National College of Ireland

Solar Sight Forecast: Deep Learning Approaches for Solar PV Power Prediction at Bui Solar Power Station Ghana

MSc Research Project
Data Analytics

Mitali Sopte
Student ID: X22198121

School of Computing
National College of Ireland

Supervisor: Jaswinder Singh

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Mitali Sopte
Student ID:	X22198121
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Jaswinder Singh
Submission Due Date:	12/08/2024
Project Title:	Solar Sight Forecast: Deep Learning Approaches for Solar PV Power Prediction at Bui Solar Power Station Ghana
Word Count:	6948
Page Count:	20

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:	Mitali Sopte
Date:	16th September 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):	

Solar Sight Forecast: Deep Learning Approaches for Solar PV Power Prediction at Bui Solar Power Station Ghana

Mitali Sopte
X22198121

Abstract

The major crisis faced by the BUI Power Authority was to cope with the consistent distribution and generation of solar energy, which is influenced by various climatic conditions like humidity, wind, ambient temperature, global irradiation, etc. This research aims to enhance the understanding of solar power generation and enable reliable energy distribution to the organisation. The previous research that deployed machine learning models like gradient boosting and random forest achieved an accuracy of 90% and a normalised mean absolute error of 1.18%. Based on these findings, the approach to understanding how a deep learning model like LSTM can be used to increase the accuracy and overall outcome was carried out in this study. These findings can result in the potential of deep learning techniques, which can help in assisting the BUI Power Authority in utilising energy appropriately. The implication of this research is to enhance the reliability of solar energy supply, resulting in the broader goal of sustainability of renewable resources.

1 Introduction

1.1 Background

Renewable energies are good sources of clean energy that don't emit greenhouse gases or other harmful gases that can affect the environment. The growth of renewable energies is growing rapidly, with the share of renewable in the global electricity supply expected to increase from 28.7% in 2021 to 43% in 2030 International Trade Administration (2022). Solar energy is one of the renewable energy resources that is crucial in achieving a sustainable environment. As the world faces issues and challenges related to climate change, solar energy comes up as a suitable solution RFI (2023). One of the essential measures for sustainable environments is solar power generation. Solar energy lessens our carbon footprint by generating electricity from the sun, which is widely available and is a free resource. The electricity which is generated is without any pollutants or greenhouse gases. Solar power systems are installed very easily, and can be available in various sizes which are compatible for small houses as well as large solar farms. The use and generation of solar power have become more and more accessible, providing a practical approach to reduce climatic change and make the world more sustainable. The production of solar power has become essential to the world's shift to sustainable energy sources. Precise estimation of solar radiation, specifically global horizontal irradiance

(GHI), is essential for efficient solar power plant planning, design, and site selection. But solar irradiance is a naturally erratic resource that changes dramatically over time due to a wide range of meteorological conditions, including temperature, humidity, and cloud cover Ela et al. (2013). Utilising semiconducting materials, photovoltaic (PV) technology is a popular renewable energy source that transforms solar radiation into electrical energy. Several variables, including solar radiation, ambient temperature, weather, and geographic location, affect the use and efficiency of photovoltaic panels. To prevent energy from renewable sources from being wasted and to guarantee that consumers have access to enough power, electrical operators must balance electrical generation and consumption. Accurate power generation prediction from PV plants is essential for future development and grid stability, as the integration of solar energy power plants with the electrical grid presents stability challenges Kim et al. (2019). This research is mainly done with the purpose of finding ways to efficiently supply power and find answers to the below questions.

1.2 Research question and Objective

- Research Objective 1: Evaluate how the performance of the LSTM model results in predicting hourly solar PV power generation at the Bui Solar Power Station in Ghana, using evaluation metrics such as RMSE and R2.

- Research Objective 2: Analyse the factors like wind speed, global irradiation, humidity, and ambient temperature that affect the performance of the LSTM model in predicting solar power generation.

Deep learning techniques like LSTM and RNN play a significant role in predicting solar power generation. These models are used since the nature of the data is a time series. These models make it convenient to forecast solar power, which is highly distorted by other environmental factors. The objective will help to find how these deep learning models can provide accurate prediction, facilitate energy management, and improve planning for solar power. The second objective will serve the purpose of comparing the traditional machine learning models like random forest and gradient boosting, which were used in the previous study by BUI Authority, with the deep learning models like LSTM. Recent studies have shown that the LSTM models outperform traditional forecasting methods like regression techniques and ARIMA. Thus, this will provide a detailed understanding of the same. Since the generation of solar power is a clean process, other factors like humidity, wind, etc. have a significant effect on solar power generation thus finding how the other environmental factors affect solar power generation would create a better understanding for the management company.

1.3 Structure of Document

This current study document is divided into seven parts, each of the sections gives detailed information on the process of predicting solar power. Section 2 is related to the literature survey this section gives us the synopsis of previous work done in related or similar fields. Section 3 consists of the methodology used in this current study, whereas section 4 gives us the design specification. Section 5 tells us about the implementation, and section 6 provides the results and evaluation, which will help to get a better understanding of the research. Finally, the last section 7 consists of a conclusion and future work.

2 Related Work

The solar power generation is crucial in the Ghana region as the country aims to tackle the climatic change issue by transitioning to renewable energy sources. Various number of studies have been carried out to understand the potential challenges that occurs during solar energy deployment. The Ghanaian aims to achieve 10% renewable energy in its power mix by 2030 with a focus on solar energy Quansah et al. (2016) which has lead to significant investments in solar energy sector. But since the generation of solar energy generation is affected by various environmental factors and weather condition it is challenging for energy management. Below mentioned subsection which explore more about the model integration and understanding machine learning and deep learning models.

2.1 Research based on Solar Power prediction using Deep learning techniques

In this section, the literature related to time series forecasting is been discussed. This research mainly focus on deep learning techniques for solar power predictions. Many studies have been done by applying deep learning techniques. The research done by Tuohy et al. (2015) investigated the present state of solar forecasting which undergoes various approaches and difficulties for accurately predicting solar power output. There are three different categories into which the solar forecasting is divided. The physical approach use numerical weather prediction models while the statistical approaches are based on machine learning model and lastly the hybrid models are used for physical and data-driven parameters. This paper highlights the importance of solar forecasting in encouraging the integration of renewable energy source into power systems despite challenges faced. In order to predict solar irradiance ,Wen et al. (2019) proposed a hybrid deep neural network model which combine various convolutional and recurrent layers. This model performs well than other traditional machine learning techniques. Various other deep learning architectures for solar forecasting was done and Voyant et al. (2017) discovered that long short-term memory (LSTM) networks were especially good at identifying temporal dependencies in solar data. Huang et al. (2021) discovered a model with deep learning approach that considered various factors as their input parameters PV solar power prediction. The result of prediction for this model was recorded apparently high when verified using data of China's 150MW plant. The study done by Wan et al. (2022) was based on how various cloud conditions affects solar power generation. It used a conventional LSTM model to predict the solar power generation with factor like cloud coverage and conditions. The study done by Adaramola (2020) analysed how the solar PV grid tied energy system for production of electricity in norther part of Nigeria. Also the analyses further was conducted on the performance of an 80 kw solar PV grid.

2.2 Research based on LSTM models for time foresting

The study done by Gers et al. (2001) explores the application of Long Short Term Memory (LSTM) network for time series forecasting which was based on time window approaches. The result obtained from this analysis stated that LSTM network are well suited for complex time series data such as weather prediction and financial forecasting. When the LSTM model was applied using the to a fix sized time window of past data point to predict the future points it result in better performance. The research done by Elsworth and

Güttel (2020) proposes a symbolic approach for time series forecasting using LSTM model combined with ABBA symbolic data. The researcher have highlighted the limitation of traditional LSTM when worked on raw numeric data. To tackle this issue a new dimension reducing symbolic representation called ABBA was used as the input of the LSTM network This approach can help to work on the limitation with the raw numeric data. Zhang and et al. (2024) introduce a hybrid model into this study which integrated the LSTM model with an attention mechanism in order to enhance the accuracy of the model. Various real time data set were used stating that the LST¡-attention-=LSTM model is much better that the LSTM model. By this attention feature the model can select most important features and efficiently allow to capture the complex temporal dependencies. This model not only resolves the issue of dependency of time series forecasting but also showcase good results.

2.3 Research based on comparison between machine learning vs deep learning models.

The study done by Makridakis et al. (2018) suggested that the Deep learning models often produced more accurate forecasting results compared to the statistical models and traditional machine learning models. It is shown that the deep leaning models can result up 6% improvement in accuracy for longer forecast. But the cost and time required for training deep learning models were comparatively higher with some of the model taken over 14 days while statistical models could achieve similar results within a minute. Janiesch et al. (2021) proposed a detailed structure comparing deep leaning models with machine learning models. They stated that machine leaning model work on the range of data to automate analytical models while deep leaning models are subset of machine learning models and uses Artificial neural networks to understand the complex data patterns. In their study Gupta et al. (2023) conducted an experiment comparing the traditional machine learning model with deep learning models for different applications. They derived that out of 8 studies conducted 7 studies showed that the deep learning model outperformed the machine learning model. When there is the high dimensional data such has satellite image or speech recognition or time series data deep learning models work better. The main finding of the research stated that deep learning models handle complex data very efficiently than compared to machine learning model

3 Methodology

This Section gives us in-dept understanding of the methodology and the dataset. Data mining is the process where the large chunk of unorganised data is analysed and organised in a certain way and pattern to retrieve expected and useful information. In this research an enhanced version of CRISP-DM methodology is been used for collecting all the necessary details for data mining process. The Cross- industry Standard Process for data mining in Figure 1 shows the old as well as the updated version which will be used for this research.

The methodology used is changed according to the requirement of this research. The first step includes Business understanding so according to the research the extraction of valid data, segregating the data and then cleaning the data according to our requirement and purpose of the research. The dataset which is been used is of a year which is

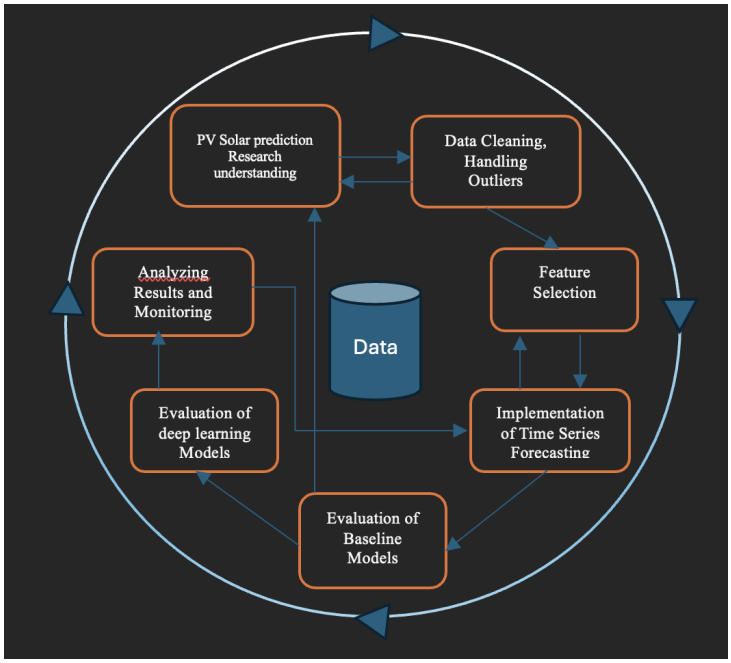


Figure 1: CRISP DM

determined as the time series data so we need to time series data aggregation and one the data is ready we will involve the data in to the implementation of time series model. The evaluation is needed to be after the implementation is carried out and lastly we need to verify the result and understand how the outcome is in line with the expected results. The dataset which is used for this research is obtained from one of the researcher on BUI solar power station. The BUI is the power plant which is responsible for distributing the electricity to the overall region. Due to the climatic changes the authorities were not able to understand the power generation and thus lack is efficient power supply. Hence to understand the power generation a study was carried out with machine learning models to determine according to the months in a time series format. In this research the data is analysed for missing data for cleaning purposes. Since the deep learning models will be used and the nature of dataset should be in time series for evaluation we check if the data set is aligned accordingly. Once the data is checked for missing and Blank values we can plot the data to check outliers using box plot, QQ plot. For this research major time series models like LSTM and RNN will be used thus we can apply the dataset to these models.

3.1 Selection of Data

Selection of appropriate dataset and having a clear understanding of the domain which we are working are important steps before evaluating any results or showcasing any solutions to real-world problems. In solar power sector it is crucial to identify the challenges such has environmental changes and impact of factors like humidity, wind, global irradiation on solar power generation. This study is mainly focused on the solar power generation at BUI Authority, Ghana where the information related to humidity, global irradiation, ambient temperature, humidity is collected

3.2 Data Source and Exploration

The data for this research is taken directly from the researcher who had worked for the same case study using machine learning models. The dataset consist of detailed parameters that are used to measure solar power generation at BUI solar power station in Ghana. It consist of 8 features such as diffuse, global irradiation, humidity along with data and time. The dataset have total 32736 records with duration of 15 minutes. The dataset has data for 11 months starting from April 25th 2021 to March 31st 2022. The data and time features are considered as an index columns to forecast the generation of solar power energy. The recorded granularity for solar power dataset is hourly. The dataset is initially processed by importing the excel file in juypter notebook. The Panda library is used for importing the dataset into a data frame format. Then by using head() command few values along with the data is printed.

3.3 Data Preprocessing and preparation

Once the data is imported to the juypter notebook then the preprocessing method of checking for missing values and cleaning the data is carried out. We have considered time as the main factor since the data is been recorded in the time series format. The data process carried starting from the exploration of data till the data is ready for use. The data was also checked for the variable types since we need integer or float variables to work for correlation matrix and for applying to the models. The date field was in object which was then converted into a datetime format for the same purpose. Once the variable type of the features were checked further process of cleaning was carry forwarded. The below Figure 2 shows the end to end process starting importing excel to the end visualisation.

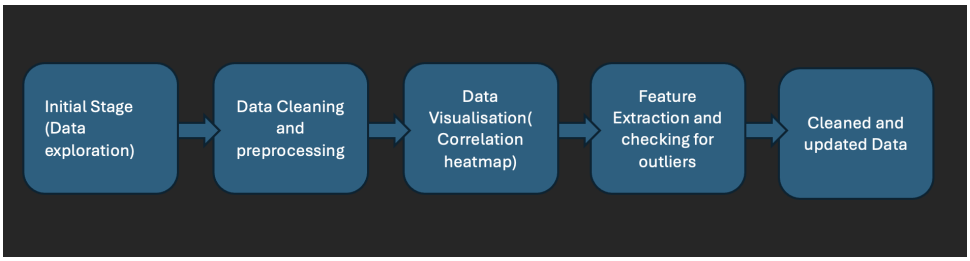


Figure 2: Preprocessing Flow of Data

3.4 Data Cleaning

The data which was obtained for this study was initially divided into each month starting from 25th of March 2021 till 31st April 2022. Each month has a time frame from 0:0 till 24:00 which was divided into slots of 15 mins. So for each month all the parameters were recorded from morning till night. This separate data was first combined into one excel sheet and imported to the jyupter notebook. Once the excel sheet was imported it was converted into the data frame which will be used for further process. The data was then checked for the missing values. From the current data after applying for missing value formula was derived that there were few columns which had missing values. As shown in Figure 3 total 385 missing values were observed which was not deleted because it can

cause abnormality in data. So a technique called interpolation was used to fill the missing values and create a timely based data. The interpolation technique is used for handling missing values for time series data where observation are recorded at regular intervals. By interpolation technique we can create function that predicts the missing values based on the existing data points surrounded by the time.

```
Missing values per column:
Diffuse                385
Wind direction         385
POWER                  385
PANEL TEMP.            385
GLOBAL IRRADIATION     385
HUMIDITY               385
AMBIENT TEMP.          385
DIRECT IRRADIANCE      385
WIND SPEED             385
Datetime                 0
dtype: int64
```

Figure 3: Missing values

3.5 Handling Missing values

Since handling the missing values can be handled in various ways starting from dropping the rows which have missing values or just inserting the average values to the empty cells. There is also one technique used to manipulate the data call interpolation. There are various type of interpolation such as Linear, polynomial, spline etc. In this study time based interpolation is used. This interpolation technique as shown in Figure 4. is ideally used for predicting missing values in time series data. The missing data in the time series forecasting can be due to various reason such as faulty sensor or mismatch in data collection. So the data is been manipulated by creating the function of the present time series data. The new data is dependent on the past data and thus the empty cell are replaced by the new manipulated data As shown in the below diagram the 'method' = 'time' shows that the interpolation is done based on the time index of Solar data frame. The interpolation will take into consideration all the time intervals while estimating missing values. The 'limit_direction' = 'forward' states the direction in which the interpolation will occur. Here the method will fill the values in the forward direction.

```
10 #  Perform time-based interpolation
11 try:
12     solarData.interpolate(method='time', limit_direction='forward', inplace=True)
13 except ValueError as e:
14     print(f"Interpolation error: {e}")
15
```

Figure 4: Timed based Interpolation

3.6 Exploratory Data Analysis

The data analysis is further done by finding out the highly correlated features. In this current study POWER is the main feature which will be monitored for the analysis. So the other feature which are correlated w.r.t to power will be taken into consideration. As shown in the below Figure 5 the scale starts from negative which is denoted by blue colour and goes till positive 1 which is denoted by red colour. The coefficients which are 0 are not correlated while coefficients below 0 shows that they are negative correlated. The

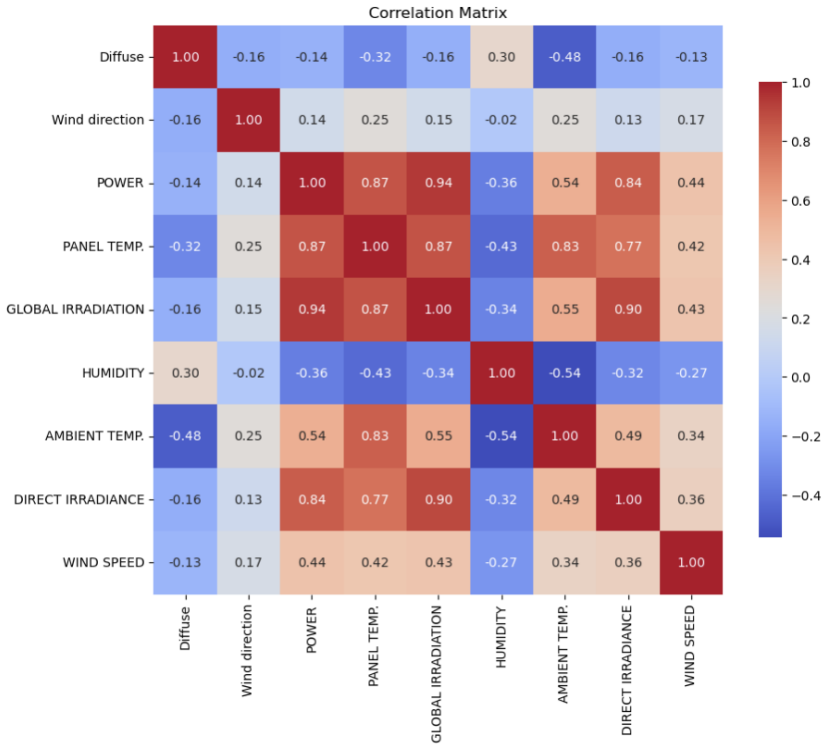


Figure 5: Correlation heatmap

GLOBAL IRRADIATION (0.94) exhibit strong positive correlation with POWER followed by PANEL TEMP. (0.86) and then DIRECT IRRADIATION (0.84). On contrary WIND SPEED (0.44) has moderate positive correlation with POWER. The generation of power shows negative correlation with HUMIDITY (-0.358) and AMBIENT TEMP. (-0.539) stating that as the humidity and ambient temperatures increases the power generation will be negatively impacted. When the parameters have high correlation there is a possibility of having multicollinearity where two or more independent variable are highly correlated. The below Table 1 shows that "GLOBAL IRRADIATION," "DIRECT IRRADIANCE," and "PANEL TEMP." are closely related with each other which may result in having complication in the results. Thus we can address this issue of multicollinearity by using Recursive feature elimination method. This method ranks the feature and keep the feature which are important and discard the function which are least important.

The Figure 6 shows the scatterplot of all the features with respect to POWER. The Diffuse vs. POWER plot shows that higher diffuse radiation responds to lower power speed resulting in negative correlation whereas Wind Direction vs. POWER do not show clear pattern or strong correlation. The HUMIDITY vs. POWER shows that the

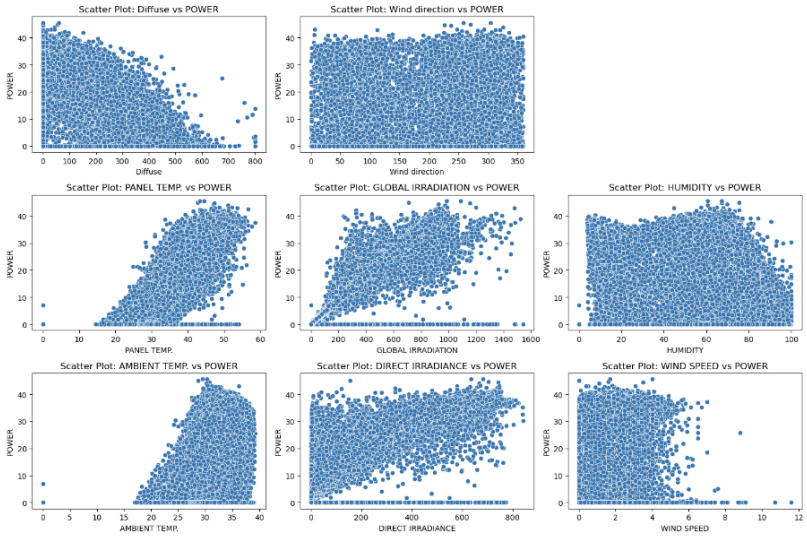


Figure 6: Scatter plot

Table 1: Multicollinearity

Independent Features	**Correlation**
GLOBAL IRRADIATION with DIRECT IRRADIATION	0.90
GLOBAL IRRADIATION with PANEL TEMP	0.87
PANEL TEMP. with DIRECT IRRADIATION	0.77
AMBIENT TEMP. with PANEL TEMP.	0.83

POWER is high at mid humidity level and Wind Speed vs POWER plot shows weak correlation though there is small clustering at low wind speed level.

3.7 Handling Outliers

Outlier are consider to be those data values which are different from the other set of data. These data variation can be caused due to various abnormality or errors and can result into misleading conclusion or skewed results . Thus handling outliers is the crucial step and can be handled using various methods. The most commonly used method is IQR which is called the Interquartile Range method. This method involves taking the Q1 quartile and Q3 quartile of the data respectively. Once the outlier are removed a new clean dataset is created which can be used to process further for deriving prominent and accurate results.

3.8 Feature Selection

For feature selection Recursive Feature Elimination (RFE) method is used. This method is useful when dealing with high dimensional dataset. RFE is usually implemented to enhance the predictive accuracy of the deep learning models used for solar power generation. This process starts by fitting a model to our dataset and then ranking the feature according to the importance score. In this study Random Forest Regression model is used to evaluate the significance of each feature. After applying RFE features namely

"GLOBAL IRRADIATION", "DIRECT IRRADIANCE" , "PANEL TEMP." , "AMBIENT TEMP." were selected and the least scored features were removed.

4 Design Specification

Once the data cleaning ,handling outliers and feature extraction methods are carried out the data is ready to be used for different deep learning model implementation.The LSTM model was chosen for this research due to its ability to handle time series data which is sequential and influenced by past observations. Since solar power generation is affected by the environmental factors LSTMs are specifically designed to capture long term dependencies and temporal pattern. For modelling Pandas library is used along with TensorFlow- keras and stats model package.

4.1 LSTM

To overcome the traditional gradient issue of conventional RNN networks a new architecture was established known as Long Short-Term Memory (LSTM) networks. Due to its capability to capture temporal connection in sequential data set it has become the convenient tool for forecasting time series data such as Solar power generation. LSTM networks are able to reproduce interaction with various other complex parameters like temperature, humidity and other environmental factors that affect the PV power generation. The below Figure 7 shown an LSTM flow which consist of input , forget and output gate. The cell acts as the memory that can be used to store the information over random time interval while the gate keeps the track of the input and output.

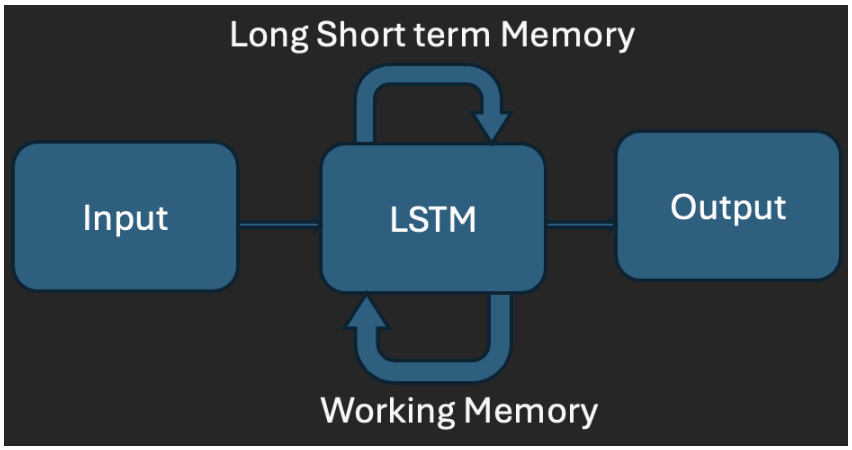


Figure 7: LSTM Model

4.2 LSTM- CNN

CNN- LSTM model as shown in Figure 8 is the combination of the convolutional neural networks with LSTM networks, which makes It convenient for both spatial and temporal datasets. The CNN network in this model is responsible for extracting the features from the input data which in our case is time series data, while the LSTM model processes all the extracted features which are captured over the period of time. This dual model

techniques helps the model to learn complex time series data or other complex image data which are often present in sequential data. In solar power generation forecasting this model can be used to analyse the historical data patterns. The CNN layer will be used to get the input data such as global irradiation, humidity, wind speed and the LSTM layer will understand how these parameter are evolved over the period of time.

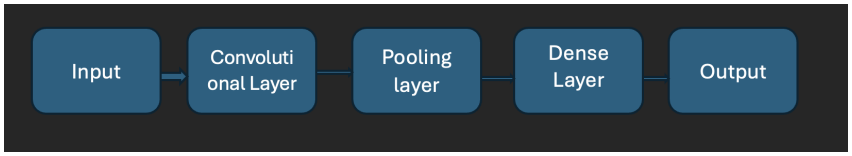


Figure 8: CNN LSTM Model

4.3 Stacked LSTM

The stacked LSTM consist of multiple hidden layer unlike normal LSTM which only has one hidden layer. The multiple hidden layer which are present in stacked LSTM contains 'n' number of memory cells. The base layer operated on basic pattern but the other layer operated on the output of the previous layer. Thus the initial layers are not complex but the preceding layer are combined to form more complex features. Since stacked LSTM can manage the inherent fluctuation and unpredictable nature of solar power generation, they are useful in this project. These models as shown in Figure 9 can identify the unpredicted patterns of the power generation and produce accurate prediction using previous solar data. This can help in improving energy management and improves the accuracy but also offers a strong model for solving issues related to solar power fluctuation.

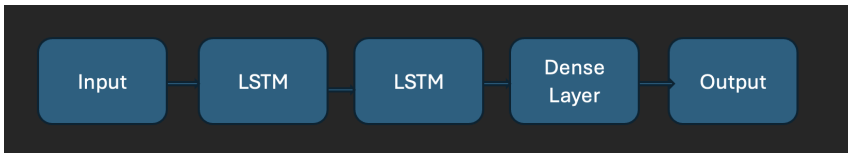


Figure 9: Stacked LSTM Model

4.4 Evaluation Technique

In this study the following Evaluation techniques are been used

- Root Mean Squared Error (RMSE): The RMSE is been calculated for the validation and training set after each epoch during the training process. It calculated the average difference between the predicted and the actual observed values in the model. A lower RMSE shows that the model is better fit to the data
- R- squared Score: The R2 is calculated after the prediction of the test set is done. R2 measures the proportion of variance in the solar power with respect to the independent variables like global irradiation and panel temp. The R2 value should ideally range from 0 to 1 which stated that the model is better fit.

5 Implementation

The implementation of the all three LSTM models was divided into 2 parts, the first part involved clean data including all the features while the other part involved only the important feature which were selected by the Recursive Feature Elimination method. This was done to understand weather all the features are important for the analysis or selected features give more better understanding of the model. As shown in below sections first the LSTM model was implemented with all the features and then the 3 features namely 'GLOBAL IRRADIATION', 'PANEL TEMP.', 'DIRECT IRRADIANCE' .

5.1 Linear Regression benchmark model

This model serve as the benchmark before implementing deep learning models as it serves the purpose to compare the performance of more complex deep leaning models. We can keep the linear model as the base model to understand whether other models are providing meaningful improvement in the analysis. The linear regression model fits a linear relationship between the target values and other independent values to capture the basic trends in the data.

5.2 LSTM implementation

The data which is imported in the python environment is then preprocessed for further model implementation. The outliers are removed and the missing values are handled using interpolation technique. After the preprocessing and exploratory data analysis the data is then used for the implementation of the models. The below Table 2 is of the LSTM model where the input layer has 50 units which means that the model will study various 50 representation of the input data. It will be recorded at each step. The second layer is hidden layer which is also 50 units ,in this layer the return_sequence is false so it generated 1D array for each input instead of sequence The dense layer is considered to be the output layer with 1 unit which is predicting solar power factor. The batch size during training is 32 which helps with the memory efficiency for fast training time. The lstm model is trained with 50 epochs meaning it will go executed the dataset for 50 time. The dropout rate is 0.2 which means 20 percent of the neurons will be zero. This is done to avoid overfitting of the dataset. The Mean squared error is used for the model evaluation and Adam optimizer is used for training the model. The below Figure 10 shows that there are total of 11800 parameters in the first LSTM layer which is followed by the dropout layer with parameter. The second LSTM layer has 20200 parameters followed by second dropout layer. There is no additional parameters in the second dropout layer and the dense layer have 52 parameters. In total there are 32051 parameters which are trainable.

5.3 CNN- LSTM implementation

In this implementation along with the traditional LSTM model steps new additional steps and layers are been added. As shown in below Table 3 the convolutional layer is added which extracts features from the layers which are useful for model evaluation. There is one convolutional layers with size of 64 filter which helps for better feature extraction and better performance of the model. The Rectified Linear Unit (RELU) activation is to include non linearity of the model and understand complicated pattern. To downsize

Table 2: Configuration of LSTM Model

Parameter	**Value**
Input Layer	50 (units in LSTM layer)
Hidden Layer	50 (units in LSTM layer)
Dense Layer	1 (output units)
Activation	None (default for Dense)
Batch Size	32
Epochs	50
Dropout Rate	0.2
Loss Function	Mean Squared Error
Optimizer	Adam

the feature pattern and lower the complexity of calculating Max pooling layers are used. Figure 11 shows the architecture of CNN-LSTM model.

Table 3: Configuration of CNN-LSTM Model

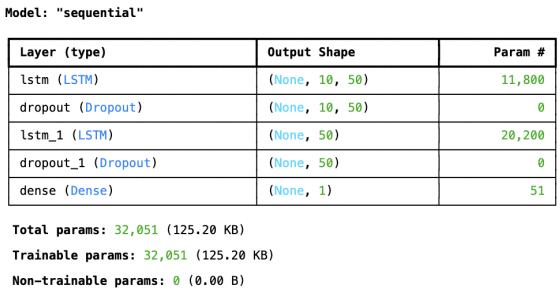
Parameter	**Value**
Input Layer	(X_train.shape, X_train.shape)
Convolutional Layer 1	64 filters, kernel size 2, ReLU activation
Max Pooling 1	Pool size 2
Dense Layer	50 units, ReLU activation
Output Layer	1 unit (no activation)
Batch Size	32
Epochs	50
Loss Function	Mean Squared Error
Optimizer	Adam

5.4 Stacked LSTM implementation

The implementation of stacked lstm is executed in the similar manner as the other two above mention models. In this model the initial as shown in below Table 4 LSTM layer is stacked upon the other LSTM layer. The two lstm layers as shown in Figure 12 have returnsequence and true but the last layer has the the returnsequence and false which indicated that the output will be single vector for each input sequence. LSTM layers give multidimensional output when the returnsequence is true so the two layers will give multidimensional output but the last layer will give 2D output since the returnsequence is set to false. The dense layer as shown in at the end outputs a 1D tensor which is appropriate for working on regression tasks.

Table 4: Configuration of Stacked LSTM Model

Parameter	Value
Input Layer	(X_train.shape, X_train.shape)
LSTM Layer 1	50 units, return_sequences=True
Dropout Layer 1	0.2
LSTM Layer 2	50 units, return_sequences=True
Dropout Layer 2	0.2
LSTM Layer 3	50 units, return_sequences=True
Dropout Layer 3	0.2
Dense Layer	1 (output layer, no activation)
Batch Size	32
Epochs	50
Loss Function	Mean Squared Error
Optimizer	Adam

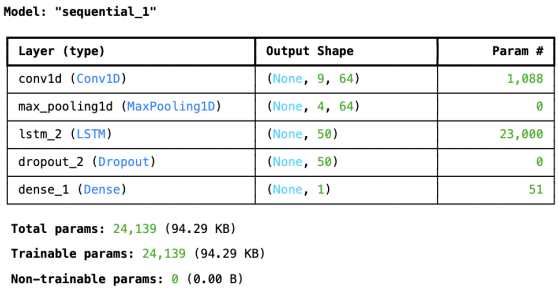


Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 10, 50)	11,800
dropout (Dropout)	(None, 10, 50)	0
lstm_1 (LSTM)	(None, 50)	20,200
dropout_1 (Dropout)	(None, 50)	0
dense (Dense)	(None, 1)	51

Total params: 32,051 (125.20 KB)
Trainable params: 32,051 (125.20 KB)
Non-trainable params: 0 (0.00 B)

Figure 10: LSTM Architecture

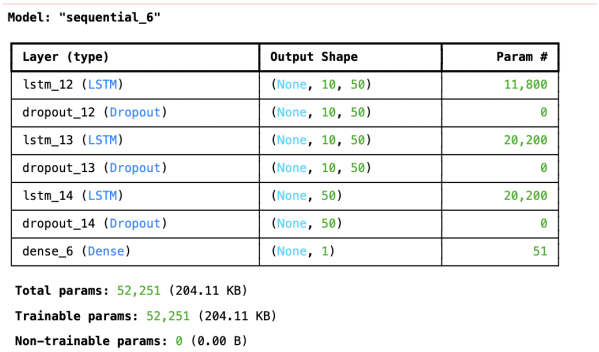


Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 9, 64)	1,088
max_pooling1d (MaxPooling1D)	(None, 4, 64)	0
lstm_2 (LSTM)	(None, 50)	23,000
dropout_2 (Dropout)	(None, 50)	0
dense_1 (Dense)	(None, 1)	51

Total params: 24,139 (94.29 KB)
Trainable params: 24,139 (94.29 KB)
Non-trainable params: 0 (0.00 B)

Figure 11: CNN-LSTM Architecture



Model: "sequential_6"

Layer (type)	Output Shape	Param #
lstm_12 (LSTM)	(None, 10, 50)	11,800
dropout_12 (Dropout)	(None, 10, 50)	0
lstm_13 (LSTM)	(None, 10, 50)	20,200
dropout_13 (Dropout)	(None, 10, 50)	0
lstm_14 (LSTM)	(None, 50)	20,200
dropout_14 (Dropout)	(None, 50)	0
dense_6 (Dense)	(None, 1)	51

Total params: 52,251 (204.11 KB)
Trainable params: 52,251 (204.11 KB)
Non-trainable params: 0 (0.00 B)

Figure 12: Stacked LSTM Architecture

6 Evaluation

6.1 Linear Regression baseline Model

The linear regression baseline model provided the result which achieved RMSE of 4.33 and R2 of 0.87 stating that the performance of the benchmark model can keep be kept as the reference for working on other deep learning models. The Residual and scatterplot are as shown in Figure 13 and Figure 14

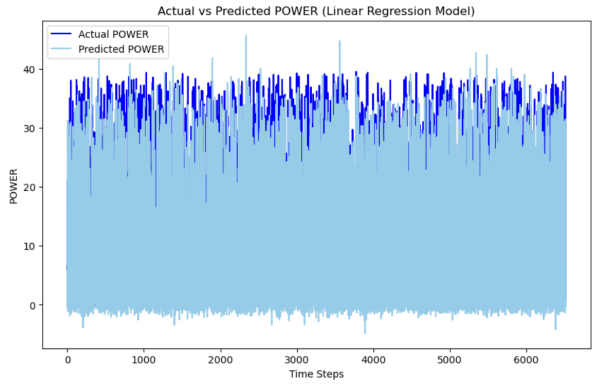


Figure 13: Residual Graph of Linear Model

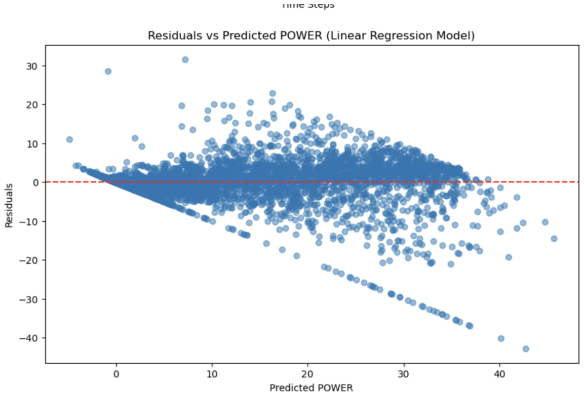


Figure 14: Scatter plot of Linear Model

6.2 LSTM Model

The LSTM model was trained on 50 epoch and achieved the RMSE of 3.51 and R2 of 0.91 as shown in Table 5. The below Figure 15 shows the Predicted vs Actual graph for LSTM model with all the independent features shows that the predicted values closely follow the actual values, which suggests that the model is performing well. The alignment between the actual and predicted lined indicate the model is able to capture the underlying trend in the data. The scatter plot in Figure 16 shows a funnel like shape where the spread of the residual increases as the power increases. This indicated heteroscedasticity, that means the model error variance is not constant across all the level of the predicted values. This model tends to have higher error for higher predicted values.

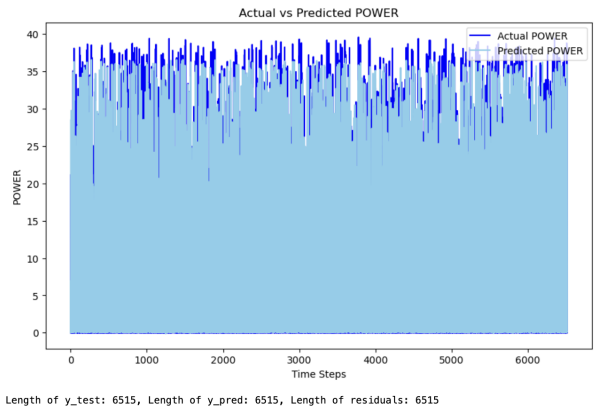


Figure 15: Residual Graph of LSTM Model

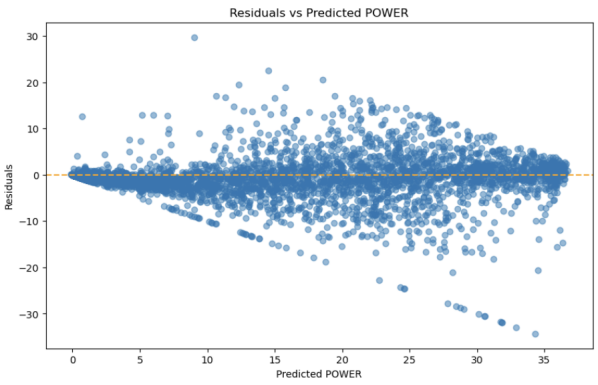


Figure 16: Scatter plot of LSTM Model

6.3 CNN-LSTM Model

The model CNN-LSTM achieved the RMSE of 3.75 and R2 of 0.90 which was less than the LSTM model which indicated that since LSTM-CNN combined the feature extraction capabilities of convolutional neural network with the temporal processing the strength of LSTM while this hybrid approach can capture complex patterns in the data that can also introduce additional complexity that could lead to over fitting of the data. The

residual graph and scatter plot of CNN-LSTM model is shown in Figure 17 and Figure 18 respectively.

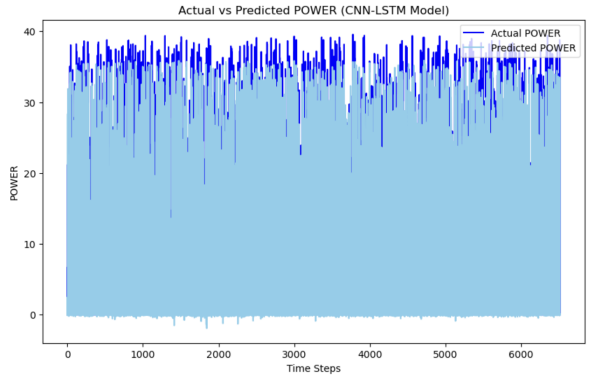


Figure 17: Residual Graph of CNN-LSTM Model

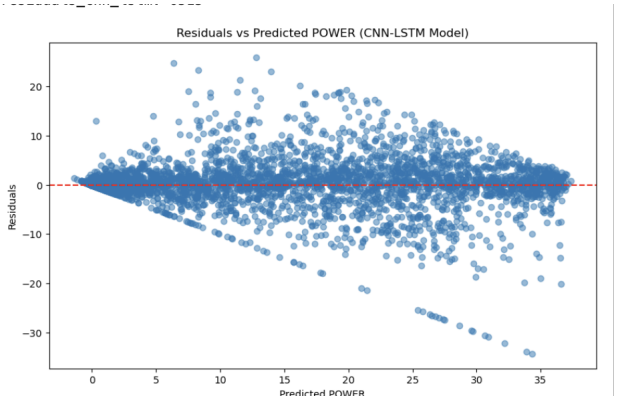


Figure 18: Scatter plot of CNN-LSTM Model

6.4 Stacked LSTM Model

The stacked LSTM is also similar to LSTM and CNN- LSTM which was trained with 50 epoch and the loss encountered was 10.58 which was less than the LSTM and CNN-LSTM. Figure 19 and Figure 20 shows the residual and scatterplot of Stacked LSTM model.

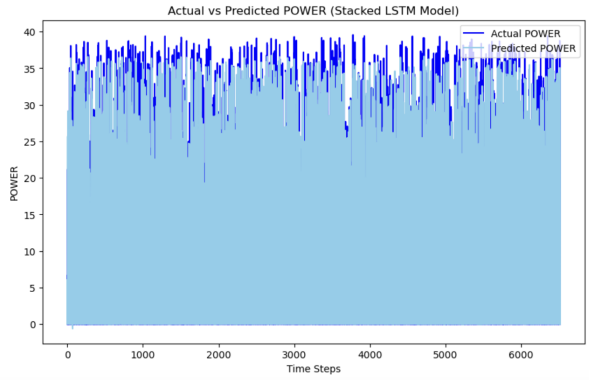


Figure 19: Residual Graph of Stacked LSTM Model

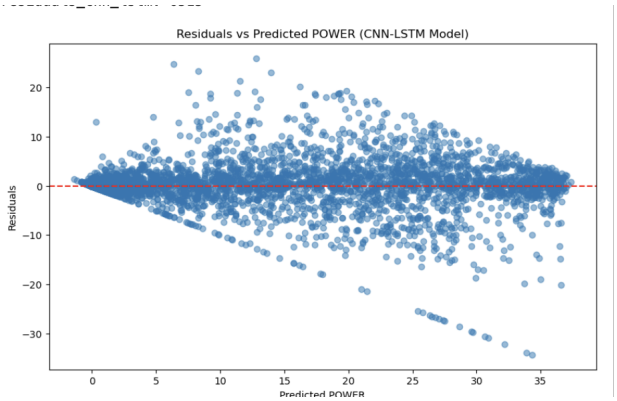


Figure 20: Scatter plot of Stacked LSTM Model

Table 5: Evaluation Results for Complete Data

Model	**All Features**		**3 Features**	
	RMSE	**R^2**	**RMSE**	**R^2**
LSTM	3.51	0.91	3.71	0.90
CNN - LSTM	3.75	0.90	3.99	0.89
Stacked LSTM	3.54	0.91	3.70	0.90

The Table 5 shows that LSTM model got RMSE of 3.71 and R2 of 0.9 follow by stacked LSTM with similar RMSE and R2 and lastly CNN-LSTM with RMSE of 3.99

and R2 of 0.89. Comparing it with all features LSTM have comparative low RMSE (3.51) followed by stacked LSTM (3.54) and then CNN- LSTM(3.75) . The R2 obtained for LSTM and Stacked LSTM was 0.91 which is higher than others. From the above table it is observed that the model LSTM works well when all the features are included rather than only 3 stating that all the parameters give information which is important for predicting power generation. While feature selection from various methods can be important for potentially improving the model accuracy and reducing overfitting but it can also lead to a loss of predictive accuracy which is provided by other features. To resolve the issue of the slant line that occurred in the scatter plot a new dataset with all the data point with 0 in power column were removed. The data was then applied to LSTM model with all feature. As shown in below Figure 21 when the 0 values are removed from the dataset the slanting line is not visible but the RSME and R2 values are low than that of the values when all the dataset is used. The below table shows the RMSE and R2 values when 0 data values are removed as shown in .

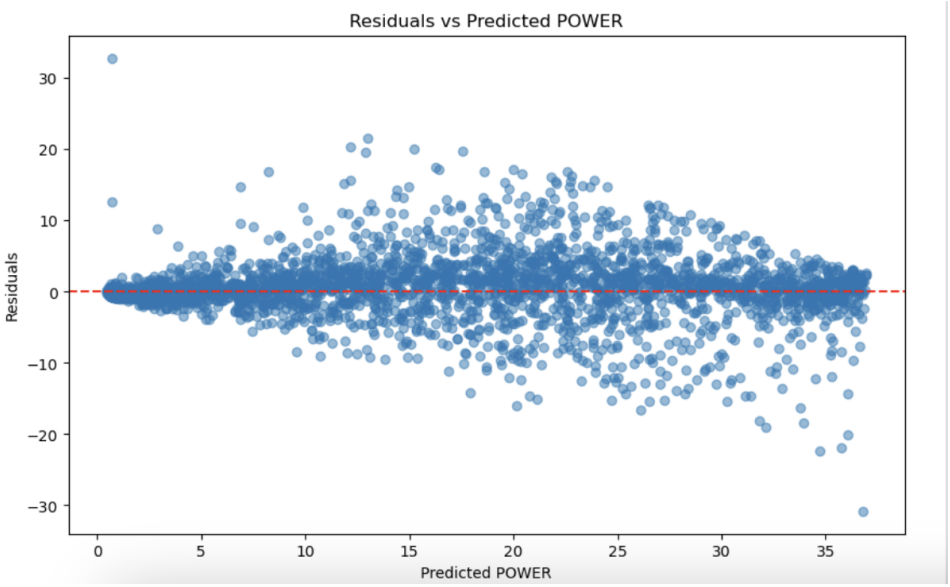


Figure 21: Scatter plot when data with 0 values are removed

Table 6: Evaluation Results when data values with 0 are Removed

Model	**All Features**		**3 Features**	
	RMSE	**R^2**	**RMSE**	**R^2**
LSTM	4.39	0.87	4.47	0.86
CNN - LSTM	5.24	0.81	5.94	0.75
Stacked LSTM	4.45	0.86	4.52	0.86

The LSTM model has performed best with 4.39 RMSE value and 0.87 $R\hat{2}$ value as shown in Table 6 but it is low than the RMSE and R2 vales which we obtained when 0 data values where not removed.

6.5 Discussion

In this research the implementation of these deep learning models for predicting solar power generation has resulted in some important and valuable insights. These model namely LSTM, CNN LSTM and stacked LSTM have been evaluated based on its ability to predict power output using historical data. The data is then been evaluated using RMSE and R2 values. The result indicated that the LSTM model gave better performance results with comparatively lower RMSE and higher R2 values across all different model architectures. This suggest that the LSTM model ability to capture all the temporal dependencies in the solar power dataset is well suited for this particular prediction task. The CNN model on the other hand while involving convolutional layers to extract spatial features did not exhibit favourable outcomes over the simple LSTM model. This may be because of the nature of the dataset where temporal patterns are more crucial than the spatial components. The stacked LSTM model which uses 2 LSTM layer considerably showed similar performance as LSTM but did not showed any substantial growth in accuracy. The residual graphs were used to understand the actual vs the predicted values of the dataset. By looking at all the model residual graphs we could identify patterns indicating areas where the models has struggled and where the model has fitted the data correctly. There are many instance for further research and enhancement of the data. Additionally exploring feature engineering factors like lagged variables might result in improving predicted accuracy. Overall, this research showcases the importance of proper model selection and analysis in development of effective predictive tools for renewable energy generation.

7 Conclusion and Future Work

The aim of this research is to understand how the deep learning models like LSTM work with the solar power dataset. This research is conducted on the basis of the previous research which was done with machine learning models. In the previous research the best performing model generated an average mean absolute error of 1.18%. The linear regression model served as the bench mark model for evaluating other deep leaning model like LSTM , LTMCNN and stacked LSTM for solar power generation. With the RMSE of 4.33 and R2 as 0.87 the model exhibited normal baseline structure which captured the relationship of the target variable with the other independent variable. However when the linear model is compared with the LSTM model it shows that the more complex model provide advance predictive capabilities.

The analysis of the model performance results in significant differences in RMSE and predictive accuracy when the complete dataset is used compared to when dataset without 0 data points are used. The result of RMSE and R2 when 0 data point are not removed stated that zero values in the dataset may provide valuable information for analysis. Additionally the Stacked LSTM and LSTM models are outperforming CNN-LSTM model highlights the latter limitation for this set of dataset.

When the dataset with all features is used LSTM model achieved the best performance with RMSE 3.51 and R2 of 0.91 indicating that the model has effectively captured all the underlying patterns of the dataset. On the other hand when the data set excluding 0 data values is been used the LSTM model resulted in RMSE of 4.39 and R2 of 0.87 indicating that the model did not performed as well as the previous one. The dataset without zero was used to tackle the issue of negative slanting line which occurred in scatterplot. But

from the analysis it is observed that 0 values may provide valuable information which is useful for analysis. The LSTM and stacked LSTM model performed well with the dataset compared to CNN- LSTM highlighting its limitation for this particular dataset.The result states that the even though deep leaning model when all features are implemented gives us a betting R2 and a low RMSE suggest that all the features are responsible for power generation. On the other hand if we remove the power values which had 0 data values the R2 achieved was less that 91% suggesting that 0 contains valuable information about the dataset.

The future work can be done by implementing model on a large dataset for period that is more that 11 months since this dataset only captures the observation of solar power generation for a 341 days. This can give a better understanding of the model and evaluating the outcomes. The dataset with higher temporal resolution cab be used for instance minute level data to enable ultra short term forecasting. Also the dataset form various other solar power plant can be taken into consideration for further research purpose.Various similar dataset can be used to analyse these three models to understand whether the models are compatible to all sort of data.

8 Acknowledgment

I would like to sincerely thank Professor Jaswinder Singh for his encouragement, untiring patience and his valuable knowledge throughout the course of my research and writing this thesis. I am extremely thankful for his guidance and expertise at each and every step which I appreciate with my utmost sincerity. The quality of thesis that I was able to write was due to his constant instruction and appropriate specification reports provided at each stage. His valuable feedback for each section was very helpful to write the thesis gaining a proper understanding of this research.

References

Adaramola, M. S. (2020). Viability of grid-connected solar pv energy system in jos, nigeria, *International Journal of Renewable Energy Research (IJRER)* **2**(3): 422–427.

Ela, E., Diakov, V., Ibanez, E. and Heaney, M. (2013). Impacts of variability and uncertainty in solar photovoltaic generation at multiple timescales.

Elsworth, S. and Güttel, S. (2020). Time series forecasting using lstm networks: A symbolic approach, *arXiv preprint* .

Gers, F. A., Schraudolph, N. N. and Schmidhuber, J. (2001). Applying lstm to time series predictable through time-window approaches, *ResearchGate* .

Gupta, V., Peltekian, A., Liao, W. K. and et al. (2023). Improving deep learning model performance under parametric constraints for materials informatics applications, *Scientific Reports* **13**: 9128.
URL: *https://doi.org/10.1038/s41598-023-36336-5*

Huang, J., Xiao, L., Wen, J. and Xie, K. (2021). Deep learning based short-term photovoltaic power forecasting, *IEEE Access* **9**: 39399–39409.

International Trade Administration (2022). Ghana - energy and renewables.
URL: *https://www.trade.gov/country-commercialguides/ghana-energy-and-renewables*

Janiesch, C., Zschech, P. and Heinrich, K. (2021). Machine learning and deep learning, *Electronic Markets* .
URL: *https://link.springer.com/article/10.1007/s12525-021-00475-2*

Kim, G. G., Choi, J. H., Park, S. Y., Bhang, B. G., Nam, W. J., Cha, H. L., Park, N. and Ahn, H. (2019). Prediction model for pv performance with correlation analysis of environmental variables, *IEEE Journal of Photovoltaics* .

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward, *PLoS ONE* **13**.
URL: *https://doi.org/10.1371/journal.pone.0202950*

Quansah, D. A., Adaramola, M. S. and Mensah, L. D. (2016). Solar photovoltaic systems in ghana: Potential and challenges, *International Journal of Solar Energy* pp. 1–8.
URL: *https://doi.org/10.1155/2016/8689097*

RFI (2023). Ghana steps up solar energy in bid to meet renewables target.
URL: *https://www.rfi.fr/en/africa/20230813-ghana-steps-up-solar-energy-in-bid-to-meet-renewables-targets*

Tuohy, A., Zack, J., Haupt, S. E., Sharp, J. and Ahlstrom, M. (2015). Solar forecasting: Methods, challenges, and performance, *IEEE Transactions on Sustainable Energy* **6**(4): 1310–1333.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F. and Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review, *Renewable Energy* **105**: 569–582.

Wan, C., Zhao, J., Song, Y., Xu, Z., Lin, J. and Hu, Z. (2022). Photovoltaic and solar power forecasting for smart grid energy management, *CSEE Journal of Power and Energy Systems* **1**(1): 38–46.

Wen, S., Lan, H., Fu, Q., Yu, D. C. and Nayar, C. V. (2019). Economic allocation for energy storage system considering wind power distribution, *IEEE Transactions on Power Systems* **33**(1): 171–180.

Zhang, Y. and et al. (2024). Time series prediction based on lstm-attention-lstm model, *IEEE Transactions on Neural Networks and Learning Systems* .
URL: *https://ieeexplore.ieee.org/document/10124729*