Student ID: x23174323

Abstract

Traditional methods in power prediction like linear regression or even simple machine learning models have struggled to handle the complexities in time series data. This study explores the prediction of solar power generation using two machine learning models: Random Forest and Long Short-Term Memory (LSTM) networks using a data set obtained from two solar power plants in India. Records in the dataset cover 34 days in total, during which, there are the power generation record per 15 minutes and the weather data or the ambient temperature, module temperature, and irradiation. The main purpose to estimate the TOTAL_YIELD of solar plant in respect of weather and power generation characteristics. The first transforming process is the data pre-processing in which data is cleaned and converted to a supervised form With the help of a feature extractor, data is divided into training and testing sets. For model implementation, the Random Forest Regressor is employed to predict the total yield and the LSTM model to analyze time series data. The performance of both models is assessed using mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE) and R² score. The outcomes reveal that the two models are reasonably accurate but the LSTM model provides better predictions with lesser error rates, which confirm the viability of time series forecasting. The final analysis is based on identifying the strengths of LSTM in terms of forecasting sequential data for renewable energy and, at the same time, the interpretability of Random Forest for features' importance.

Keywords: Solar Power Generation, Renewable Energy, Time-Series Data, Machine Learning, Long Short-Term Memory (LSTM)

1 Introduction

1.1 Background

Solar energy is clean energy harnessed from the radiation of the sun in the provision of power Hosseini and Wahid (2020). Solar energy is probably the most abundant and renewable source of energy to ever exist, thus holding a lot of promise in boosting the supply of energy globally as it minimizes the use of fossil fuels. Solar power is harnessed primarily through two methods: There are two types of solar systems: photovoltaic (PV) systems, which are built of semiconductor materials directly converting sunlight into electricity Abou et al. (2022), and the solar thermal systems which utilizes mirrors or lenses to focus sunlight and heat up, before creating electricity. As one of the sources of renewable energy, solar power plays a huge role in the shift to the cleanest power generation regimes in that they released little to no GHG emissions during their use. The demand to decrease carbon emissions and combat climate change has driven much innovation in solar product technology making it all the more affordable and efficient. Renewable power generation refers to generation of power from other sources such as wind, hydropower, geothermal, and even biomass solar power takes a central position among all the renewable power generation sources. Renewable power generation means the generation of power supply from renewable resource Maradin (2021) which is available naturally and hence does not require any exhaustible or renewable resource and has no long term adverse effect on environment. Traditional sources of energy such as electricity from the power grid has adverse effects on the environment with emissions of toxic compounds, pollutants, greenhouse gases and has finite reserve this makes the conservation of renewable energy sources seen as environmentally friendly, green, efficient as they help in cutting energy emissions on the environment.

Renewable electricity generation brings the benefit of diversification of the energy supplies, energy security decreased by national grids' dependence on imported fossil fuels Ayoo (2020), and new employment in renewable electricity industries. Solar power particularly has been growing at a very fast pace due to the following reasons; the technology is easily scalable, easily to implement, and the cost of the photovoltaic system continues to decline. It can be applied at residential level and at commercial level such as rooftops to large solar city farms that power some cities. The implementation of solar systems by governments, organizations as well as individuals is on the rise bearing in mind the current fight against global warming and scarcity of resources. Furthermore, off-setting environmental effects, renewable power from photovoltaic system provides chances on economic generate, energy security, and sustainable advancement making solar power a core feature in future energy mixes.

1.2 Research Motivation

At present, the demand for the clean and sustainable source of energy has raised awareness of renewable power for addressing environmental issues and for minimizing carbon emissions where renewable energy particularly solar energy concluded a significant role. Solar power generation though, is volatile in nature, as it depends closely on temperature, solar intensity or irradiation and other atmospheric conditions that vary with time so prediction of these is crucial for proper energy planning. This research is motivated by the lack of sophisticated models to accurately predict solar power generation for improved energy grid stability and resource management. Such data has high temporal dimension and traditional analyses are incapable of considering the rich interactions between variables or temporal dependencies for energy planning systems. More specifically, Exploratory Modeling techniques such as Random Forest as well as Earthquake Archetypes are the possible solutions through deep learning architectures such as Long Short-Term Memory (LSTM). Moreover, unlike other similar research studies, this shall endeavour to connect knowledge generated from academic innovation to the practical world to support the attainment of energy sustainability. Through the improvement of forecast accuracy of solar power, the study seeks to contribute to the process of reinventing the energy sector through encouraging use of renewable energy resources, thereby minimizing the use of fossil based energy, as well as reducing emission of greenhouse gases thus reducing climate change'sffects and consequently attaining a sustainable power sector.

1.3 Aim of the study

The aim of this study is to use machine learning models to evaluate the effect of weather factors to solar power output. More specifically, this research aims at deriving and assess machine learning models to predict power output from solar PV systems, using historical weather data and generation data sets with the help of Random Forest and LSTM models. In this context the investigation of the interaction between temperature, irradiation, and ambient conditions should allow for improving accuracy of solar powers' prediction and therefore the management and optimization of solar energy systems. The purpose of this paper is to find better solutions to current issues of forecasting for making optimal decision in renewable energy production.

1.4 Research Question

Which machine learning model—Random Forest or LSTM—provides the most reliable predictions for solar power output in varying weather conditions, and how can their performance be optimized for real-time energy management?

1.5 Research Objectives

There are some research objectives in this study which are as follows:

1. To design and implement robust machine learning models, specifically Random Forest and Long Short-Term Memory (LSTM), for accurately forecasting solar power generation using real-world data from solar plants.
2. To examine how key environmental factors, such as ambient temperature, module temperature, and solar irradiation, influence solar power generation, and incorporate these variables into the predictive models.

1.6 Structure of the Report

This study provides a comprehensive analysis of solar power generation prediction using machine learning models. The structure is outlined as follows:

Chapter 1 Introduction: Introduction to the study, problem definition, objectives, and significance of solar power generation forecasting.

Chapter 2 Literature Review: A review of existing research on solar power prediction using machine learning and deep learning models.

Chapter 3 Methodology: Data collection, preprocessing, feature engineering, model selection, and implementation details.

Chapter 4 Design Specification: Workflow diagram and design methodology for model implementation.

Chapter 5 Implementation: Detailed implementation of Random Forest and LSTM models, including code, hyperparameters, and training process.

Chapter 6 Evaluation: Performance evaluation of the models using metrics like MSE, RMSE, MAE, and R^2.

Chapter 7: Conclusions and Future Work: Summary of findings, implications for renewable energy, limitations, and proposed future research.

2 Related Work

2.1 Renewable Power Generation

Renewable energy technology means the use of power from the sun, wind, water and bio mass to generate electricity thereby providing clean energy for use instead of conventional sources such as coal Hafezi and Alipour (2021). Among renewable energy resources, solar energy occupies a leading position and operates with the use of photovoltaic (PV) facilities Izam et al. (2022). These systems are affected factors like irradiance, ambient temperature, and shading occasions, and their efficiency differ especially during seasons and geographical region. In a similar manner, wind harness pulls its power from the kinetic energy of the winds which is converted by turbines. Regarding the energy output it is particularly sensitive to the wind speed and direction, air density and turbulence. Hydropower harnesses water movement to produce electricity, this making its yields influenced by seasonal and climactic changes, biomass energy is obtained from biological processes which if efficiently properly managed, is renewable and environmentally friendly. Nonetheless, fluctuation in power production emanating from renewable energy sources which rely on weather factors becomes a major challenge in relation to grid stability and prediction Medina et al. (2022). However, to overcome these challenges, the renewable energy systems in present time vary with actual weather conditions, accurate predictive models, and have analytics tools for better output. The introduction of renewable energy resources into electricity networks has necessitated drawing up of methods such as batteries and pumped hydro storage for energy storage Bhayo et al. (2020) during optimal production to be used during high demand. However, current complexity issues that affect the efficiency of renewable power systems include intermittency, regional limitations, and infrastructure. As the new studies show, the need to advance the technologies that will allow for better predictions for the weather conditions, application of the combined systems of the power supply, and the flexibility of the grid to increase the share of renewables. As advances in technology occur more regularly, renewable power generation is expected to assume an increasingly important component in satisfying the global energy needs and at the same time reducing greenhouse gas emissions which are causing climate change hence leading to the enhancement of sustainable development universally.

2.2 Impact of Weather Conditions on Power Generation

The impact of these weather conditions on power generation is high as several renewable resources such as solar, wind and hydro power are influenced by these natural conditions Zhang et al. (2022). In solar power generation, the most critical factors affecting photovoltaic (PV) is solar radiation, ambient temperature, and cloud cover which makes it to be most efficient under clear sky conditions and moderated temperatures. Seasonal and diurnal changes are also very significances, with shorter daylight hours and lower intensity of sun light in winter reducing the systems' efficiency Bellia et al. (2020). As with any institution utilizing atmospheric conditions wind energy has its key variables with wind speed, direction, air density, and turbulence being paramount to turbine performance. Wind speeds are usually associated with increased energy generation but gustiness or calm spells are known to disturb efficiency. For the hydropower resource, these conditions of weather would affect water supplies which in turn will affect the turbines and the reliability of the energy supply Borowski (2022). Weather factors are also unpredictable which cause fluctuations in energy generation; there is a need to predict forthcoming weather patterns for renewal energy sources. Recent techniques use NWP models, satellite imagery, and machine learning to improve the precision of forecasts and subsequently increase the matching of power production with consumption patterns. Climate change, on the same note, increases variability due to storms, heat and fluctuating precipitation patterns that may pose threat to renewable energy assets Xu et al. (2024). Any of these challenges needs the incorporation of reliable prediction models with renewable energy resources, enhancements of the flexibility of the distribution network, and the utilization of storage devices to counter the impacts of intermittency. On the same note, the dependency on weather also present probabilities of tapping a variety of renewable sources adequately through the integration of solar, wind, and hydro power for diverse and stable resource base. Therefore, the necessary implication is to work on improving or at least effectively predicting the correlation between weather conditions and power generation for improved integration of renewable power into the energy sector.

2.3 Machine Learning Models in Renewable Power Generation

Artificial neural networks have also played a critical role in improving the forecast of and interface between integration of renewable energy systems. The Random Forest, GBM, SVM, kNN, Decision Trees, Extra Trees, LSTM, and Multilayer Perceptron Regression models have been common with wind and photovoltaic power generation forecasting. Stacking of Random Forest and X GBOOST Combination Some of the other ensemble methods values include: These models, with efficient algorithms such as the dragonfly algorithm and the Cauchy mutation operator, shown enhanced assuredness in the power prediction system, which assists in power integration and system reliability in the unpredicted renewable power systems.

First study which is given by Singh et al. (2021) who suggested a comparative analysis of five most powerful robust regression machine learning algorithms – random forest, gradient boosting machines (GBMs), k-nearest neighbor (kNN), decision trees, extra tree regression for improving the accuracy of the short-term forecasting of wind energy generation in Turkish wind farms in the western part of this country. Historic wind speed and wind direction as inputs on the models, wind speed and contribution through polar diagram and wind speed turbine output through scatter curves. Another difficulty was to meet the high accuracy requirement of the forecasts because of the volatility of the wind patterns. The outcome revealed that of the employed strategies, gradient boosting machine regression algorithm generated the best forecast of the average wind energy efficiency to which forecasting error revealed low deviations from the actual turbine outputs.

Another study by Mahmud et al. (2021) who put forward a machine learning methodology for predicting the PV power generation to respond to the problem of generation intermittency for complicating the grid control and operation. Utilizing different environmental factors, and analyzing various machine learning models the study concentrated on Alice Spring, Australia a place abundant in solar energy Linear Regression, Polynomial Regression, Decision Tree Regression, Support Vector Regression, Random Forest Regression, Long Short-Term Memory, and Multilayer Perceptron Regression models were tested. The approach involved comparison of performance under baseline and uncertainty contexts while assessing effects of normalisation on 3-way forecast accuracy across various measures of performance. One such hardship has been generation volatility to support dispatch with greater confidence at day- and hour-ahead basis. It was established that Random Sampling with Regression analysis provided better results than other models for the dataset to forecast time – ahead PV power variability for grid operators to consider in selecting proper forecasting algorithms.

In a similar manner Martinez et al. (2021) gave a background of status of application of ML in the manufacturing sectors of concerns to sustainability and environment such as renewable energy technologies for storage and conversion including the solar energy, wind energy, hydro energy, bio-mass energy, smart grid, catalysis and power storage and distribution. The study based its projection on the artificial neural networks because of their ability to generalize their lessons; and there seems to be growing interest on the ML techniques as science, mathematics, and engineering explore the realm of artificial intelligence. Such issues relate to generation, storage, and protection of data, which is crucial for expanding the use of highly effective ML algorithms in the spheres of energy among the representatives of the energy sector for the purpose of implementing challenging energy management initiatives. The findings indicated that improvements in the classes of unsupervised and reinforcement learning will improve the capability of the systems to interface with the ML systems along with advances in the data science domains of big data analytics, new sensors and reliable algorithms. However, it was agreed that 5G and specific algorithms will fasten the process of integrating sustainable ML in energy solutions.

Deb et al. (2020) introduced a new approach for handling congestion in distribution systems due to the uncoordinated charging of Plug In Electric Vehicles (PEVs), which is affecting both transport and power domains. The study proposed a charging scheduling management technique for both G2V and V2G for analyzing high PEV integration into the distribution system. For state-of-charge prediction of PEV batteries, the gradient boosting regression tree technique was used. The goal was to bring down the overall cost as close to the lowest level as possible while at the same time allowing for as many PEVs as possible into the distribution network so as not to load the feeders. The above strategy was implemented at the workplace car park where the chosen car park was supplied by electricity from the grid and grid connected photovoltaic generation in an industrial area of IEEE 38 bus radial distribution system. The paper has qualitatively compared the performance of the system based on the designed system together with the solar-powered car park verses without the PEV and found out that the proposed strategy would go a long way in the society to posing an efficient way of discouraging penetration of PEVs and at the same time reduce the over-concentration of the distribution lines.

To mitigate the deficiencies, which the wind power short-term forecasting possesses, non-stationary and stochastic in nature Li et al. (2020) put forward a short-term forecast model built with support vector machine (SVM) integrated with an enhanced dragonfly algorithm. The work contributed a new learning factor and differential evolution to enhance the dragonfly algorithm application. This new algorithm was then applied to fine-tune other parameters of the SVM in order to increase the prediction accuracy. The empirical part of the proposed model was implemented using real data derived from the La Haute Borne wind farm at France. The findings highlighted the superiority of the model under evaluation to other methodologies, back propagation neural network and Gaussian process regression in terms of prediction efficiency and preparedness for short term power forecasting.

Lin et al. (2020) developed a new moth flame optimization for support vector machine for accurate prediction of photovoltaic energy as the output of solar power flexibilities due to variability as grid connected solar generation increases. For the purpose of improving the model, an inertia weighting strategy was incorporated into the search of the population location to optimize the ratio of search and mining throughout the optimization progress and external Cauchy mutation operator to enhance the population difference and avoid falling into local optimal solution. The study also enhanced the input data via Grey Relational Analysis in respect to many factors affecting PV power generation involving meteorological factors. The above characteristics were tested using the number of orthogonal test functions and actual data from a photovoltaic power station in Australia, where the model exhibits higher optimization performance than other models. The proposed method enhances accuracy of photovoltaic energy forecasting, minimizes the effect of photovoltaic power integration on the grid, and provides support for the reliability of the system.

Finally, Banik (2024) developed a novel ensemble model consisting of Random Forest and XGBoost for different issues related to the integration of renewable energy into the grid; to improve the predictive error in case of BP fluctuation. That is why work emphasized the importance of forecast actions needed to mitigate environmental and economic consequences of integration of renewable energy into networks along with the impact on stability of grids and users' life. The proposed ensemble was successfully applied on the dataset from Agartala City, Random Forest model to predict target variable based on input parameters and XGBoost was used to boost the Random Forest output. An initial meta-model followed by a logistic regression, further fine-tuned these predictions to achieve best accuracy. When tested for R2 and RMSE, the proposed model secured a success rate of 99%

proving the model to be most efficient and accurate for both short duration and long duration renewable energy forecasting.

2.4 Deep Learning Models in Renewable Power Generation

There are few research studies for deep learning in renewable power generation. The first study presented by Saini et al. (2020) where the author suggests that a diverse selection of machine learning algorithms be applied for the precise prediction of Wind speed, and the assessment of the produced energy as vital in applying Wind energy in the grid system. Because of the variability in wind speed, precise prediction is critical to avoid an unreliable power supply. The research methodology includes using several machine learning models on hourly wind speed data for Jodhpur, Rajasthan (India). Among the major limitations in this research work is the accuracy of predicted wind speeds, which is a common problem with many works in this area. The numerical outcomes revealed that the Gated Recurrent Unit (GRU) algorithm had better prediction results than other models based on the measurements obtained.

Another study stated by Miraftabzadeh et al. (2023) who provides a data-driven approach for identifying daily PV power generation profiles through deep learning and clustering-based methodologies in order to overcome the vagaries that characterise solar power generation. The approach introduces high dimensionality of temporal features of daily PV output power and maps them to a lower dimension latent space using a deep learning autoencoder. Subsequently, more specified dominant pattern detection is performed with the application of the Partitioning Around Medoids clustering technique, and six different patterns in PV power generation are defined. To this end, a new algorithm is presented to reconstruct such patterns from the noisy measurements in their original subspace. When tested on two datasets, four of the identified patterns correlate highly (greater than 95%) with the specific group of weather conditions, including sunny, mostly sunny, cloudy, and negligible power generation days that is observable within approximately, 61 per cent of the analyzed period. Such patterns are expected to be valid in other locations, as well. The mentioned patterns can be integrated into forecasting models, optimize energy consumption, or help with the introduction of energy storage or demand response programmes. The problem that emerged in this research was to capture and model the temporal weather variabilities for reconstructing PV power generation.

2.5 Conclusion

The surveyed papers also discuss the possibility of utilising Random Forest, SVM, and Gradient boosting to predict renewable energy production adequacy and integrate it into the grid. Issues like ensemble learning that involves supplementing the elements of several learning algorithms, and optimization problems like the dragon-fly algorithm of data preprocessing Grey Relational Analysis. Possible enhancements for the future may include the use of unsupervised and reinforcement learning, as well as inclusion of real time data from the smart grids into the system and modification of the hybrid models in order to minimally decrease the prediction errors and to increase the scalability of renewable energy systems in the complex conditions.

3 Methodology

3.1 Importing Libraries

In this analysis, a number of important libraries in Python have been imported for use in the analysis of the data. One of the packages used is pandas library and this is helps in loading of data, cleaning, and preprocessing of data for ready analysis. NumPy is used in the manipulation of numerical properties more especially in array mathematics and mathematical functions critical in data transformation. As for the visualization, matplotlib.pyplot allows one to make static/interactive/aninated plots or even all together and seaborn focuses on more statistical graphic features including heatmaps & pair plots that may help to gain a better understanding of your data and investigate it more thoroughly.

3.2 Data Preprocessing

In this section, the first step toward feeding the data to a machine learning algorithm is done through preprocessing. The remaining features in the dataset are more valuable and include, therefore features like 'PLANT_ID', 'SOURCE_KEY_gen', 'SOURCE_KEY_weather', and 'hour' are dropped. Data obtained as a result is divided into training and test samples using the train_test_split function from the sklearn.model_selection module here we take 80% of the data for training and 20% for the test. This means that by setting the random_state to 42, the split can be replicated again and again.

3.3 Data Transformation for Time-Series Modelling

To facilitate time-series forecasting and convert the dataframe, a special function df_to_X_y is made. Accepts: dataframe, window_size – the number of observations to look back in time before making the prediction. The given dataframe is then converted into a NumPy array and sliding windows of size window_size are created, which acts as input features (X) and the related labels, the target values (y). This conversion is important since the temporal data needs to be fed into machine learning where temporal dependencies have to be learned by the model.

3.4 Dataset Description

The dataset is made up of solar power generation and weather sensor data from two solar plants in India, with data recorded every 15 minutes, with the study spanning 34 days. Power generation details of each plant are provided in the Plant_1_Generation_Data.csv file and Plant_2_Generation_Data.csv file containing similar columns as DATE_TIME, PLANT_ID, SOURCE_KEY, DC_POWER, AC_POWER, DAILY_YIELD and TOTAL_YIELD. The DC_POWER and AC_POWER indicate the direct current and alternating current from the inverters during each 15 min interval reported in kW. DAILY_YIELD is another column that is a running total of watts produced per day; TOTAL_YIELD is the total wattage produced by that particular inverter in accord with date and time. Further, the first dataset from two plants, namely Plant_1_Weather_Sensor_Data.csv and Plant_2_Weather_Sensor_Data.csv contain the weather parameters collected for each plant, though they contain DATE_TIME, PLANT_ID, SOURCE_KEY, AMBIENT_TEMPERATURE, MODULE_TEMPERATURE and IRRADIATION. These weather meters detect the temperature of the surrounding environment, the temperature of the solar panels, as well as amount of radiation on the solar power plant location which affect efficiency of solar power generation. These data are useful in the study of how the weather influences the productivity of solar power and enhance climate modeling for electricity production.

3.5 Data Visualization

To analyse the performance of Plant 1, Plot 3.1 illustrates a time-series of the DC power and irradiance levels in Plant 1 over the production period. The x-axis is the date and y-axis to the left is the DC power and the y-axis to the right depicts the irradiation. The blue line means the DC power and the red line means the irradiation. Figure 1 is a line graph that plots DC power against irradiation throughout the time span in question. This kind of joint line graph is used to describe the changes in two correlated factors over time as the recipient is able to compare as well as interpret the patterns and characteristics of the data trends. The simultaneous use of these two parameters allows explaining the impact of irradiation to DC power generation and providing effective control over the photovoltaic plant performance.

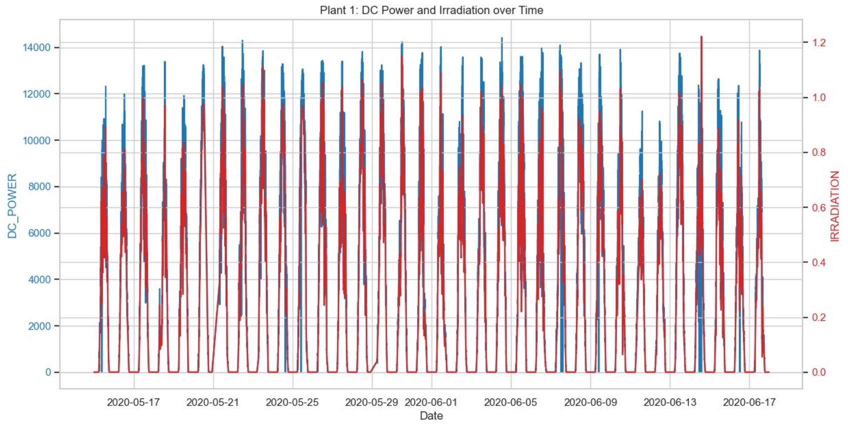

Figure 3.1: Plant 1: DC Power and Irradiation over Time

This presents a time-series plot in figure 3.2 below that shows air temperature and module temperature of plant 1. The horizontal axis is the date whereas the vertical axis to the left is the ambient temperature and the right axis is for the module temperatures. The blue line indicates the ambient temperature while the red line on the figure indicates temperature of the module. This joint dual plot provides the means to visualize the correlation between the ambient and module temperatures during the analyzed time frame. In the case of the photovoltaic modules, there is a need to monitor both the ambient and the module temperature since the current through the module depends on the temperature difference between the module and the ambient temperatures. When both of these related variables are parallelly plotted, the user can find out any trends, patterns or shifts that might be useful when considering the operation and maintenance of the plant.

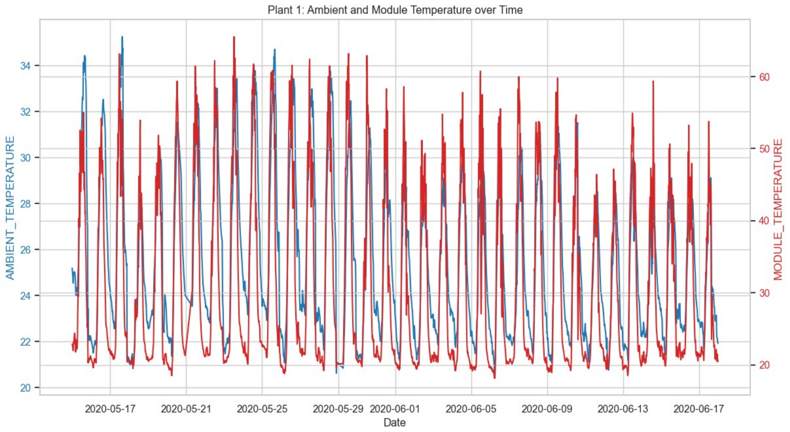

Figure 3.2: Plant 1: Ambient and Module Temperature over Time

Figure 3.3 presents information on the trend of the DC power against irradiation for Plant 2. The line chart with the combined data is defined by the DC power values represented on the left y-axis and irradiation values represented on the right y-axis, with the date represented on the abscissa. Blue line shows the DC power and the red line shows the irradiation level. This way the variation in the DC power generation that is due to pulsation in irradiance levels within the stipulated time period is visualizable. The availability of these two related figures simultaneously allows for certain patterns, trends and even possible correlations to be observed and analysed, which are extremely significant towards enhancing performance of the photovoltaic plant.

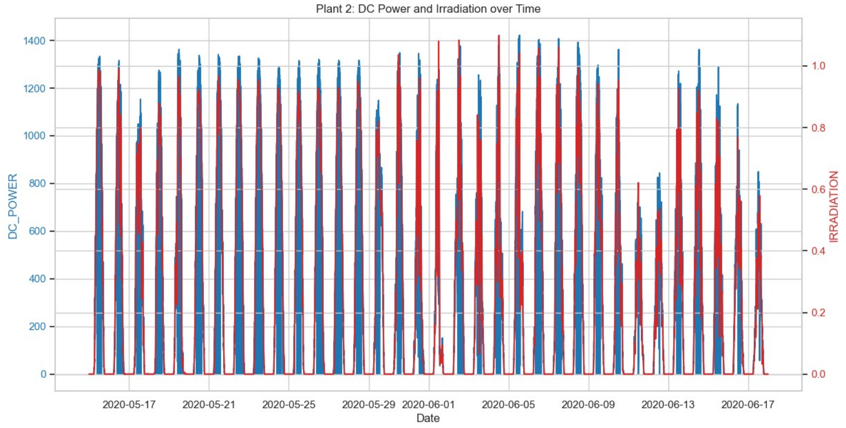

Figure 3.3: Plant 2: DC Power and Irradiation over Time

Figure 3.4 shows the trend of the ambient temperature against the module temperature for Plant 2. The x-axis is the date while the left y-axis represents ambient temperature, right y-axis represents thermo module temperature. The blue line represents the ambient temperature, where as the red line represents the module temperature. With the help of this combined line plot it becomes possible to control both temperature lines at the same time which is essential to control and investigate the efficiency of the photovoltaic modules because the module temperature affects the amount of power being fed in the line remarkably.

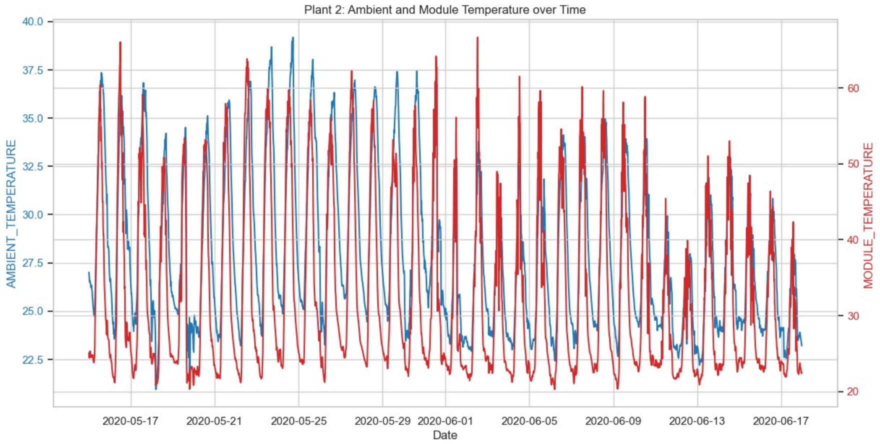

Figure 3.4: Plant 2: Ambient and Module Temperature over Time

Figure 3.5 shows correlations between Plant 1 and other parameters, and the correlations between all the parameters are presented in the form of a correlation matrix in Figure 3.5. The matrix shows the correlation coefficient figures ranging between -1.0 and +1.0: where -1.0 is negative correlation; 0.0 is no correlation; + 1.0 is positive correlation. All the variables of the investigated power plant like DC power, AC power, daily production, total production, temperature, module temperature, and irradiation can be easily depicted and their inter relation can also be recognized instantly. The correlation matrix helps them know how some of these factors are related and the information from the matrix can be useful in trying to analyze and improve the performance of the plant.

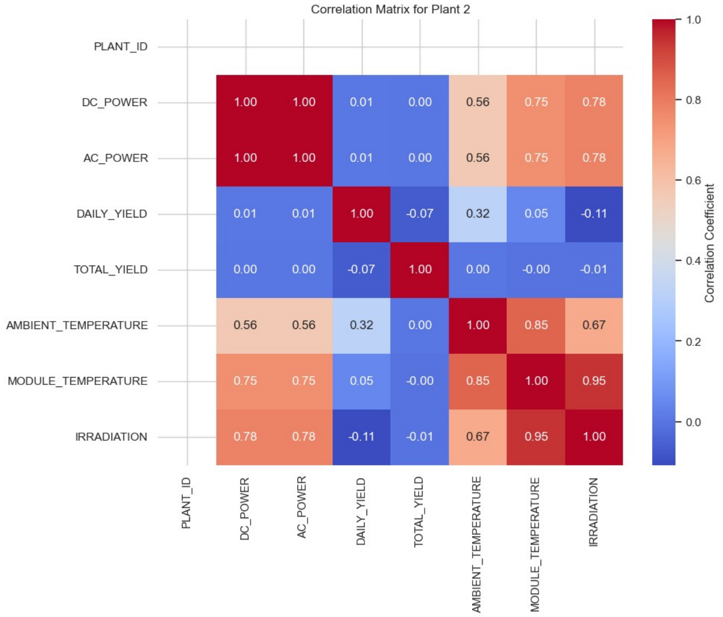

Figure 3.5: Correlation Matrix for Plant 1

The matrix shown in figure 3.6 represents the Plant 2 and shows the interconnection between two parameters. The matrix gives the correlation coefficients, which lie between minus one and plus one with minus one meaning negative correlation and plus one meaning the positive correlation. This simplifies the means by which the relationship between the DC power, AC power, daily yield, total yield, ambient temperature, module temperature and irradiation levels can be determined. The correlation matrix gives information which can be useful for further analysis and improvement of the efficiency of Plant 2.

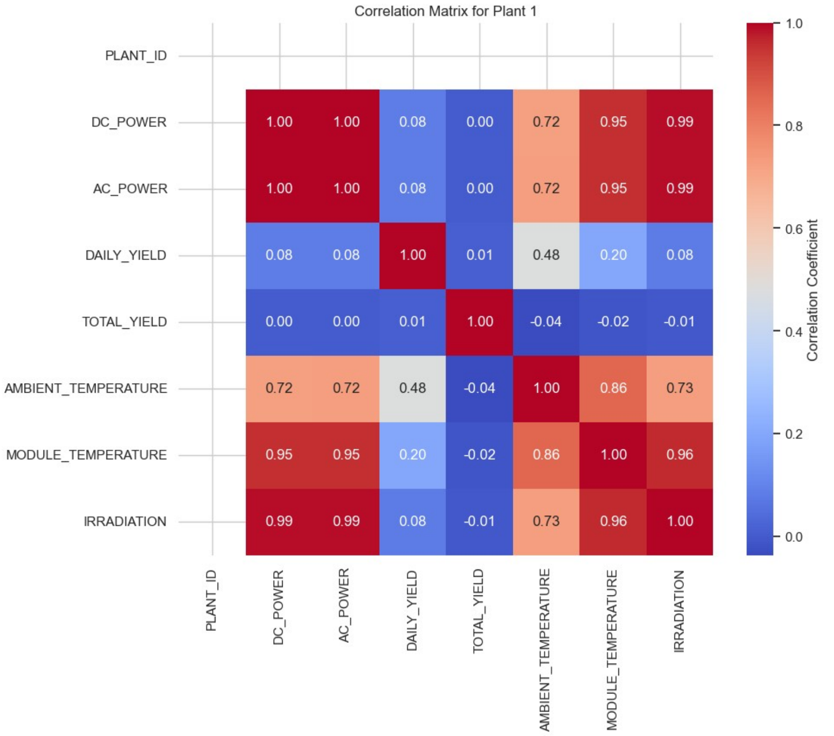

Figure 3.6: Correlation Matrix for Plant 2

4 Design Specification

The proposed workflow diagram has been shown in Figure 4.1 indicates how analysis is done on a solar power generation data set. The first of these include data preprocessing in which; we eliminate columns with no significance, we manage cases of missing values and we resample the data if required. This is succeeded by feature engineering in which related characteristics from the database are extracted with the view of training a machine learning model.Subsequently, to construct samples for training the model, a sliding window strategy is incorporated with the appropriate window length to capture temporal dynamics in the data. For more well-managed machine learning, data is divided into the training set and evaluation set, whereas the training data set for fitting the model and the evaluation data set for evaluating its performance.

In the model training step, the adopted Random Forest algorithm is a common instance in the ensemble learning category. Last but not the least, hyper tuning is followed by the evaluation of the trained model using correct and relevant metrics. The workflow may also consist of the LSTM component – Long Short-Term Memory, recurrent neural network designed for time series data analysis.In short, this workflow is a systematic work flow of facilitating data preprocessing, feature engineering, model training and model evaluating in order to improve the accuracy and robustness of the predictive model of future solar power generation.

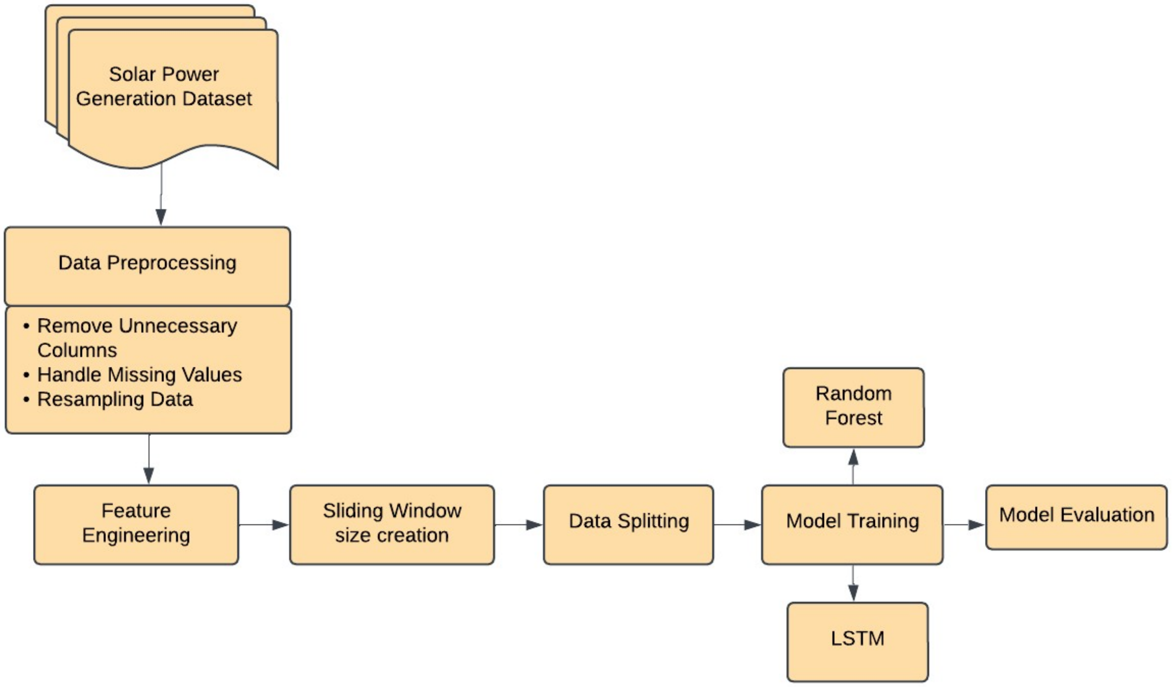

Figure 4.1: Proposed Workflow System

5 Implementation

5.1 Implementation on Random Forest

The Random Forest model is used namely RandomForestRegressor from the scikit-learn library. First, all the non-pertinent variables are omitted to exclude such and only the Select Features column is utilized in the target column, TOTAL_YIELD. The dataset is split into training and testing sets using an 80:20 ratio with train_test_split I note that the authors have used the train test split ratio of 20 from this model. It involves choosing the training set in order to fit the model called Random Forest with 100 estimators which are decision trees, every estimator is a separate decision tree and it is trained on a boot strap random sample of the training set. The model is then used to predict the total yield on the test set based on the share of yield of each product. Performance is evaluated using multiple metrics: The results also reveal the generalization capabilities of the Random Forest model to unseen data. Furthermore, for the insights of which features are more important in the predictions, the feature importance of the model is calculated. This makes the model versatile in handling relations as well as non-linear in the dataset, and a perfect one for predicting the power generated from solar.

5.2 Implementation on LSTM

The application of LSTM model starts with the pre-processing of the time- series data. Although using a simple dataset, the data is transformed into a sequential form where the model learns a mapping from a window of past observation (for instance 5 values) to the next values. LSTM Model implemented using Keras Sequential API is used in this data set. Two LSTM layers are employed to identify the temporal dependencies recurring in data, and are succeeded by Dropout layers to minimize over-learning. To add non-linearity a Dense layer is inserted followed by output layer using sigmoid hence predicting a continuous value of total yield. This is is built using Adam optimizer with learning rate set to be 0.001 and mean squared error loss function. This is used so that training can be stopped when validation loss does not decrease for a fixed number of epochs While this reduces the learning rate when the validation loss plateaus. This model is trained using the fit function, 64 batch size for 100 epochs.

5.3 Model Performance Evaluation and Visualization

The frequency distribution of the residuals is given by a histogram of the residuals which is shown in figure 5.1. The residuals are calculated as the discrepancies between the endogenous states which have been forecasted and the outcomes that have been realized in a given set of data. Where the x-axis shows a range of residual values and the y-axis shows the frequency or count values of those residuals.

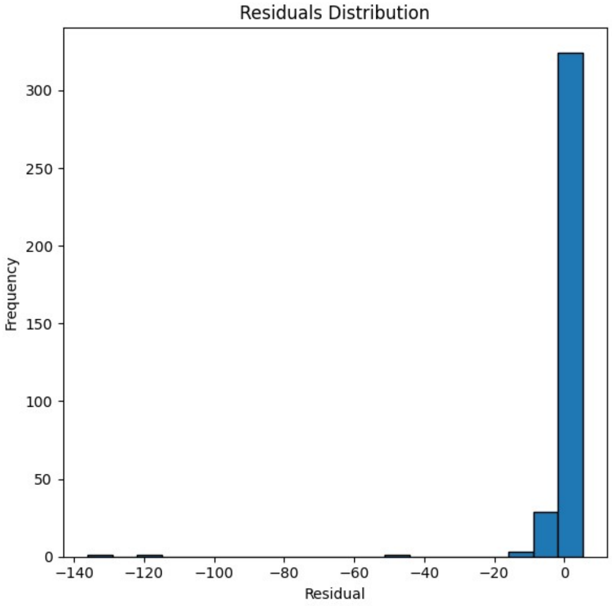

Figure 5.1: Residuals Distribution

Figures 5.2 shows the scatter plot of the predicted values and the actual values of the variables. The x-axis depicts the actual prices as the real scales and y-axis depicts the predicted prices as the scales. Such a type of plot is used most frequently for the assessment of the performance of any given predictive model since it offers a graphical way of monitoring the degree at which actual values differ from the predicted values. The higher the values of κ distributions are close to the line with the slope of 45 degrees, the higher is the ability of the model to perform the target variable forecasting.

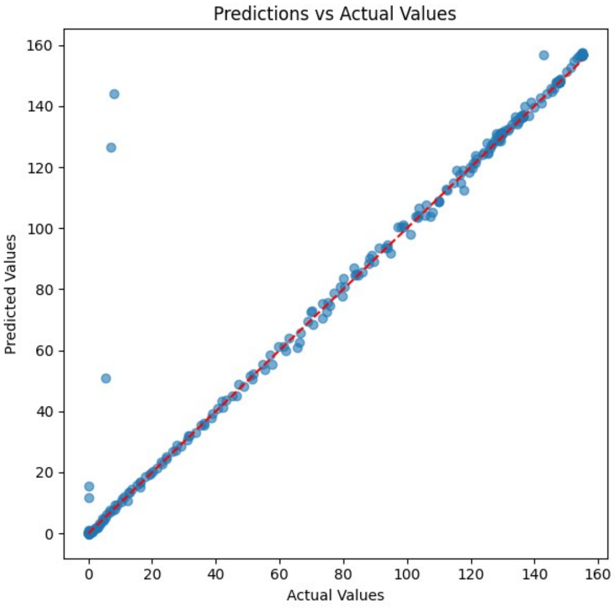

Figure 5.2: Predictions vs Actual Values

As shown in figure 5.3, the actual values are plotted against the predicted values over the time. The first axis will be the set of data points labelled by a index or sequence number, the second axis will be values corresponding to the former ones. The blue-coloured line focuses on the actual values while the red line refers to the predicted ones. This line plot shall enable us to compare between the predicted outcome by the model and the true observed outcome throughout the length of the data.

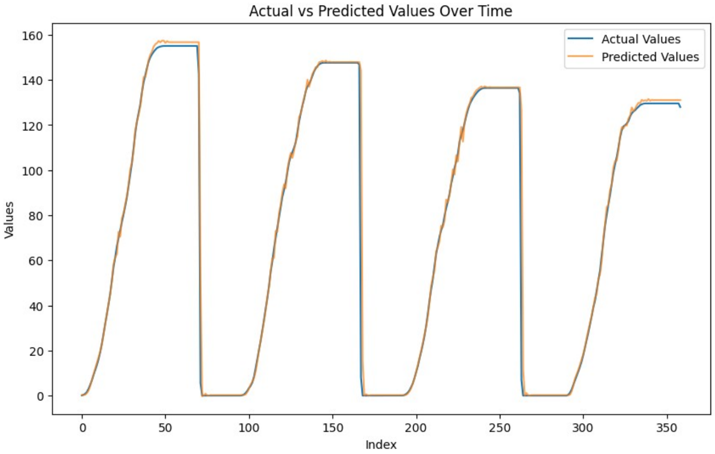

Figure 5.3: Actual vs Predicted Values Over Time

6 Evaluation

6.1 Case Study 2: Random Forest

Random Forest Regressor was applied by means of the scikit-learn library with 100 estimators for the model and the fixed random state to ensure the result is consistently rebuildable. The obtained features were used to train the model and provide predictions of the test set. An assessment of the performance of the proposed model yielded impressive statistics; MSE 3.37 trillion, MAE 931,211.59, and a coefficient of determination of 0.997. The calculative R^2 standing suggests that the model accounted for 99.7% of variance of the data explaining its efficiency and effectiveness. Moreover, the expressed meaning of error metrics provides evidence of Random Forest model in terms of capacity to capture intricate interdependencies existing in the data: by examining their low values. The Random Forest approach is also interpretable, and generates a feature importance, which is a significant advantage in Solar power generation analysis. For the given non-linear and multi-variate dependencies, the Random Forest has been proved to be effective and thus it becomes a suitable algorithm for use in renewable energy prediction. The performance of the model clearly demonstrates the possibility of its practical application in real-world problems that require accurate prediction of power generation to improve energy management systems and the use of resources, as well as to increase the reliability of the grid.

6.2 Case Study 2: LSTM

The analysis of the solar power generation prediction based on the Long Short-Term Memory (LSTM) model was enabled by the model's capacity to consider the temporal dependency in the time series. Data preprocessing in this research involved partitioning the dataset into training set and test set, and constructing the model as a supervised learning model by forming input-output pairs where each five consecutive time steps were used to predict future values. An LSTM model was built with the Keras Sequential API stacked with two LSTM layers of 128 and 64 units and two Dropout layers to prevent overfitting. It also contained dense layers to extract features from and the final linear layer that predicts the value of the continuous target variable TOTAL_YIELD. The model used the Optimizer Adam, a Learning Rate 0.001 for mean_squared_error loss to try and reduce the general prediction errors. Training was done over a hundred epochs with a batch size of 64 together with callbacks that include the EarlyStopping when the model is overfitting, ModelCheckpoint for saving the best model and ReduceLROnPlateau for the learning rate when the model's accuracy is plateauing. During the final evaluation, in training we achieved a training loss of 183.7992 and in validation a validation loss of 24.1852. On the test set, the model has given MSE=100.3860, RMSE = 10.0193, MAE = 1.8173 and $R^2 = 0.9731$, which means that this model is right 97.3% of the time in predicting the data variance in the test set. Autopilot of the actual values versus predicted values and residual analysis also supported the proficiency and accuracy of the model designed. The results indicate LSTM ability to capture temporal dependencies inherent in the solar power generation data, bringing LSTM as a valuable tool in renewable energy forecasting and improving decision-making processes in the renewable energy production and grid stability management.

Table 6.1: Performance Comparison of Random Forest and LSTM Models

Metric	Random Forest	LSTM
Mean Squared Error (MSE)	3,370,311,740,309.5015	100.3860
Root Mean Squared Error (RMSE)	1,836,055.97	10.0193
Mean Absolute Error (MAE)	931,211.59	1.8173
R² Score	0.9970	0.9731

6.3 Discussion

From the results obtained in this study, Random Forest was slightly better in detecting the feature importance with the $R^2 = 0.997$ while LSTM was better placed in the time series with the RMSE = 10.0193. Both models succeeded in responding to the set research question in the sense that accurate predictions of solar power under different weather conditions was achieved. Random Forest provided high interpretability of the important features, while LSTM provided the capability of handling the time series information crucial for forecasting the renewable energy. Such models' synergy confirms the possibilities of developing real-time energy management, optimizing the integration of facilities with solar power, and stabilizing the sustainable energy systems.

7 Conclusions and Future Works

Conclusion

Consequently, Random Forest and LSTM models used in this research provided the ability to predict solar power generation. Random Forest model was best suited for variant explanation with an R^2 of 0.997 while LSTM model proved best for error minimization with RMSE of 10.0193. Since LSTM has made a significant consideration to temporal dependencies, it is ideal for time-series cases like the predictions shown here; Random Forest is strong and easily explainable to show which features impact solar energy generation significantly. Aid of these models can improve renewable energy management systems with the assistance of efficient estimating.

Implications

The practical meaning and application of these models has great importance. Weather prediction of solar electricity production enables the enhancement of grid reliability, diversification of energy sources and efficient utilization of resources for maintenance purposes. In this study, machine learning models, such as Random Forest, as well as sequence-focused models, such as LSTM, are considered to achieve an optimal level of accuracy that can be explained using a hybrid approach in renewable energy forecasting.

Future Work

As for the limitations of this study, future work can pay more attention to the way of importing real-time weather data to enhance the accuracy of the model and carry out more comparative studies of the model in different regions. Building models that incorporate the two strategies explored here – Random Forest and LSTM- could lead to even better results. For even more progress in solar energy forecasting it is also worth to employ other improved deep learning architectures, for example Transformer-based models.

Limitations

Imperatively, those include the fact that the results of the study are based on a modest sample set which encompassed only thirty-four days. The models also presuppose constant context and activity, environmental or operational, which might fail to address certain scenarios, for example, having a faulty sensor or a storm. To overcome these limitations of datasets, the technique of anomaly detection, and domain-specific model tuning are expected with larger datasets for real-world implementation.

References

1. Singh, U., Rizwan, M., Alaraj, M. and Alsaidan, I., 2021. A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments. *Energies*, *14*(16), p.5196.
2. Mahmud, K., Azam, S., Karim, A., Zobaed, S., Shanmugam, B. and Mathur, D., 2021. Machine learning based PV power generation forecasting in alice springs. *IEEE Access*, *9*, pp.46117-46128.
3. Rangel-Martinez, D., Nigam, K.D.P. and Ricardez-Sandoval, L.A., 2021. Machine learning on sustainable energy: A review and outlook on renewable energy systems, catalysis, smart grid and energy storage. *Chemical Engineering Research and Design*, *174*, pp.414-441.
4. Deb, S., Goswami, A.K., Harsh, P., Sahoo, J.P., Chetri, R.L., Roy, R. and Shekhawat, A.S., 2020. Charging coordination of plug-in electric vehicle for congestion management in distribution system integrated with renewable energy sources. *IEEE Transactions on Industry Applications*, *56*(5), pp.5452-5462.
5. Li, L.L., Zhao, X., Tseng, M.L. and Tan, R.R., 2020. Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm. *Journal of Cleaner Production*, *242*, p.118447.
6. Lin, G.Q., Li, L.L., Tseng, M.L., Liu, H.M., Yuan, D.D. and Tan, R.R., 2020. An improved moth-flame optimization algorithm for support vector machine prediction of photovoltaic power generation. *Journal of Cleaner Production*, *253*, p.119966.
7. Banik, R. and Biswas, A., 2024. Enhanced renewable power and load forecasting using RF-XGBoost stacked ensemble. *Electrical Engineering*, pp.1-21.
8. Hafezi, R. and Alipour, M., 2021. Renewable energy sources: Traditional and modern-age technologies. In *Affordable and clean energy* (pp. 1085-1099). Cham: Springer International Publishing.
9. Izam, N.S.M.N., Itam, Z., Sing, W.L. and Syamsir, A., 2022. Sustainable development perspectives of solar energy technologies with focus on solar Photovoltaic—A review. *Energies*, *15*(8), p.2790.
10. Medina, C., Ana, C.R.M. and González, G., 2022. Transmission grids to foster high penetration of large-scale variable renewable energy sources–A review of challenges, problems, and solutions. *International Journal of Renewable Energy Research (IJRER)*, *12*(1), pp.146-169.
11. Bhayo, B.A., Al-Kayiem, H.H., Gilani, S.I. and Ismail, F.B., 2020. Power management optimization of hybrid solar photovoltaic-battery integrated with pumped-hydro-storage system for standalone electricity generation. *Energy Conversion and Management*, *215*, p.112942.
12. Zhang, Y., Cheng, C., Yang, T., Jin, X., Jia, Z., Shen, J. and Wu, X., 2022. Assessment of climate change impacts on the hydro-wind-solar energy supply system. *Renewable and Sustainable Energy Reviews*, *162*, p.112480.
13. Bellia, L., Acosta, I., Campano, M.Á. and Fragliasso, F., 2020. Impact of daylight saving time on lighting energy consumption and on the biological clock for occupants in office buildings. *Solar Energy*, *211*, pp.1347-1364.
14. Borowski, P.F., 2022. Water and hydropower—Challenges for the economy and enterprises in times of climate change in Africa and Europe. *Water*, *14*(22), p.3631.
15. Xu, L., Feng, K., Lin, N., Perera, A.T.D., Poor, H.V., Xie, L., Ji, C., Sun, X.A., Guo, Q. and O'Malley, M., 2024. Resilience of renewable power systems under climate risks. *Nature Reviews Electrical Engineering*, *1*(1), pp.53-66.
16. Hosseini, S.E. and Wahid, M.A., 2020. Hydrogen from solar energy, a clean energy carrier from a sustainable source of energy. *International Journal of Energy Research*, *44*(6), pp.4110-4131.
17. Abou Jieb, Y. and Hossain, E., 2022. Photovoltaic Systems.
18. Maradin, D., 2021. Advantages and disadvantages of renewable energy sources utilization. *International Journal of Energy Economics and Policy*, *11*(3), pp.176-183.

19. Ayoo, C., 2020. Towards energy security for the twenty-first century. *Energy policy*, pp.15-40.

20. Saini, V.K., Bhardwaj, B., Gupta, V., Kumar, R. and Mathur, A., 2020, December. Gated recurrent unit (gru) based short term forecasting for wind energy estimation. In *2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)* (pp. 1-6). IEEE.

21. Miraftabzadeh, S.M., Longo, M. and Brenna, M., 2023. Knowledge Extraction from PV Power Generation with Deep Learning Autoencoder and Clustering-Based Algorithms. *IEEE Access*.