DEEP LEARNING PREDICTION OF SHORT-TERM HOURLY POWER OUTPUTS OF WIND AND SOLAR PV FARMS FOR REAL-TIME SMART BUILDING HVAC CONTROL

By

JAYASURIYA NANDHAGOPAL (501085084),

A Major Research Project Report

Presented to Toronto Metropolitan University

in partial fulfilment towards the requirements for the degree of

Master of Engineering

In

Mechanical, Industrial and Mechatronics Engineering

2022-2024

Toronto, Ontario, Canada, 2024

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Toronto Metropolitan University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Jayasuriya Nandhagopal

DEEP LEARNING PREDICTION OF SHORT-TERM HOURLY POWER OUTPUTS OF WIND AND SOLAR PV FARMS FOR REAL-TIME SMART BUILDING HVAC CONTROL

Jayasuriya Nandhagopal

Master of Engineering 2024

Mechanical and Industrial Engineering

Toronto Metropolitan University

ABSTRACT

Electrifying space heating of residential and commercial buildings through the use of Air Source Heat Pumps (ASHP) over Natural Gas Furnace (NGF) had been enhanced by the development of Smart Dual Fuel Switching Systems (SDFSS), a cloud based IoT controller in buildings that enables automatic switching between ASHP and NGF based on price optimization criterion to reduce the operation cost of ASHP while reducing the greenhouse gas emission (GHG). In the current work five machine learning algorithms, 1) Multiple Linear Regression, 2) MLPRegressor, 3) Random Forest Regressor, 4) Support Vector Regressor, and 5) LSTM algorithms were used to develop hourly energy output forecasting of wind farms and solar PV farms, in Ontario, Canada. The two hours ahead prediction model trained with LSTM algorithm predicts the solar and wind energy output with 91% and 90% accuracy without overfitting, which could be used in SDFSS system.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude towards **Dr. Alan Fung** for providing me with the opportunity to work and collaborate with his team at the Centre for Sustainable Energy Systems (CSES) at Toronto Metropolitan University. The experience of working on a real-world problem with the aim of tackling a pressing issue like climate change has helped me gain perspective towards the urgent need of such research.

I am grateful to **Dr. Fung** for his constant support and assistance towards the progress and completion of this project. As my supervisor and mentor, Dr. Fung has been monumental throughout the term to guide and direct my research and provide valuable feedback for me to improve on.

I am walking away with an invaluable experience and motivation to use my knowledge and skills towards the betterment of the world we live in.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1. Introduction

De-Carbonizing the society to achieve the Net-Zero emission by 2050 is the primary target of the Paris Agreement drafted by United Nations Framework Convention on Climate Change (UNFCCC). In resonance, the Canadian Federal Government's Pan Canadian Framework dictates the reduction of Canada's greenhouse gas (GHG) emission by 30% by 2030 and by 80% by 2050 [1, 2]. The deep examination of Canadian GHG emissions by sector shows that the residential sector is the third-largest contributor to national GHG emissions, accounting for a total contribution of 17%, of which 80% arises from space heating applications due to cold climatic conditions and large living space buildings, with an average house consuming 63% of its total energy for space heating by using natural gas as the predominant source of fuel for heating [1, 2, 3].

The switching of space heating from more polluting traditional natural gas fired furnace (NGF) to electricity driven air source heat pump (ASHP), reduces approximately 46-54% of annual GHG emission, while increasing the operational cost of space heating [4]. The high operational cost of ASHP could be reduced by using hybrid heating systems. The cloud based smart dual fuel switching system (SDFSS) reduces the operational cost and improves the efficiency of hybrid heating systems by enabling automatic switching between ASHP and NGF depending on multi-variable optimization process using the following variables time-of-use (TOU) pricing, fuel cost, weather forecast, equipment efficiencies and capacities [1, 3, 5]. There is a further opportunity to improve the efficiency of SDFSS system to reduce the operational cost of heating by integrating the heating system with renewable energy output from the local nearby solar PV farms and wind farms forming a local communal energy pooling. The more precise short-term prediction of energy output from local solar PV farms and wind farms is of paramount importance while integrating the renewable energy output to the SDFSS systems. The renewable energy should be predicted up to 6-hour head to be used with the SDFSS system.

The artificial neural network (ANN) models with an average correlation coefficient of 0.99 was developed using the weather data at a 2 min and 1 hour resolution to predict the outdoor temperature of residential building without requiring high precision sensors [6]. ANN model was developed using the Levenberg-Marquardt algorithms to predict the performance of heat pump systems in row-house with retrofits [7].

Two major methods are used in forecasting solar irradiance that could be used in predicting solar PV output the cloud imagery combined with physical models, and the machine learning models. Machine learning models to predict the solar irradiance is not limited to artificial neural networks, support vector regression, regression tree, random forest, gradient boosting [8]. The weather data classified into four types clear sky, cloudy day, foggy day, and rainy day was used to develop support vector machines (SVM) to forecast the output power of solar PV one-day-ahead [9]. The total annual solar irradiance, land slope, land use, land accessibility, air borne dust risk map, air temperature at 2 m, relative humidity at 2 m and wind speed at 10 m were used to determine the land relative suitability index for the implementation of solar PV farms [10]. The radial bias function network (RBFN) was developed using the combination of classified weather points into rainy, cloudy and sunny in the first layer and weather parameters in the second layer to determine the solar PV output 24 hours ahead [11]. Long Short-Term Memory (LSTM) model was developed using the weather parameters to predict the solar PV output 24 hours ahead with 18% more accuracy in terms of RMSE than back-propagation neural network (BPNN) algorithm [12]. SVM based forecasting model was trained to predict the solar PV power output using the locally available weather forecast data [13].

The wind speed sequence was decomposed into sub-sequences using Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), and the Least Squares Support Vector Machine (LSSVM) algorithm, optimized by the Artificial Bee Colony algorithm, was then applied to predict the wind speed [14]. Ensemble regressors and decision tree algorithms like k-nearest neighbors and Support Vector Regression (SVR) were developed to predict the energy output of wind farms, and it is found that the combination of decision tree with SVR outperforms the ensemble predictors [15]. Six different algorithms were developed to forecast the energy output of wind farms, namely XGBoost, Random Forest regressor, LightGBM, CatBoost, AdaBoost, and M5-Prime and among these algorithms, XGBoost produced the best performance in forecasting [16]. In predicting the wind farm's energy output for varying time horizons (1-hour, 1-week, and 12-month), six algorithms—Elastic Net Regression, Random Forest Regression, SARIMA, XGBoost, Prophet, and a combined Prophet and XGBoost model—were employed, with the SARIMAX model excelling in short-term forecasting and XGBoost delivering superior long- term predictability [17].

The present work focuses on the prediction of wind farms power output for varying time horizons (1-hour, 2-hour, 3-hours, 4-hours, 5-hours, 6-hours) using five algorithms: 1) MLPRegressor, 2) Support Vector Regression (SVR), 3) Random Forest Regressor, 4) multiple

linear regressor, and 5) LSTM. The models were developed for three wind farms at Zurich, Erieau, Amaranth and two solar PV farms at Southgate and Windsor Airport in the province of Ontario, Canada. Each model is developed using multiple test-train splits (70/30, 80/20, 60/35, 50/50) to determine the robustness of the model.

2. Machine Learning Architecture

The machine learning problem consists of four stages, as illustrated in Figure 1, with each stage being critical for formulating a precise and robust artificial intelligence (AI) model for forecasting wind farm and solar PV farm power output.

Stage 1: This stage involves gathering hourly power output data from power plants and their corresponding weather data obtained from nearby weather stations. The data is then cleaned to remove inconsistencies, transformed into a standardized format, and stored in the database, as depicted in Figure 1.

Objective Function Selection: Next, the objective function is selected, which entails predicting the power output for 1, 2, 3, 4, 5, and 6 hours ahead, as shown in Figure 1. For each prediction horizon, a series of machine learning (ML) models were developed, which will be discussed later.

Production AI: The ML algorithm with the highest average correlation coefficient (r^2) is chosen as the production AI model as indicated in Figure 1.

Within each objective function, the dataset is divided into multiple test-train splits (70/30, 80/20, 65/35, 50/50), as depicted in Figure 2. Each test-train split undergoes training and testing using five different algorithms: multiple linear regression, MLPRegressor, SVR, random forest regressor, and LSTM. Their performance is evaluated in terms of root mean squared error (RMSE), mean squared error (MSE), and correlation coefficient (r^2).

The mean correlation coefficient (r^2) for each algorithm across different test-train splits is computed. The algorithm with the highest average correlation coefficient (r^2) is selected as the best machine learning model for predicting the power output with a specific time delay, as illustrated in Figure 2.

Similarly, the objective function is altered for forecasts ranging from 1 to 6 hours ahead, and all previous procedures are repeated to identify the optimal algorithm for power output forecasting. Finally, SHAPLEY values are applied to this chosen model to develop an explainable AI.

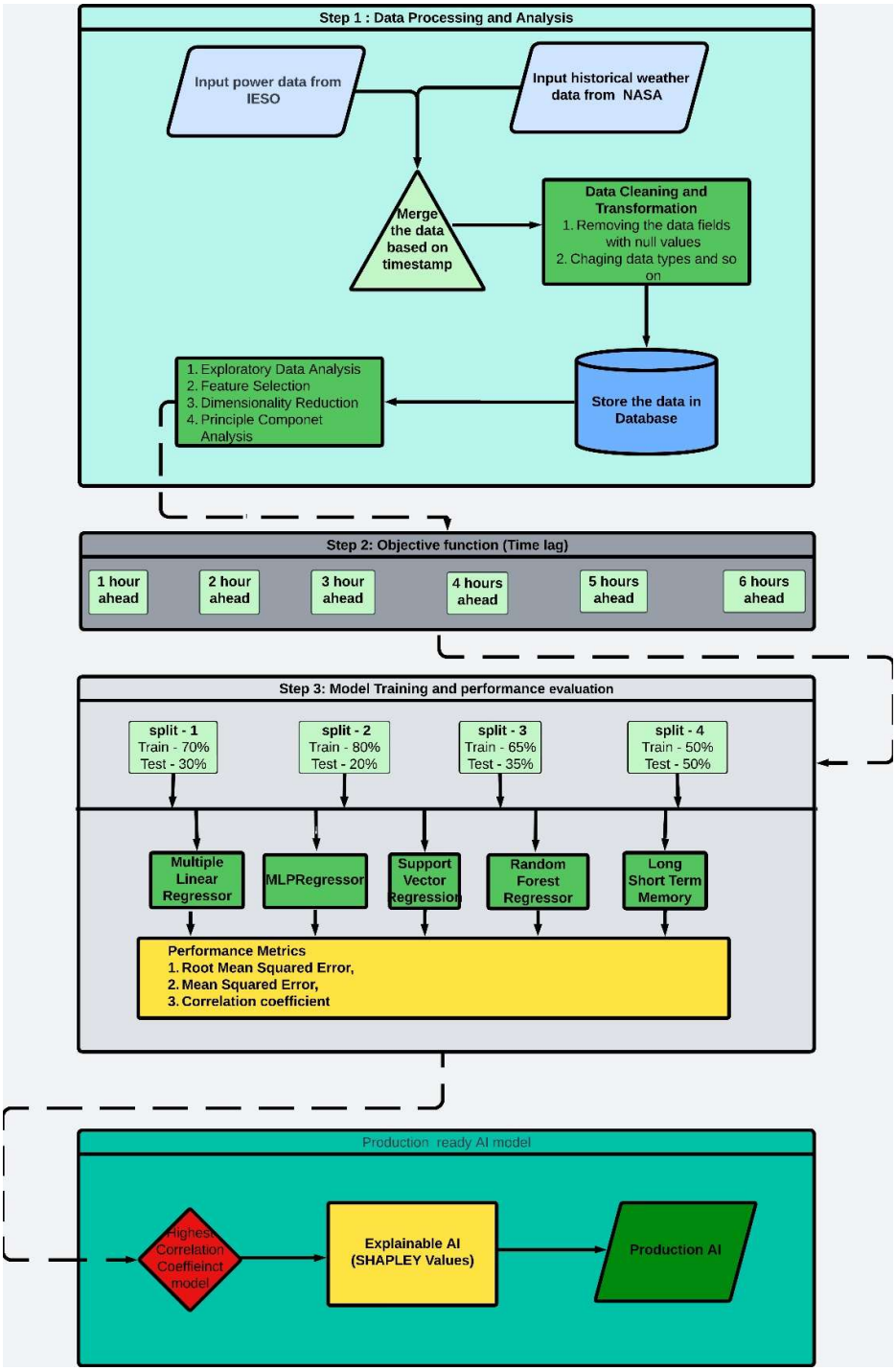

Figure 1. Machine learning architecture to predict wind and solar PV farm power output

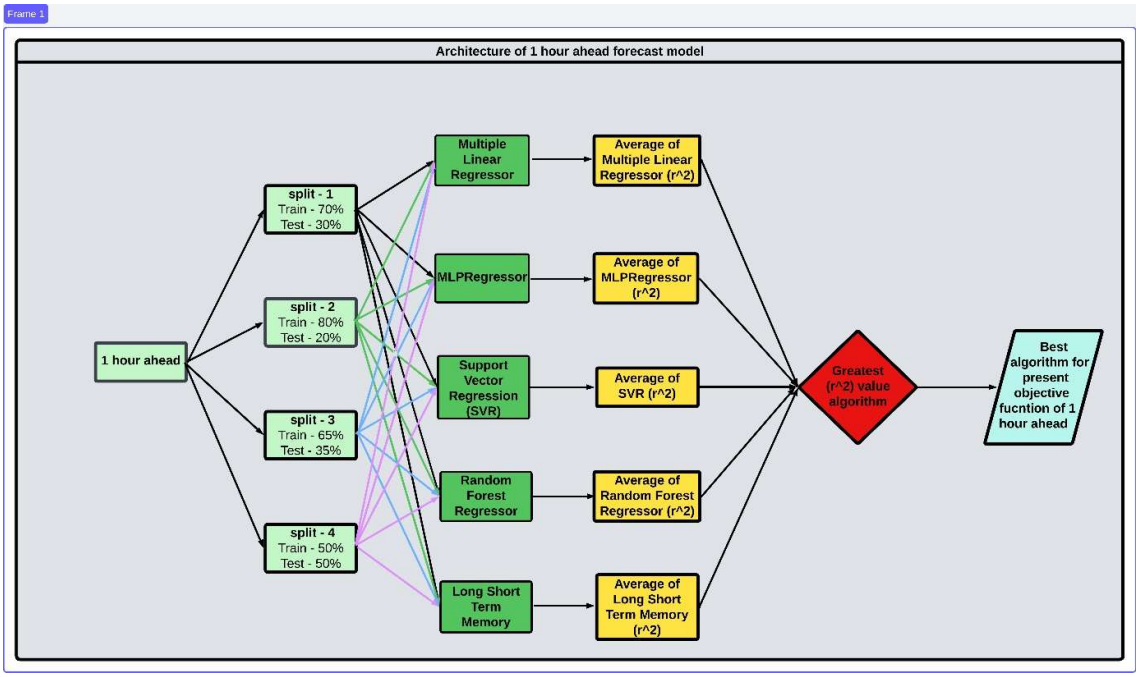

Figure 2. Architecture of single iteration of unique multilevel machine learning mode

2.1 Data

2.1.1 Input data

The prediction of energy output for wind and solar PV farms requires historical power output data from wind farms and solar PV farms, along with corresponding weather data. The historical hourly energy output data from all types of power stations across Ontario, Canada, is publicly available through the Independent Electricity System Operator (IESO) portal [18]. This dataset spans a period of 3 years and 6 months, from June 2019 to January 2023, encompassing all types of power generation, including solar, wind, nuclear, biomass, and other generators, providing hourly power output information for each day. The weather data for the corresponding latitude and longitude of the power plant is obtained from NASA POWER LARC for the same time period [19].

2.1.1 Power plant selection

The details of three wind farms and two solar farms located across Ontario, Canada, are in Table 1 and the location visualized in map is shown in Figure 3.

Table 1. Power plant under study details

Generator name	Type of generator	Capacity	Latitude	Longitude	Operated by
Amarnath, ON	Wind	199.5 MW	44.1001	-80.2709	Transalta [20]
Erieau, ON	Wind	99 MW	42.4	-82.183333	Engie [21]
Zurich, ON	Wind	100 MW	42.30143	-82.16119	Samsung renewable energy and Pattern Canada [22]
Southgate, ON	Solar	50 MW	44.1243696	-80.588280	Samsung renewable energy [22]
Windsor Airport, ON	Solar	50MW	42.28397	-82.93226	Samsung renewable energy [22]

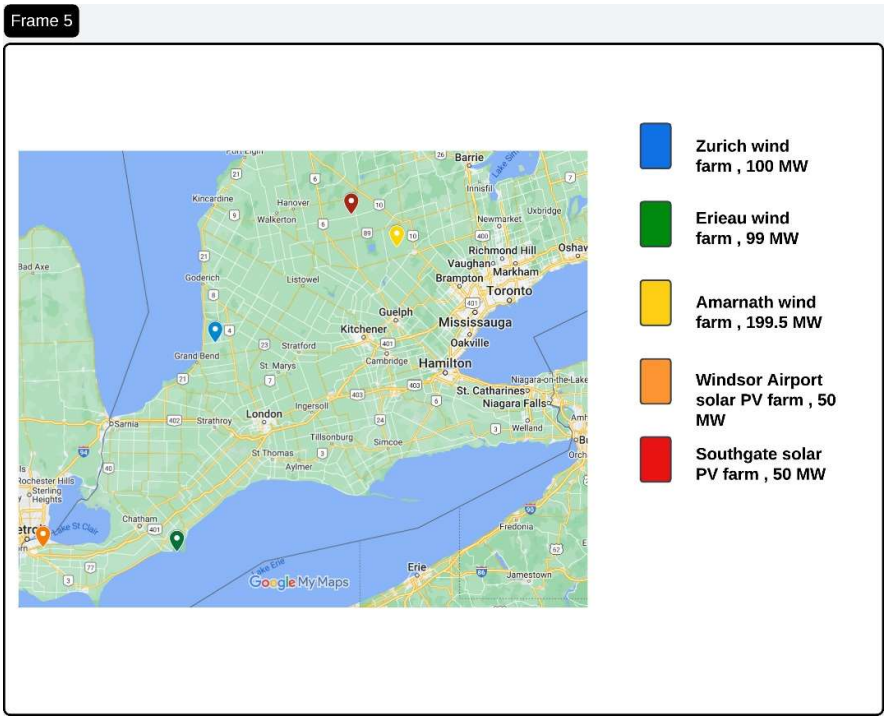

Figure 3. Power plant's location and capacity

2.1.2. Handling null values

The null values in the columns are handled by using different techniques as listed below

1. Exponential weighted moving average
2. Rolling window,
3. Filling last hour value
4. Moving average

The input data set had missing values as shown in Figure 4. The number of missing values in the data set is high, therefore further analysis is required before handling the missing values. Since most data missing were related to solar radiation as shown in Figure 4, seasonal solar radiation analysis was performed.

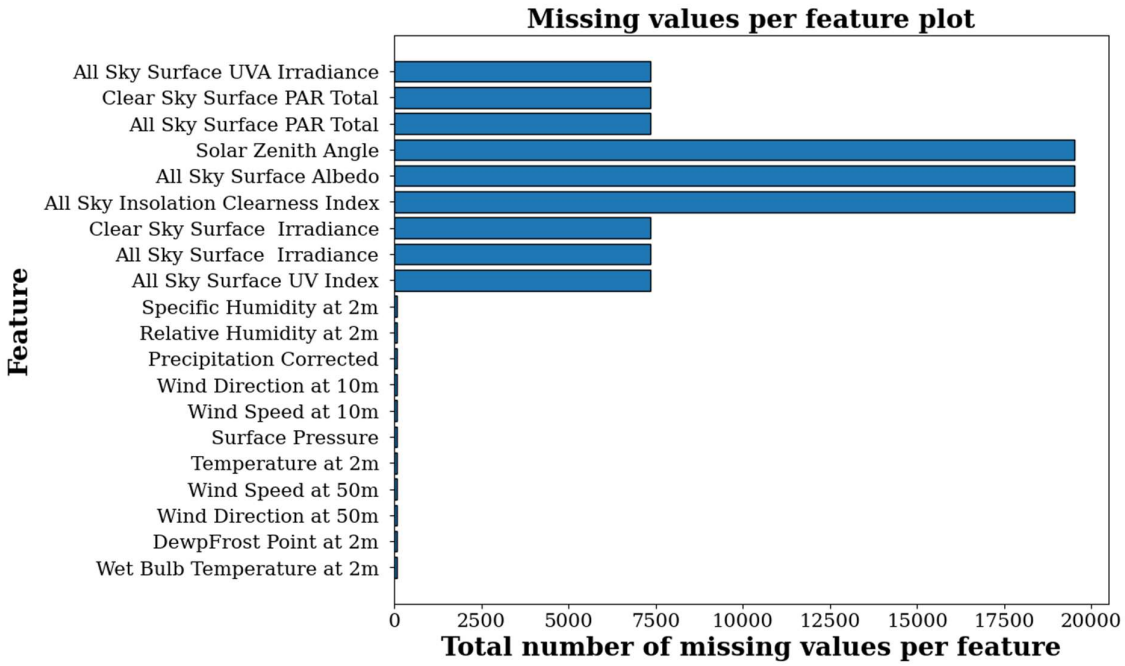

Figure 4. Missing values per feature count

The input data was split into two based on timing as day time dataset (morning 6 to evening 6) and night time data (evening 6 to morning 6). The sum of solar radiations in each df is analysed. Further the data is split into summer and winter data to examine the effect of missing data as shown in Figure 5. It is seen that the total annual solar radiation contribution is maximum during summer months and minimal during winter months. Therefore the missing data for solar radiation could be removed from winter months in the evening time could be removed. Then

the rolling average of 6 hours is done. On further analysis its found that the data was missing for a consecutive 3 month period, which cannot be filled by any means.

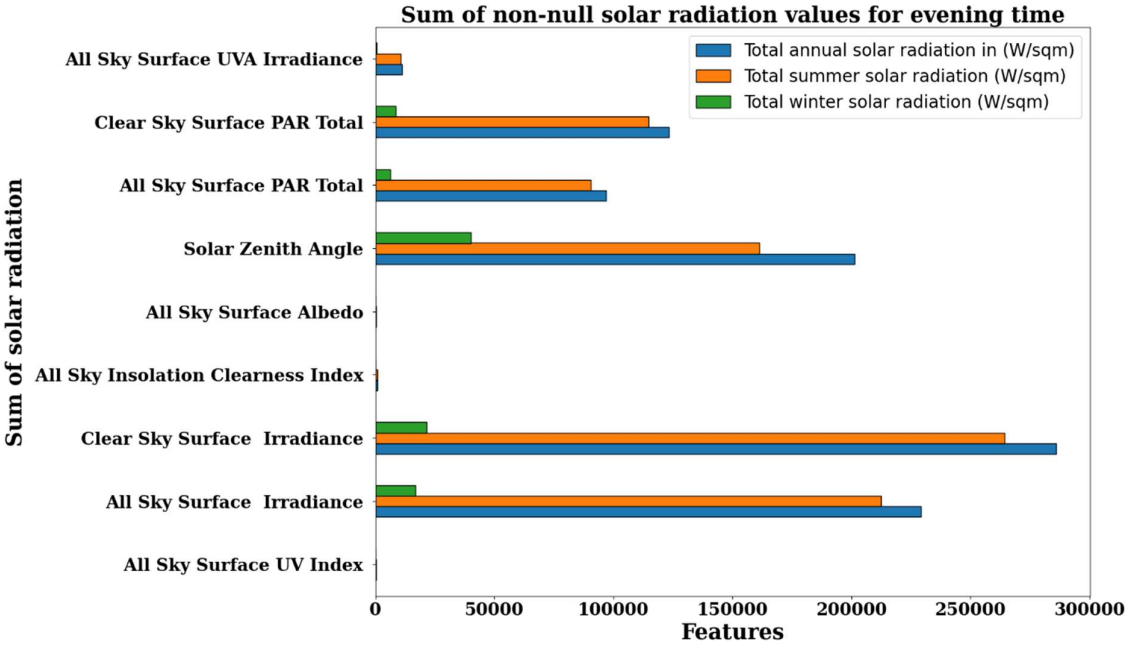

Figure 5. Sum of solar radiation seasonally

Before dropping the feature pearsons corelation matirix is developed to determine the interdependancy among the features as shown in Figure 6. It is seen that almost all the solar radiations are highly dependent on each other with correlation coefficient 0.99. Therefore those features could be dropped.

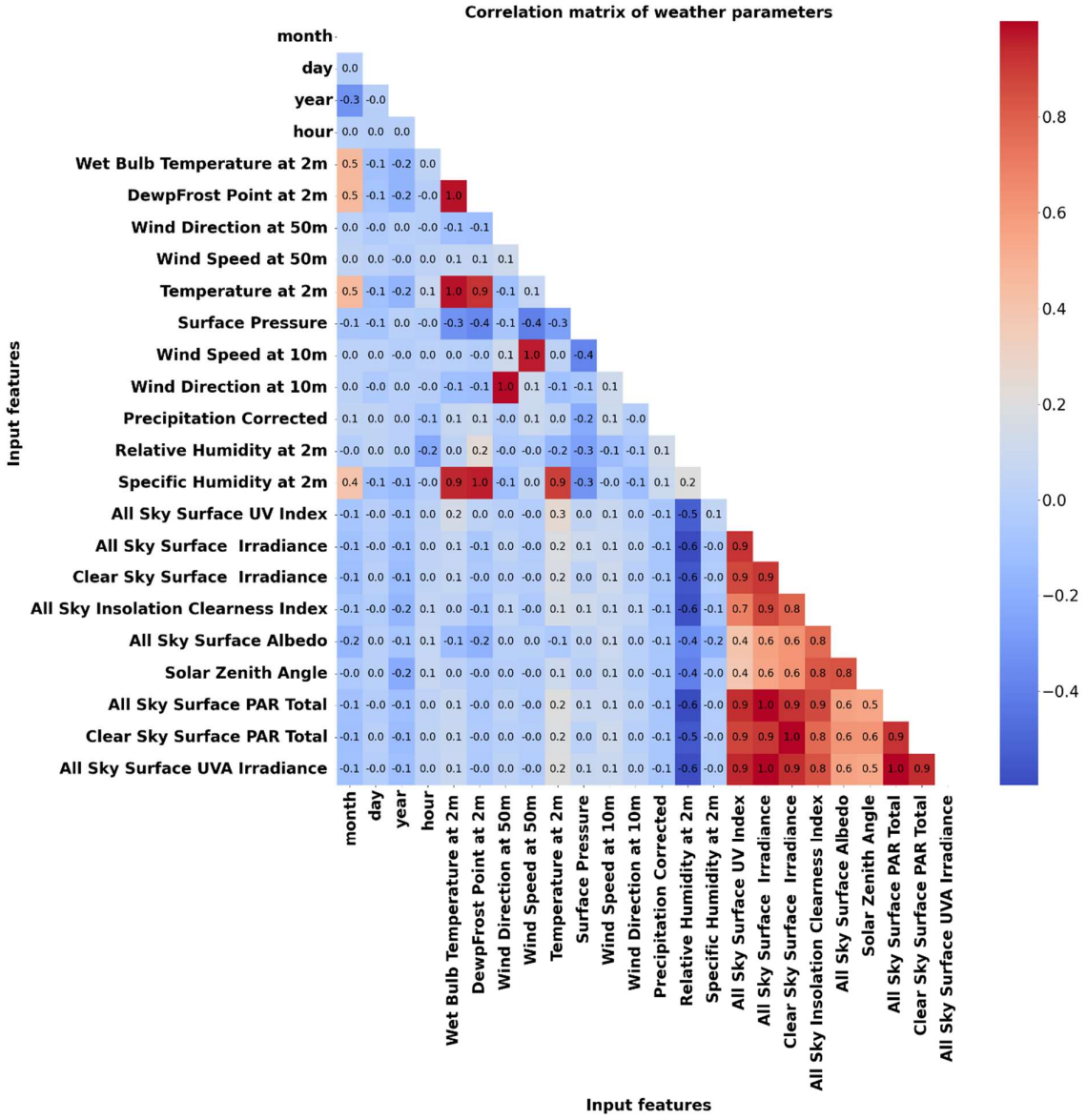

Figure 6. Pearson's correlation matrix for weather parameters

2.1.3. Time series generator

The prediction of power output for solar PV and wind farms in advance requires special data transformations. This enables the forecasting of future power output based on weather forecast data. The hour of the day for which the prediction is made is denoted as P(t), and the input weather parameters for the prediction include Temperature(t), Windspeed(t), and more.

For a three-hour ahead power output prediction, weather forecast data from the previous three hours must be included, such as Temperature(t-1), Windspeed(t-1), Temperature(t-2),

Windspeed(t-2), and so on increasing the number of rows as shown in Figure 7. The data for 1, 2, 4, 5, 6 hours ahead predictions are transformed in the same manner as shown in Figure 7.

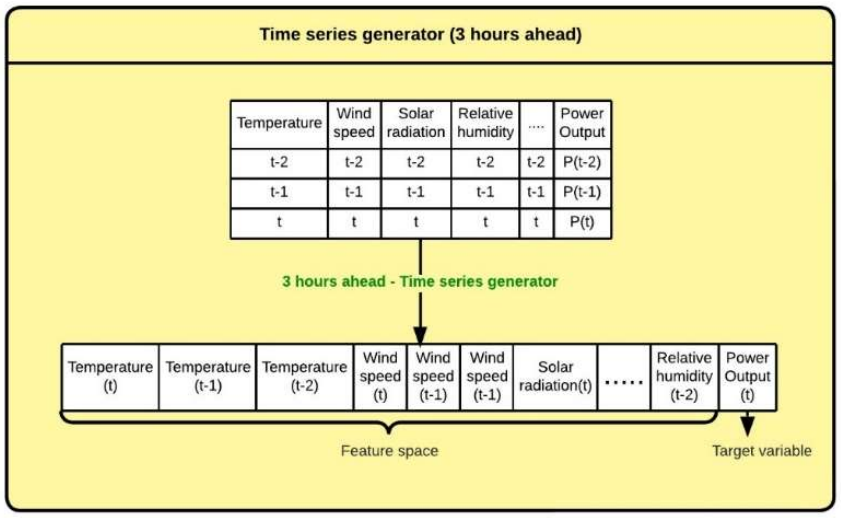

Figure 7. Time series generator 3 hours ahead

2.2 Exploratory data analysis

2.2.1 Wind farm energy output analysis and feature selection

The analysis of power output trend of wind farm at Erieau, between June-2019 to January-2023 along with average temperature, average windspeed and total solar radiation are shown in Figures 8-11. The power output trend shows a stationary seasonality with high power output in winter months (October to April) and low power output in summer months (May to September) each year . The temperature vs power output shown in Figure 8, shows that with the increase in temperature the wind power output decreases.

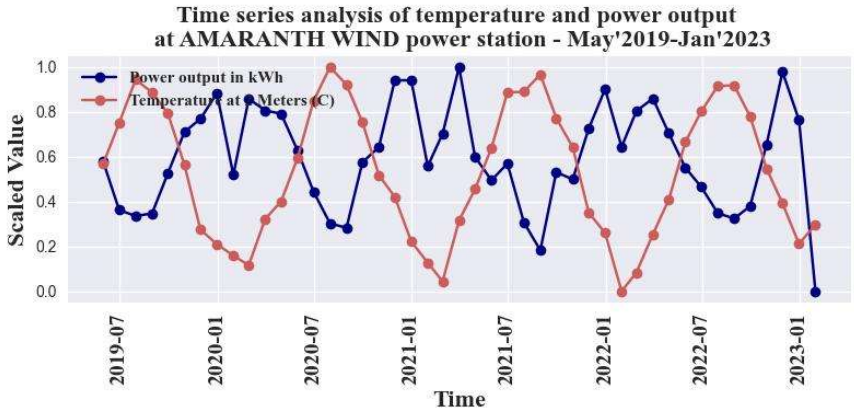

Figure 8. Timeseries analysis of wind farm power output and temperature between May 2019 – January 2023

The wind speed vs power output shown in Figure 9, shows that with the increase in wind speed the wind power output increases.

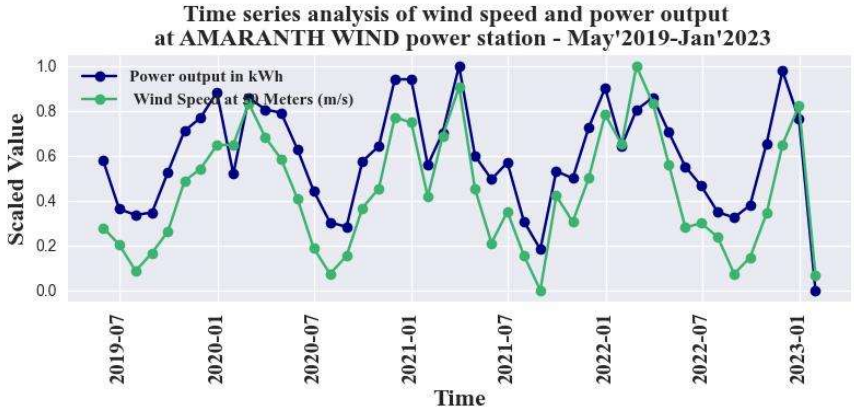

Figure 9. Timeseries analysis of wind farm power output and wind speed between May 2019 – January 2023

The solar radiation vs power output plot shown in Figure 10, shows that with the increase in solar radiation the wind power output decreases. This may be due the correlation between temperature and solar radiation which share a direct relationship among them and inverse relationship with the wind power output.

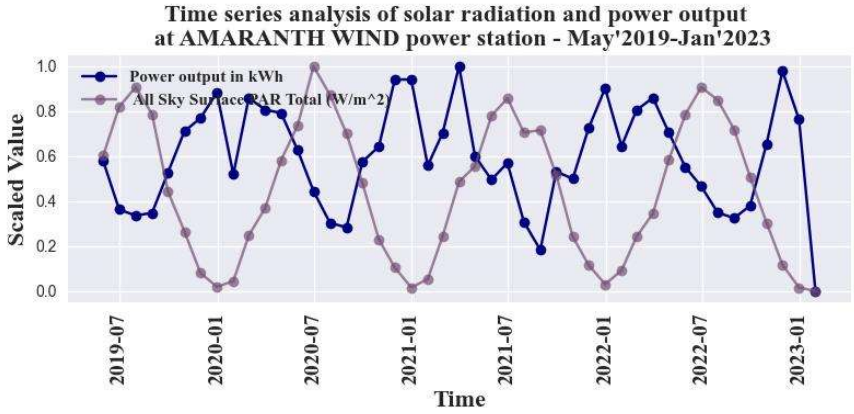

Figure 10. Timeseries analysis of wind farm power output and solar radiation between May 2019 – January 2023

The relative humidity vs power output plot shown in Figure 11, shows that with the increase in relative humidity the wind power output increase. This doesn't confirm any correlation

between relative humidity and wind speed but shows both the variables exhibits a direct relationship with the wind power output.

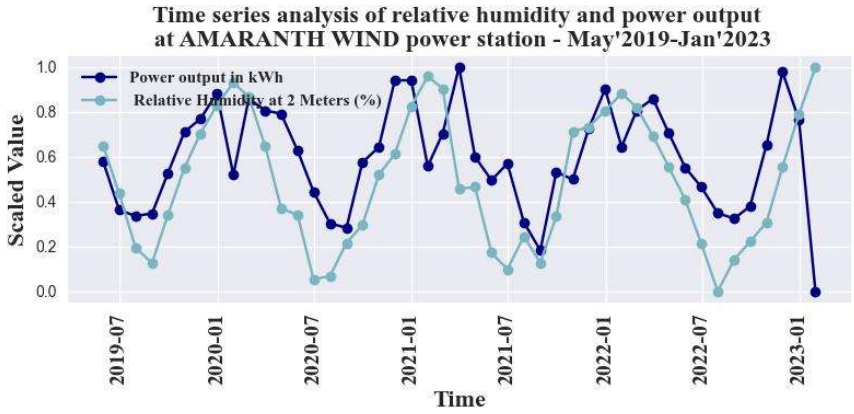

Figure 11. Timeseries analysis of wind farm power output and relative humidity between May 2019 – January 2023

The total monthly power output trend of the wind farm at Erieau is shown in Figure 12, along with the average monthly temperature, windspeed and total monthly solar radiation. The analysis indicates that during the winter season (October – April), wind farm power output is high, corresponding to high wind speed, relative humidity, demonstrating a direct relationship among these three variables. This pattern is consistent across all three wind farms. In addition, the analysis shows an inverse relationship between temperature and solar radiation with the power output from the wind farms.

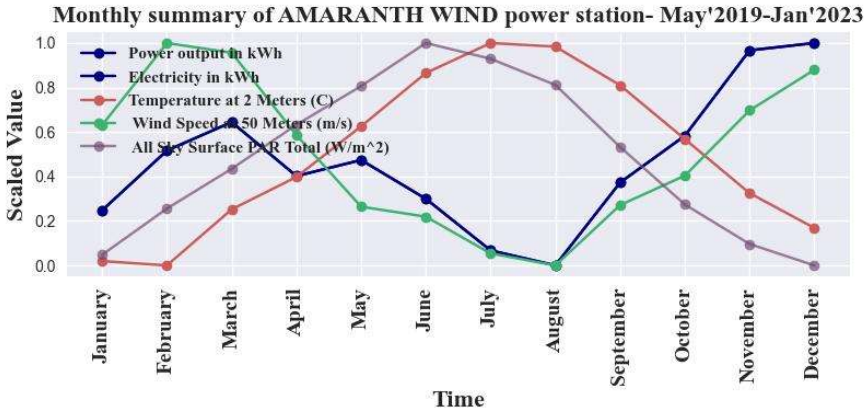

Figure 12 . Monthly trend analysis of wind farm at Erieau

The hourly total power output of the Erieau wind farm exhibits a cyclic pattern as shown in Figure 13. During night hours when wind speed and relative humidity are high and temperature and solar radiation are low, the power output is at its peak. Conversely, wind power output is

high when wind density and wind speed are elevated, while high temperature and solar radiation result in reduced wind density, leading to lower energy output during daylight hours when both are high.

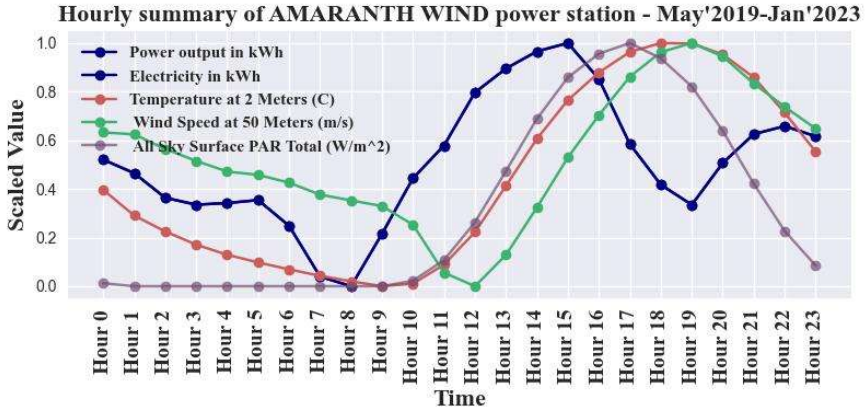

Figure 13. Hourly summary of wind farm at Erieau

The feature importance chart, as shown in Figure 14, illustrates the correlation between wind farm at Erieau's power output and its corresponding weather parameters. This chart highlights the parameters with the highest correlation to the target power output variable. Specifically, the wind speed and relative humidity exhibit the strongest correlations with the output power.

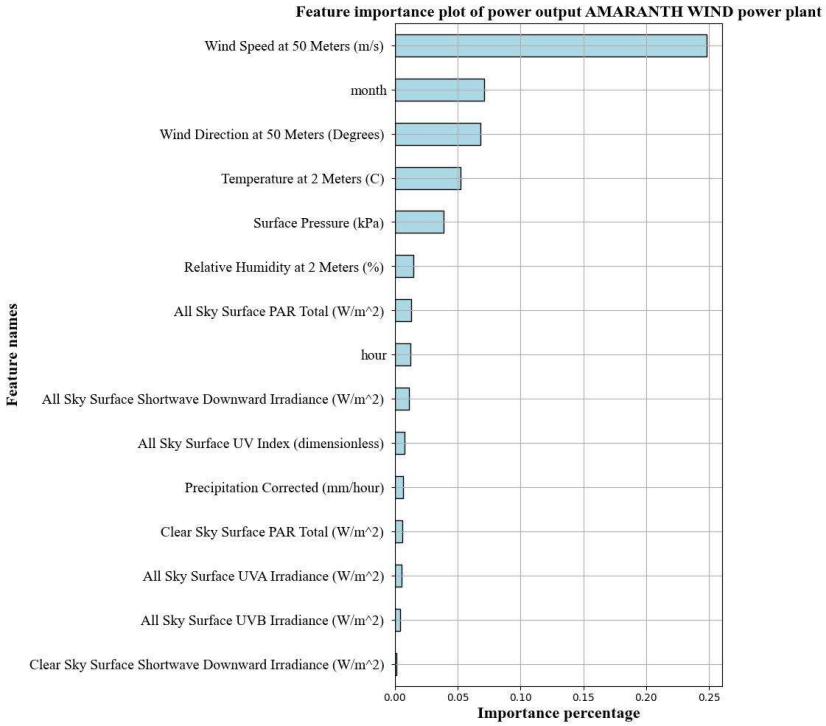

Figure 14. Feature importance of wind farm power output with respect to weather parameters

2.2.2 Solar PV energy output analysis and feature selection

The analysis of power output trend of Southgate PV farm between June-2019 to January-2023 along with average temperature, average windspeed and total solar radiation are shown in Figure 15-17. The power output trend shows a stationary seasonality with high power output in summer months and low power output in winter months. The temperature vs power output shown in Figure 16, shows that with the increase in temperature the solar PV power output increases.

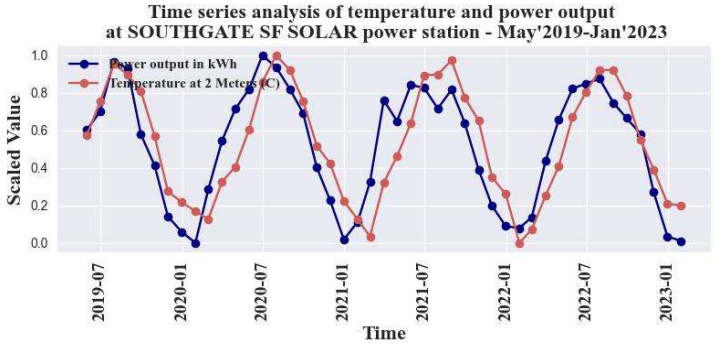

Figure 15. Timeseries analysis of solar PV power output and temperature between May 2019 – January 2023

The wind speed vs solar PV power output shown in Figure 16, shows that with the increase in wind speed the solar PV power output decreases.

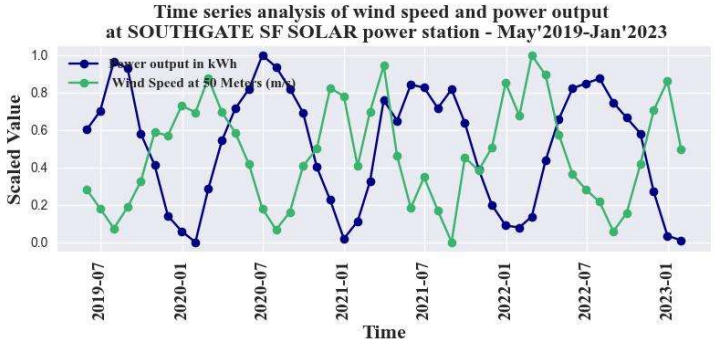

Figure 16. Timeseries analysis of solar PV power output and wind speed between May 2019 – January 2023

The solar radiation vs solar PV power output plot shown in Figure 17, shows that with the increase in solar radiation the solar PV power output increases. This may be due the correlation

between temperature and solar radiation which share a direct relationship among them and also a direct relationship with the solar PV power output.

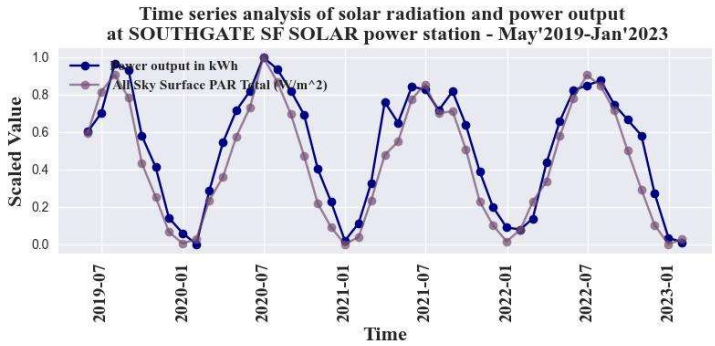

Figure 17. Timeseries analysis of solar PV power output and All Sky Surface Radiation between May 2019 – January 2023

The relative humidity vs power output plot shown in Figure 18, shows that with the increase in relative humidity the solar PV power output decrease. This doesn't confirm that both relative humidity and wind speed has indirect relationship with the solar PV power output.

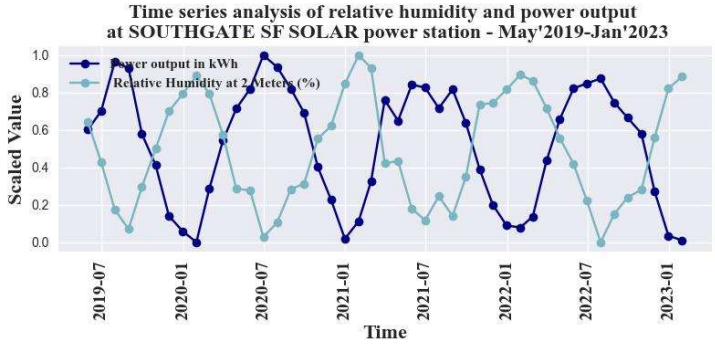

Figure 18. Timeseries analysis of solar PV power output and relative humidity between May 2019 – January 2023

The total monthly power output trend of the Southgate PV farm is shown in Figure 19, along with the average monthly temperature, windspeed and total monthly solar radiation. The analysis indicates that during the summer season (May – September), solar PV farm power output is high, corresponding to high total solar radiation and mean temperature, demonstrating a direct relationship among these three variables. All the solar PV farms show similar pattern

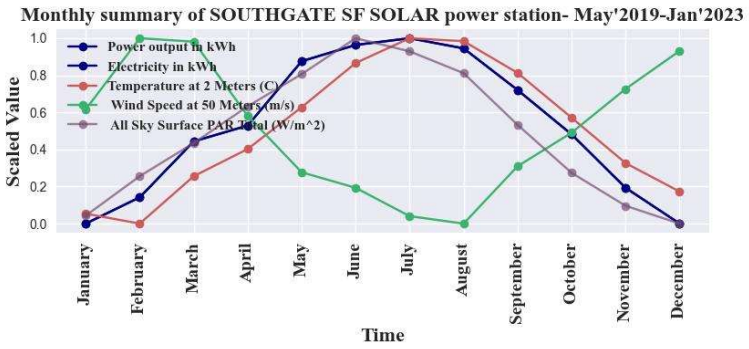

Figure 19. Monthly trend analysis of Southgate PV farm power output

The hourly total power output at the Southgate PV farm, as shown in Figure 20, exhibits a cyclic pattern, with power output being high during daylight hours when solar radiation is abundant and decreasing during the night when solar radiation is minimal.

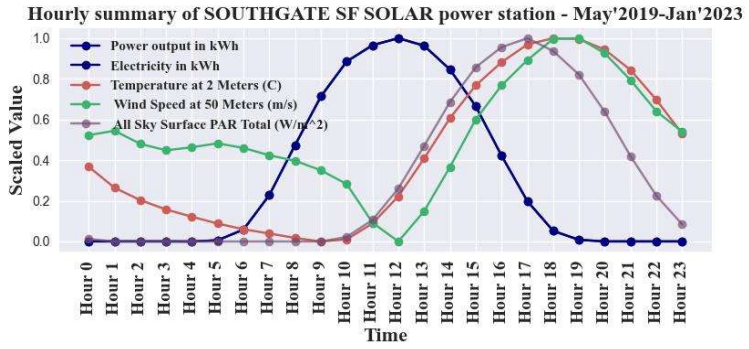

Figure 20. Hourly summary of Southgate PV farm power output

The feature importance chart, as shown in Figure 21, illustrates the correlation between Southgate PV power output and weather parameters. This chart highlights the parameters with the highest correlation to the target power output variable. Specifically, the hour of the day, month of year and solar radiation exhibit the strongest correlations with the output power.

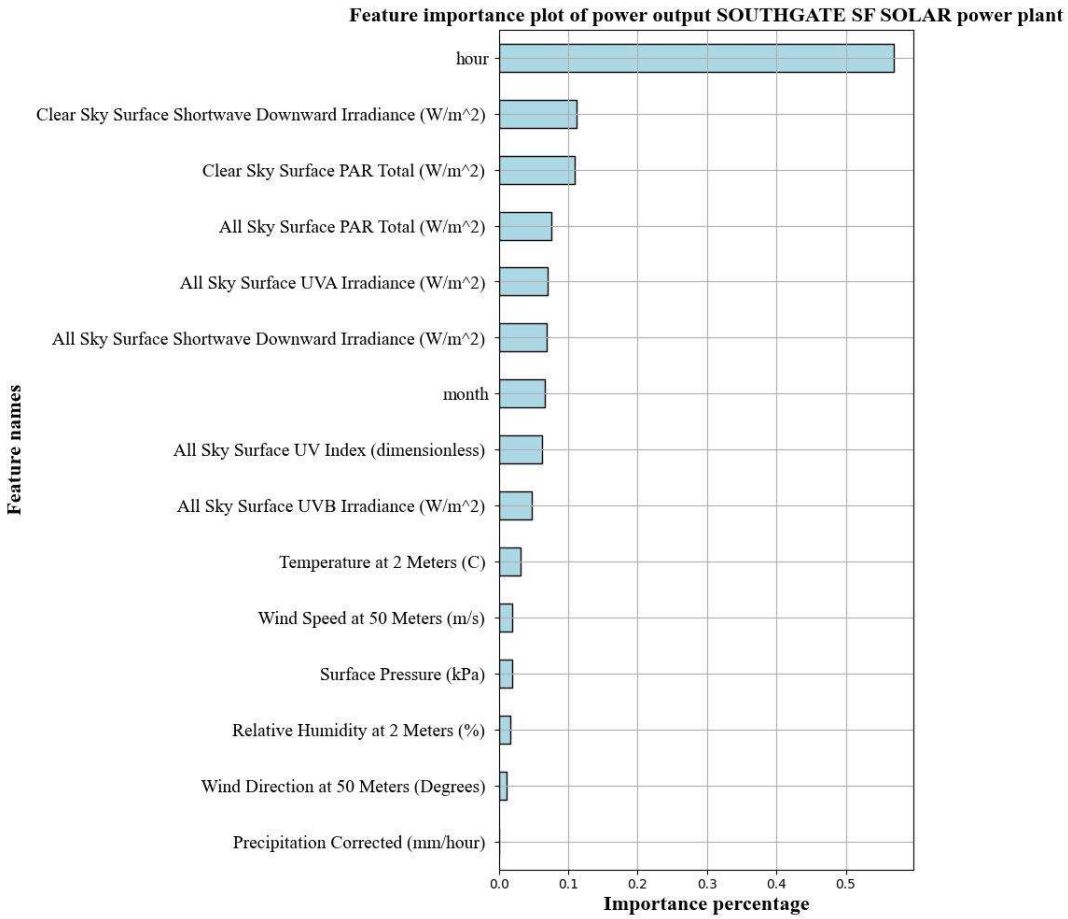

Figure 21. Feature importance of weather parameters with respect to power output

3. Machine learning algorithms

This section describes the mathematics of ML algorithms involved in the development of forecast model for solar PV and wind farm power output. The models discussed are namely, multiple linear regression, MLPRegressor, Support Vector Regression (SVR), random forest regressor, LSTM. The hyperparameter values for the above algorithms were arrived from the review papers [6-8, 10-12, 15,16]. Based on the base parameters each algorithm was trained for a trail run with hyper parameter optimization and the best parameters for the current dataset is determined and is discussed in the below section.

3.1. Multiple linear regression

The model assumes a linear relationship between the input variables and formulates the equation for the line of best fit based on the input variables to determine the target variable. The equation is given below

$$x = \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \ldots + \beta_p y_p + \epsilon \quad \text{Eq.1}$$

Where $\beta_1 \ldots \beta_n$ is the coefficient of input parameters, β_0 is the bias term and ϵ is the error term of the equation. The hyperparameter used for linear regression are given in Table 2.

The value for the parameter is achieved through trial-and-error method, since the algorithm complexity is less

Table 2. Hyperparameters for multiple linear regression

SN	Hyperparameters	Value
1	Learning rate	0.1

3.2 Random Forest regressor

The Random Forest regressor is an ensemble learning technique that makes predictions by aggregating the average result of training multiple decision trees using the input features and splitting rules. The hyperparameters used for random forest regressor are given in Table 3.

The hyper parameters base values are arrived [8,10, 15, 16]. Based on the hyperparameter range hyperparameter tuning is performed to achieve the below value used for model training.

Table 3. Hyperparameters for random forest regressor

SN	Hyperparameters	Value
1	n_estimators	100
2	max_depth	20
3	min_samples_split'	4
4	Max_features	'auto'
5	bootstrap	True
6	Criterion	MSE
7	'max_samples'	0.7

3.3 Multi-Layer Perceptron Regressor

Multi-Layer PerceptronRegressor (MLPRegressor) is a feed forward artificial neural network model used develop forecasting models in which the inputs features and the target variables have a non-linear relationship. The architecture of MLPRegressor is shown in Figure 13. The MLPRegressor made of four-layer, first layer (input), second layer (hidden), third layer (hidden), output. The hyperparameters used for MLPregressor are given in Table 4.

Frame 3

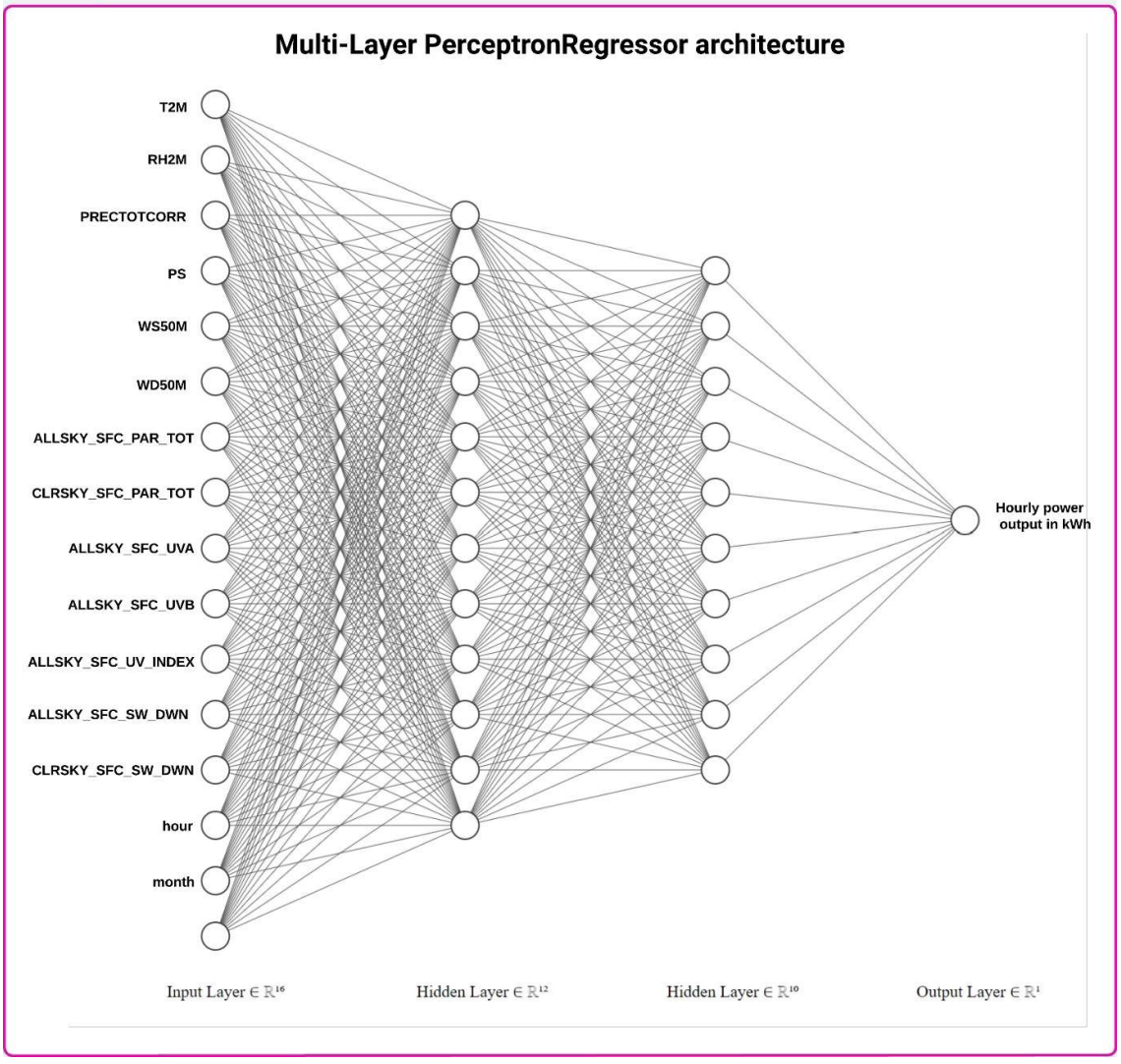

Figure 22. MLPRegressor architecture

The hyper parameters base values are arrived [6-8, 10, 11]. Based on the hyperparameter range hyperparameter tuning is performed to achieve the below value used for model training.

Table 4. Hyperparameters for MLPRegressor

SN	Hyperparameters	Value
1	hidden_layer_sizes	(60, 60,)
2	activation	'relu'
3	solver	Adam
4	learning_rate	constant
5	alpha	0.001
6	max_iter	1000

3.4 Support vector regression

Support vector regression (SVR) is a variant of support vector machines (SVM) used to determine continuous target variable (regression). A tolerance level is set for the line of best fit, i.e., SVR allows for some error is predictions as mentioned by epsilon. Unlike other regression models SVR not fits the data to the line of best fit, there is a position for a tube around the line of best fit which is consider to be good predictions. SVR handles outliers well, because the position of the tube depends on the points on the tube and outside the tube (support vectors)

The hyper parameters base values are arrived [8, 10, 11, 15, 16]. Based on the hyperparameter range hyperparameter tuning is performed to achieve the below value used for model training.

The hyperparameters used for support vector regression is shown in Table 5.

Table 5. Hyperparameters for support vector regression

Sno	Hyperparameters	Value
1	Kernel	Rbf
2	C	1
3	Epsilon	0.2

3.5 Long Short-Term Memory (LSTM)

LSTM is a variant of recurrent neural network used to make predictions out of sequential data, identify the pattern in the time series data. The LSTM model has a provision to store the information from the previous data points (memory cells). LSTM model controls the

information entry into the memory cells through input gate, controls the hidden state and prediction through output gate and use forget gate to throw away or store information.

$$[forget\_gate(f_t) = \sigma\left(W_f \cdot \left[h_{\{t-1\}}, x_t\right] + b_f\right)] \quad \text{Eq. 2}$$

$$[input\_gate(i_t) = \sigma\left(W_i \cdot \left[h_{\{t-1\}}, x_t\right] + b_i\right)] \quad \text{Eq. 3}$$

$$[candiate\_cell\_state(\tilde{}\{C\}_t) = tanh\left(W_C \cdot \left[h_{\{t-1\}}, x_t\right] + b_C\right)] \quad \text{Eq.4}$$

$$[output\_gate(o_t) = \sigma\left(W_o \cdot \left[h_{\{t-1\}}, x_t\right] + b_o\right)] \quad \text{Eq. 5}$$

$$[hidden\_gate\ (h_t) = o_t \cdot tanh(C_t)] \quad \text{Eq. 6}$$

4. Wind energy prediction results

4.1 Multiple linear regression

Multiple linear regression models were employed to predict power output for various time horizons, ranging from 0 to 6 hours ahead. These models were trained using hourly power output data and corresponding weather features from three wind farms: Erieau, Amarnath, and Zurich. For each wind farm and specific time horizon, the prediction model underwent multiple iterations, each with different test-train splits (70/30, 80/20, 65/35, 50/50). The average performance across these four splits was considered the model's evaluation for that specific wind farm and time lag. In Figure 23, the combined R-squared (r^2) values for the multiple linear regression forecasting models for all three wind farms across various time horizons are depicted. The model's predictive performance is suboptimal when using only current-hour weather data (i.e., 0 hours ahead), with an average r^2 value of 0.42. However, when the data from the previous hour is included in the analysis, the model excels, achieving an average r^2 value of 0.91 for all time horizons, as shown in the Figure 23. The higher predictive performance with the previous hour datum, included could be misleading when evaluating the

model's performance. Therefore, the model is further investigated below to determine whether the model is free from underfitting and overfitting.

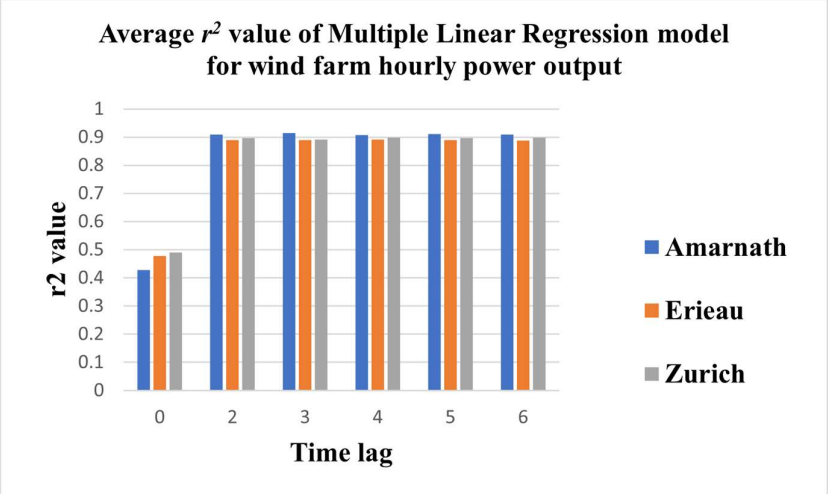

Figure 23. Multiple linear regression model's r^2 value for wind energy power output

The scatter plot of the actual power output of the Amarnath wind farm and the three-hour ahead power forecast using multiple linear regression is shown in Figure 24. The distribution of points in the scatter plot exhibits a linear trend, with most points positioned closely to the line of fit and a relatively low number of outlier points, as shown in Figure 24. This pattern suggests a good indication of predictivity for the multiple linear regression model, a conclusion further supported by the higher R-squared (r^2) value of 0.912 and the lower Mean Absolute Error (MAE) value of 9.06 for the multiple linear regression model.

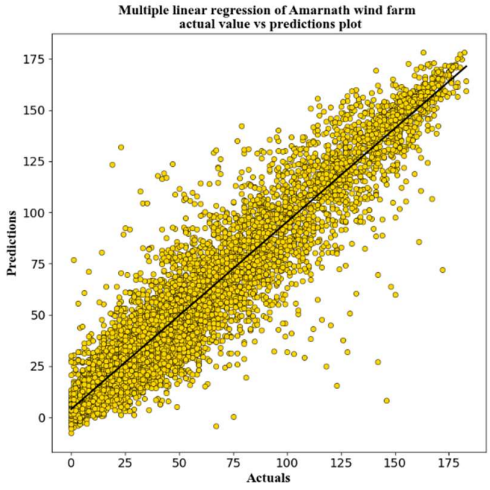

Figure 24.Multiple linear regression actuals vs predictions scatter plot

Residuals are the differences between the actual values and the prediction output, they measure how far data points deviate from the line of best fit. Residuals are also used to calculate the Mean Absolute Error (MAE), which is the square root of the sum of all residuals. However, it's essential to note that having a lower MAE with skewed residuals can be misleading.

The distribution plot of residuals from the three-hour ahead power forecast multiple linear regression model for the Amarnath wind farm fits a normal distribution curve, as illustrated in Figure 25. This distribution shows a nearly zero mean of residuals and a standard deviation of 1, indicating low variance in residuals. The narrow, normally shaped residual distribution without skewness suggests good predictivity of the model without overfitting. Conversely, if the residual curve is wider, it implies that the model's performance is lower, as the variance of the residuals is greater.

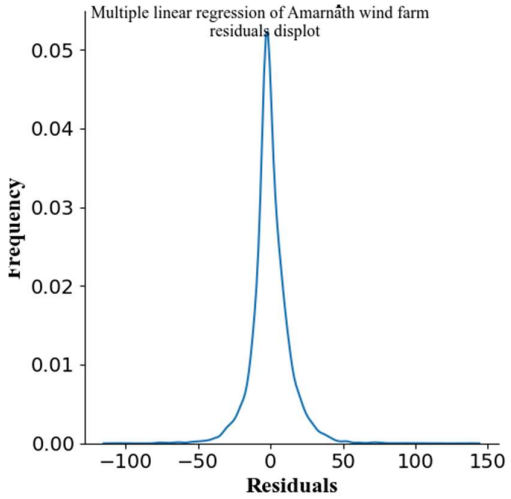

Figure 25. Multiple linear regression predictions vs actuals distribution plot of Amarnath wind farm

The prediction and their corresponding residual plot of the above model shown in Figure 26, displays a decreasing tunnel shape with major cluster of points around 0 and having high value outliers on the right side of the plot and low value outliers in the left side of the plot. The presence of outliers on both sides of the plot exhibits heteroscedasticity in the model's performance i.e., the variance of residuals is not constant across the range of predicted values, the model performs better or worse than the overall in certain cases. The higher r^2 value of 0.91 and lower MAE value of 9.06 could possible due to overfitting of the data to the multiple linear regression model and this is confirmed by the skewness of residuals in the predicted values as

shown in Figure 13. Therefore, multiple linear regression model is not a good fit for forecasting the power output of wind farms in different time horizons ahead.

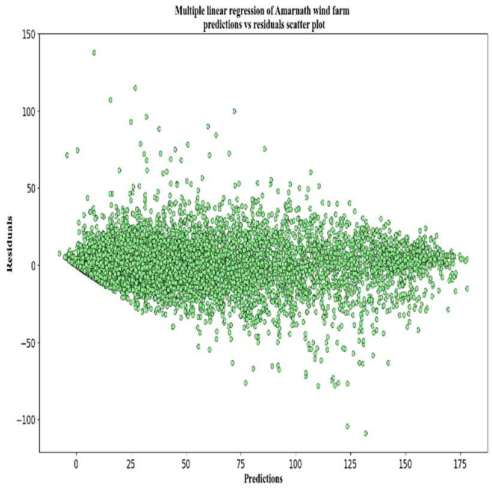

Figure 26. Multiple linear regression predictions vs residuals scatter plot

The r^2 values of the three wind farms power output forecasting using multiple linear regression for the time lags of 0, 2, 3, 4, 5, 6 hours is shown in Table 6.

Table 6. Average r^2 values of wind farms power output forecast models using multiple linear regression model

Time lag	Amarnath	Erieau	Zurich	Average for all plants
0	0.578936872	0.603426719	0.603185361	0.595182984
2	0.909964496	0.888140003	0.892300109	0.896801536
3	0.907665677	0.88491968	0.884048588	0.892211315
4	0.899781484	0.877985279	0.883137126	0.886967963
5	0.900682932	0.876246463	0.879033246	0.88532088
6	0.895704116	0.864521426	0.875162897	0.878462813
Average for each plant	**0.848789263**	**0.832539928**	**0.836144554**	**0.839157915**

4.2 MLPRegressor

MLPRegressor models for hourly power output forecasting were developed for three wind farms under study using the datum as discussed in earlier section. The models were designed

to predict power output at various time horizons, and the inclusion of weather data from the respective previous hours depended on the specific forecast horizon. Similar to the multiple linear regression models discussed earlier, each forecasting model for a particular wind farm and time horizon underwent training with multiple test-train splits, including scenarios of 70/30, 80/20, 65/35, and 50/50. The performance results were averaged across these splits to determine the overall model performance.

The overall averaged r^2 values for power output forecasting using MLPRegressor models for the three wind farms at various time horizons are presented in Figure 27. Notably, the MLPRegressor models exhibit enhanced predictivity when forecasting the power output for the current hour by utilizing weather data from the same hour. This results in a notably improved average r^2 value of 0.65 compared to the multiple linear regression model discussed earlier. Conversely, there is a slight decrease in r^2 values for the MLPRegressor models when predicting power output at wind farms for different time horizons, specifically in the range of 2 to 6 hours ahead. The average r^2 value for this range is 0.89, as depicted in Figure 27. The higher values of r^2 must be further examined to determine if there is any overfitting of data or the model's capacity in capturing the data accurately.

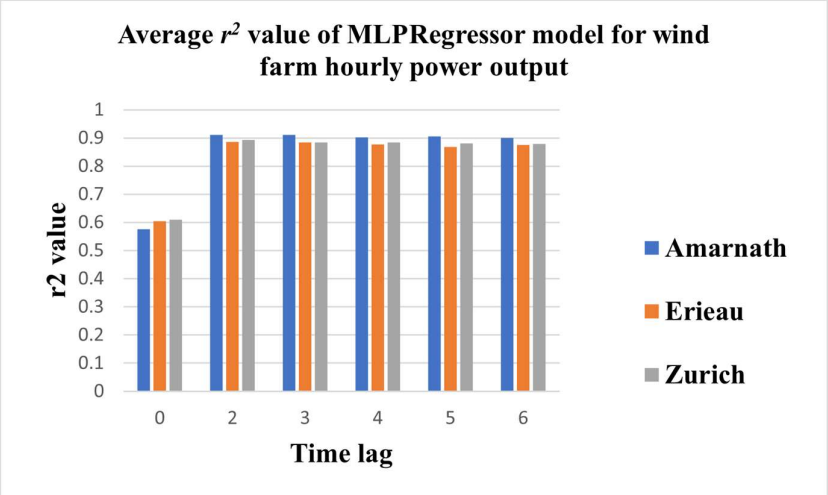

Figure 27. Average r^2 value of MLPRegressor models for predicting the power output in wind farms

The scatterplot in Figure 28, illustrates the actual power output of the Erieau wind farm alongside the three-hour ahead power output forecasts generated using the Multi-Layer Perceptron Regressor. The data points on the plot closely align with the line of best fit, with some points positioned above and below the line, reflecting prediction errors. The presence of

outliers on both sides of the line of best fit suggests that the model excels in some instances while performing less effectively in others. Notably, the plot also reveals a few predictions that fall below zero, implying a power output less than zero, and others exceeding the plant's capacity of 100MW. It's vital to acknowledge the physical impossibility of a power plant consuming electricity or producing more electricity than its capacity allows, as seen in the Figure 28.

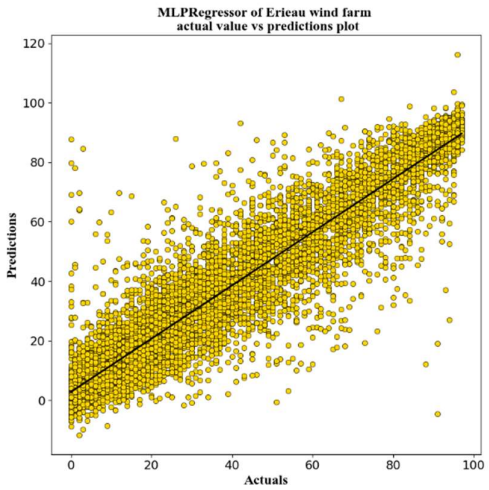

Figure 28. MLPRegressor actuals vs predictions scatterplot for wind farm at Erieau

The residuals distribution plot of the MLPRegressor model for forecasting the Erieau wind farm power output three hours ahead is shown in Figure 29. The distribution plot fits a normal curve with almost zero mean and standard deviation 1. The plot reveals minimal skewness in the residuals which suggest the good predictivity of the model. However, the predictions vs residuals plot must be accessed to determine the variance of the residuals with respect to predictions before concluding about the overall model performance.

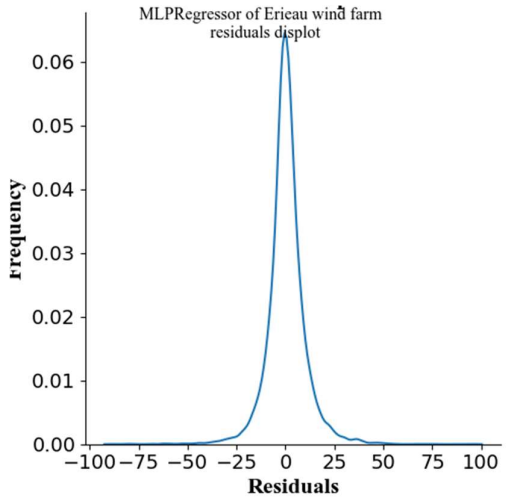

Figure 29. Residual plot of wind power output forecast using MLPRegressor.

The examination of the prediction and residual plot of the Erieau wind farm power output forecasting model using MLPRegressor model shown in Figure 30, reveals a downward tunnel pattern as observed in multiple linear regression prediction and residual plot. The downward tunnel pattern with most points clustered around 0, shows most points are exactly predicted, with some points above in the left side with predictions having more residuals (prediction higher than actual) and some points lower in the right side with predictions having less residuals (predictions less than actuals). This higher values of residuals and lower value of residuals in Figure 30, reveals higher variance in residuals and the residuals are skewed. Therefore, the higher value of r^2 is due to overfitting of the data rather than the model capturing the actual behaviour of the pattern of the variables.

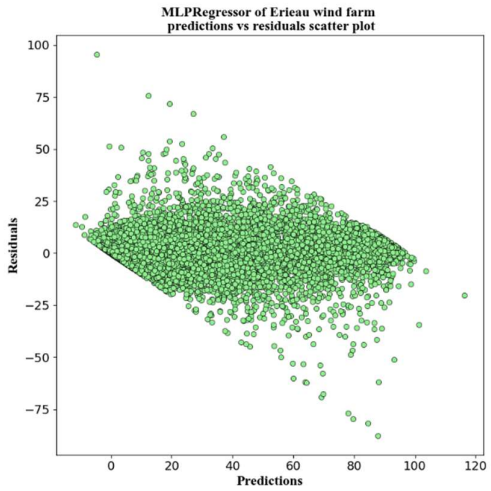

Figure 30. MLPRegressor predictions vs residuals plot

The r^2 values of the three wind farms power output forecasting using MLPRegressor for the time lags of 0, 2, 3, 4, 5, 6 hours is shown in Table 7 .

Table 7. Average r^2 values of wind farms power output forecast models using MLPRegressor model

Time lag	Amarnath	Erieau	Zurich	Average for all plants
0	0.526686472	0.56742843	0.603185361	0.558283033
2	0.890442629	0.879695356	0.892300109	0.886515216
3	0.883996413	0.872079129	0.884048588	0.879239934
4	0.876303621	0.864118126	0.883137126	0.872796124
5	0.872422135	0.859013009	0.879033246	0.868380707
6	0.864787501	0.84931344	0.875162897	0.860672956
Average for each plant	**0.819106462**	**0.815274582**	**0.836144554**	**0.820981328**

4.3 Random Forest regressor

The ensemble learning technique, random forest regressor was used to train the models to forecast hourly power output of wind farms with various time horizons ahead. The averaged r^2 values of hourly power output forecast models of the three wind farms for 0 to 6 hours ahead time horizon is shown in the Figure 31. The ensemble learning technique has significant increase in r^2 when comparing the multiple linear regression and MLPRegressor models for predicting the current hour power output with an r^2 value of 0.68. These models predict the wind energy output of farms with an average r^2 value 0.89 as shown in Figure 31. The model output is further investigated to determine the presence of overfitting.

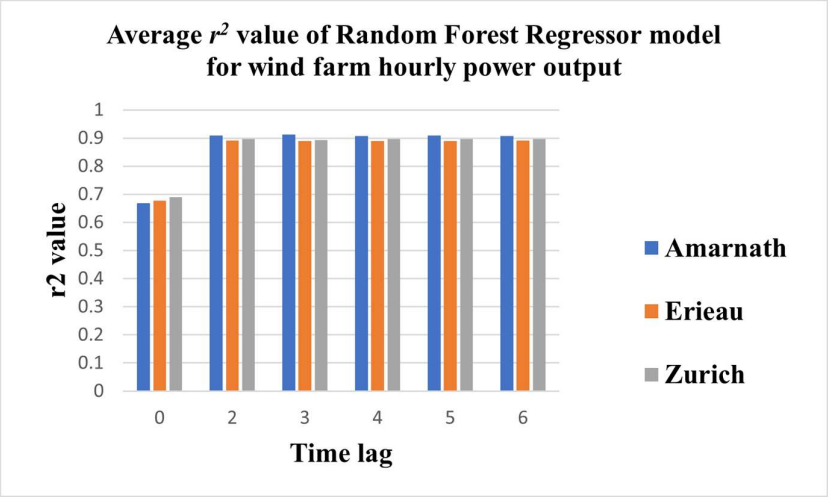

Figure 31. Average r^2 value of random forest regressor models for predicting the power output in wind farms

The scatterplot of the hourly wind energy output and the predictions for the random forest regressor model is shown in Figure 32. The scatter plot shows a linear trend with the predictions almost equalling the actual values. The point aligns themselves toward the line of best fit, with some points above and below the line. The Figure shows that all the predictions were between 0 to 199, which is well within the range of the wind farm capacity (0 to 199 MW capacity). That is either the wind farm produced no output nor the energy output was maximum not exceeding the maximum.

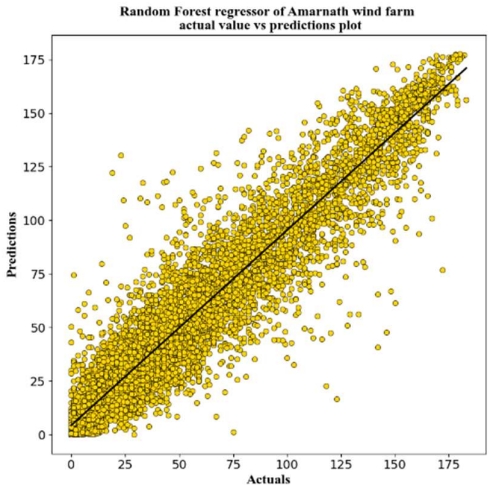

Figure 32. Actual power output vs Random Forest regressor power predictor of Amarnath wind farm

These difference in residuals is examined in the residual plot shown in Figure 33. The normal cure, 0 mean and 1 standard deviation of the residual distribution curve shows that the residuals are normally distributed and the model predictivity is good.

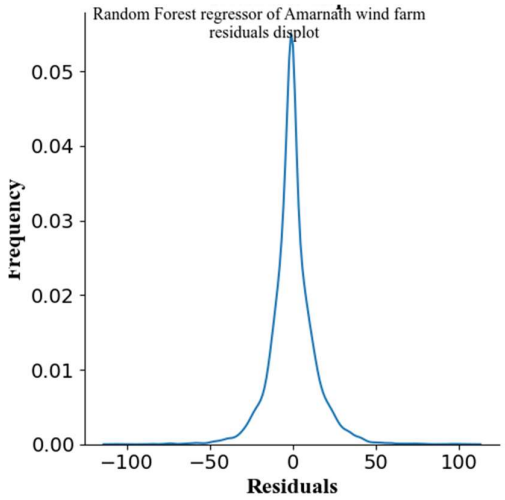

Figure 33. Residual distribution plot of wind farm power predictor using Random Forest regressor

The predictions vs the residuals plot is illustrated in Figure 34, shows the presence of heteroscedastic, outliers and patterns in the model. The points are distributed in a random pattern centred around 0 rather than like a tunnel pattern observed in multiple linear regression and MLPRegressor. This conforms that the model is free of heteroscedastic, i.e., variance of residuals is low. But they are still outliers in the model with points well above and below the graph. This conforms that Random Forest Regressor is one of the good fits for predicting the wind energy output of farms with an average r^2 value of 0.89.

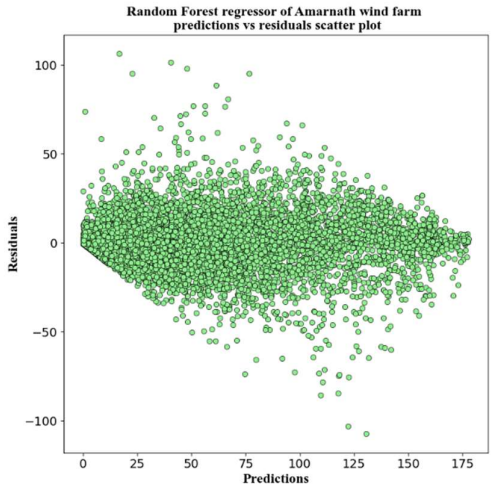

Figure 34. Prediction vs residuals scatter plot of random forest regressor predictor for wind energy

The r^2 values of the three wind farms power output forecasting using random forest regressor for the time lags of 2, 3, 4, 5, 6 hours is shown in Table 8 .

Table 8 Average r^2 values of wind farms power output forecast models using random forest regressor model

Time lag	Amarnath	Erieau	Zurich	Average for all plants
2	0.909964496	0.888140003	0.892300108	0.896801536
3	0.907665677	0.88491968	0.884048588	0.892211315
4	0.899781484	0.877985279	0.883137126	0.886967963
5	0.900682932	0.876246463	0.879033246	0.88532088
6	0.895704116	0.864521426	0.875162897	0.878462813
Average for each plant	**0.902759741**	**0.87836257**	**0.882736393**	**0.887952901**

4.4 Support Vector Regression

The hourly wind energy output forecasting model for various time horizons ahead was developed using Support vector regression (SVR) algorithm. The average r^2 value of the hourly power output SVR model for the three windfarms is shown in Figure 35. The overall model r^2 value is around .80, which is much less when compared with the other two models. The r^2

confirms that the model is not capturing the behaviour of the input problem. The scatter plots of residuals and predictions, actuals and predictions, distribution plot of residuals was examined as shown in Figure 36, 37, 38, respectively.

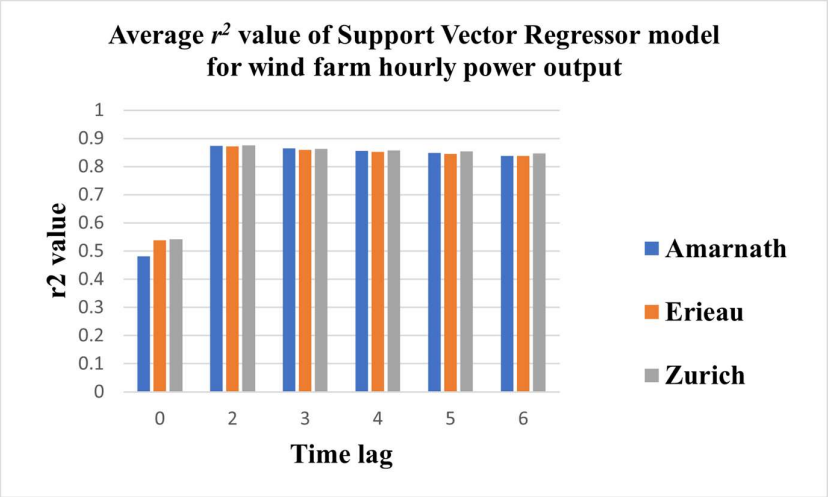

Figure 35. Average r^2 value of support vector regression models for predicting the power output in wind farms

The scatterplot of actual hourly energy output of Amarnath wind farm and the SVR model predicted power output 3 hours ahead is shown in Figure 36. The points align themselves toward the line of best fit, but all the points are slightly away from the line of best fit, i.e., for an actual value predicted value is not accurate and is subjected to error.

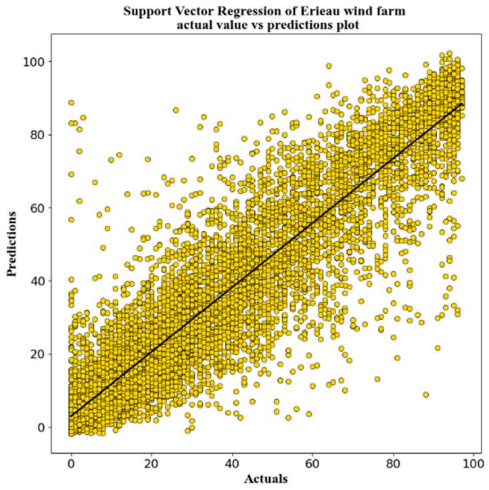

Figure 36. Actual power output vs support vector regressor power predictor of Erieau wind farm

The residual distribution plot (a way to visualize the error in predictions) shown in Figure 37, exhibits a normal curve with wide area under the curve, signifying greater error in predictions, which the higher MAE value confirms.

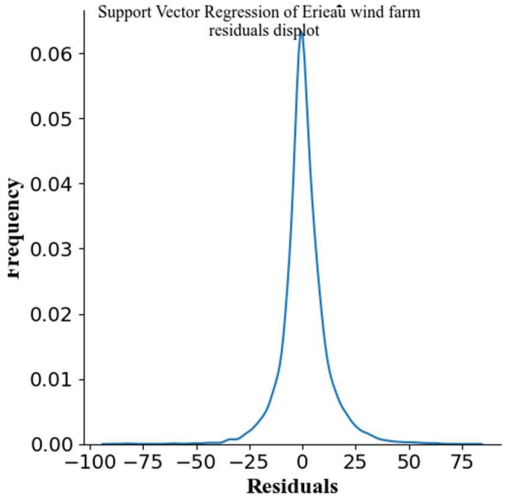

Figure 37. Residual distribution plot of wind farm power predictor using support vector regressor

The predictions vs residuals plot shown in Figure 38, exhibits a tunnel shape, which confirms heteroscedastic in the predictions. The points clustered around 0 with outliers in both the sides shows that the residuals are more skewed and the model doesn't capture the actual behaviour of the input datum.

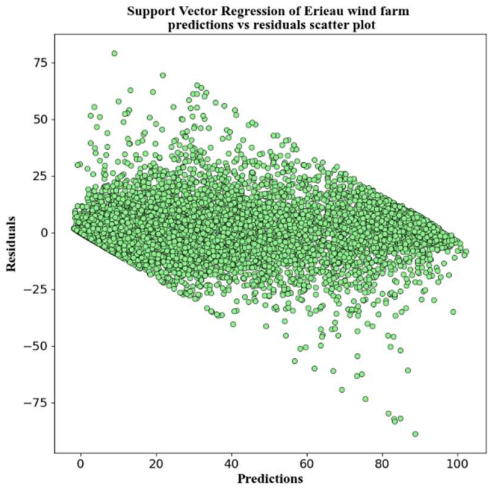

Figure 38. Prediction vs residuals scatter plot of support vector regressor predictor for wind energy

The r^2 values of the three wind farms power output forecasting using support vector regression for the time lags of 2, 3, 4, 5, 6 hours is shown in Table 9.

Table 9. Average r^2 values of wind farms power output forecast models using support vector regression model

Time lag	Amarnath	Erieau	Zurich	Average for all plants
2	0.870920763	0.871250708	0.873661613	0.872076457
3	0.860327149	0.859238578	0.862488508	0.860684745
4	0.852825758	0.850250973	0.85766762	0.852219394
5	0.844161339	0.841779555	0.853177225	0.844104533
6	0.833870887	0.834105455	0.847277984	0.835464817
Average for each plant	**0.852421179**	**0.851325054**	**0.864698744**	**0.854832876**

4.5 Long Short-Term Memory Model (LSTM)

The recurrent neural network model LSTM was used in forecasting the hourly power output of wind farms for various time horizons ahead. The average r^2 values of deep learning LSTM model trained for forecasting the power output of three wind farms on various time horizons ahead. The figure illustrates that the LSTM model has an average r^2 value of 0.92 for predicting the output of all wind farms and for all the time horizons ahead. The analysis of scatterplots of actual values and predictions, residuals distribution plots, predictions and residuals scatterplot in Figures 39, 40, 41, respectively, provides a deeper intuition towards the model's performance.

The scatterplot of the actual power output and the LSTM model predicted power output for three hours ahead time lag is showing a linear trend, with a smaller number of outliers as shown in Figure 39. The residual distribution plot shows a normal curve with zero mean and 1 standard deviation shows that the residuals are not skewed and properly distributed. The residual distribution plot show that the good predictability of the model.

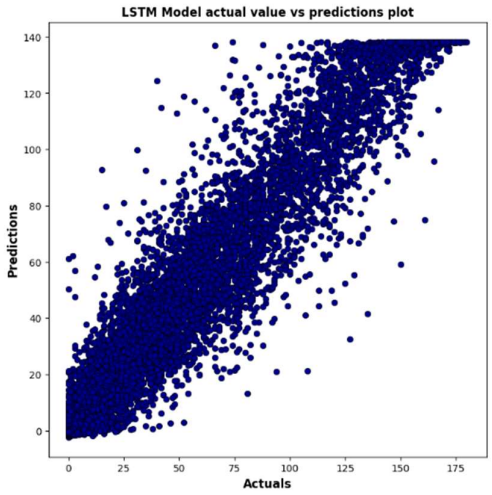

Figure 39. Actual power output vs LSTM power predictor of Amarnath wind farm

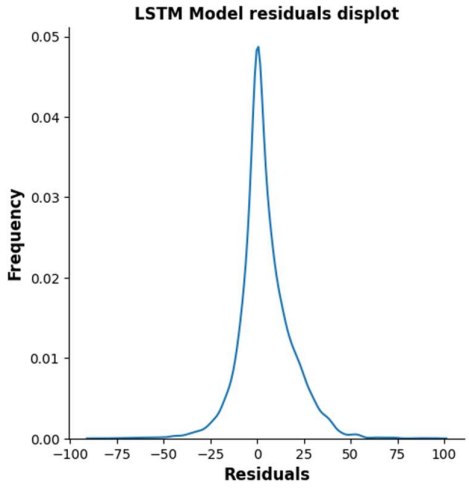

Figure 40. Residual distribution plot of wind farm power predictor using LSTM

The scatter plot of predictions and the residuals of the LSTM model shown in Figure 41, depicts a random pattern rather than a tunnel pattern as observed in other models. The random pattern with most points clustered above and below zero shows that the LSTM models were free of heteroscedastic and the residuals was not skewed and the variance of the residuals was low. The model has a very good predictability with higher r^2 value and lower MAE. The predictions and residuals plots show that the model captures the behaviour of the input datum rather than overfitting the input data.

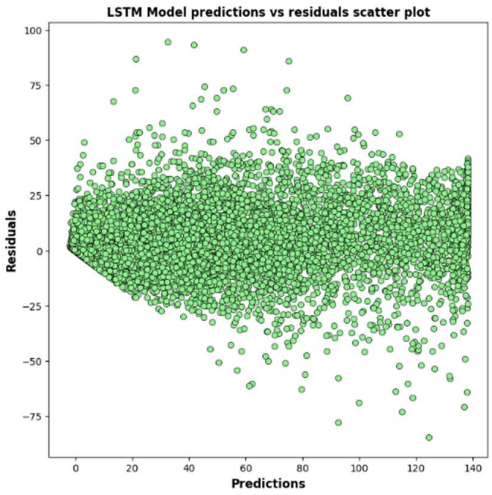

Figure 41. Prediction vs residuals scatter plot of LSTM predictor for wind energy

The r^2 values of the three wind farms power output forecasting using LSTM for the time lags of 2, 3, 4, 5, 6 hours is shown in Table 10.

Table 10. Average r^2 values of wind farms power output forecast models using LSTM model

Time lag	Amarnath	Erieau	Zurich	Average for all plants
2	0.912964464	0.881694328	0.889999976	0.899405808
3	0.907628366	0.862787343	0.884309326	0.884908345
4	0.901089754	0.880293224	0.883776779	0.888386586
5	0.8897349	0.864594235	0.88877164	0.881033591
6	0.891515985	0.887902419	0.874009774	0.884476059
Average for each plant	**0.902649656**	**0.87545431**	**0.884173499**	**0.888377311**

5. Solar PV farm hourly energy output prediction results

5.1 Multiple Linear Regression

Multiple linear regression models were applied to forecast power output across various time horizons, spanning from 0 to 6 hours in advance, using hourly power generation data and corresponding meteorological attributes from two solar PV facilities situated at Southgate and Windsor Airport. For each solar PV facility and specific time horizon, the prediction model underwent several iterations with distinct test-train splits (70/30, 80/20, 65/35, 50/50). The

model's performance evaluation for each solar PV facility and time lag was based on the average performance across these four splits. In Figure 42, the combined R-squared (r^2) values are displayed for the multiple linear regression forecasting models encompassing two solar PV facilities across various time horizons. Notably, the model's predictive efficacy is less optimal when exclusively relying on current-hour weather data (i.e., 0 hours ahead), yielding an average r^2 value of 0.20. However, when the data from the previous hour is incorporated into the analysis, the model exhibits substantial improvement, achieving an average r^2 value of 0.89 for all time horizons, as depicted in Figure 42. The enhanced predictive performance arising from the inclusion of the prior hour's data may potentially lead to misinterpretation when assessing the model's overall performance. Consequently, the model is subject to further scrutiny below to ascertain its susceptibility to underfitting and overfitting.

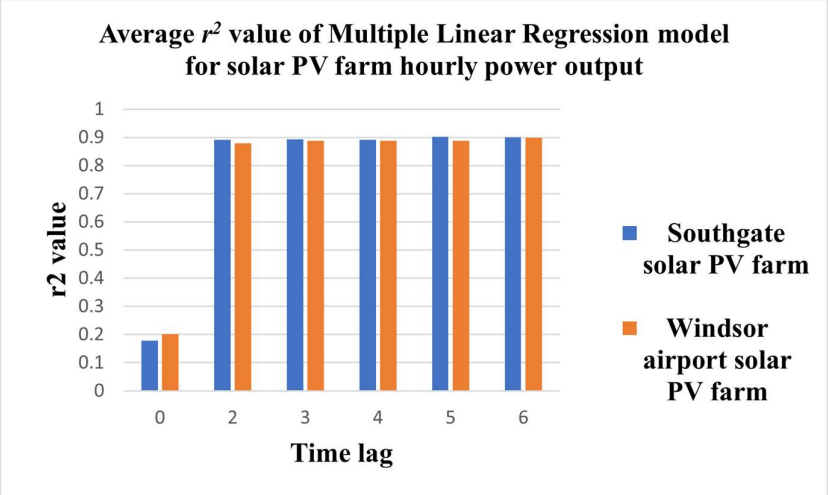

Figure 42. Multiple linear regression model's average r^2 value for solar PV power output forecasting

The scatter plot in Figure 43 illustrates the actual power output of the Southgate PV farm and the five-hour ahead power forecast using multiple linear regression. It reveals a distribution of points with a clear linear trend, where most points are closely aligned with the line of fit and only a relatively low number of outlier points, as displayed in Figure 43. This pattern suggests a good indication of predictivity for the multiple linear regression model, a conclusion further supported by the model's higher R-squared (r^2) value of 0.90 and the lower Mean Absolute Error (MAE) value of 9.06.

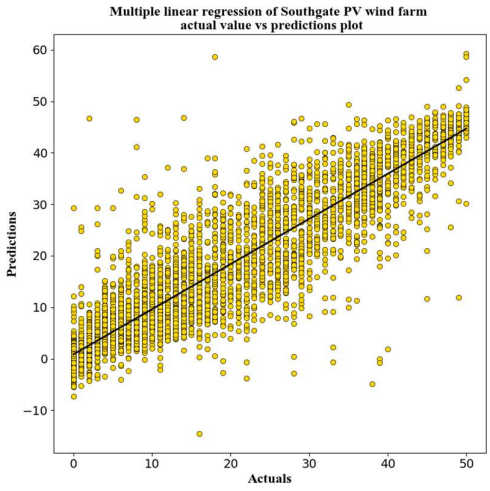

Figure 43. Multiple linear regression actuals vs predictions scatterplot for solar PV farm at Southgate

The distribution plot of residuals from the five-hour ahead power forecast multiple linear regression model for the Southgate PV farm fits a normal distribution curve, as illustrated in Figure 44. This distribution shows a nearly zero mean of residuals and a standard deviation of 1, indicating low variance in residuals. The narrow, normally shaped residual distribution without skewness suggests good predictivity of the model without overfitting. Conversely, if the residual curve is wider, it implies that the model's performance is lower, as the variance of the residuals is greater.

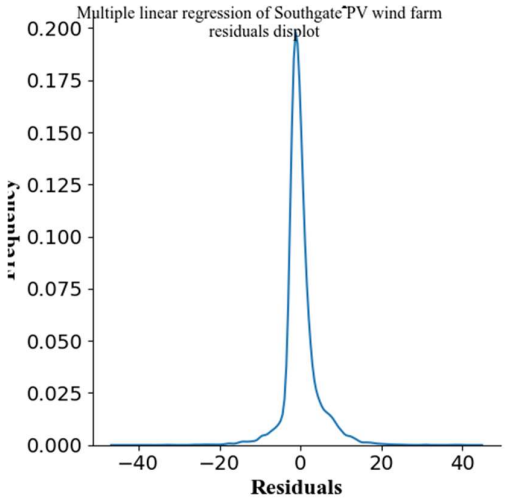

Figure 44. Residual plot of Solar PV power output forecast using Multiple Linear Regression

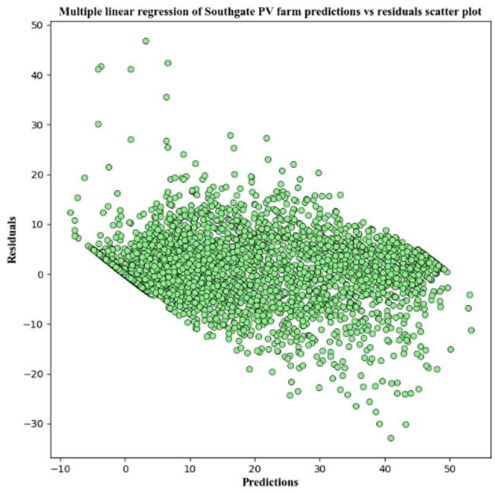

Figure 45. Multiple Linear Regression predictions vs residuals plot

Table 11. Average r^2 values of solar PV farms power output forecast models using Multiple Linear Regression

Time lag	Southgate solar PV farm	Windsor airport solar PV farm	Average for all plants
0	0.177914135	0.20162793	0.189771033
2	0.890631739	0.878579434	0.884605587
3	0.893017948	0.888153615	0.890585782
4	0.892237831	0.887454016	0.889845924
5	0.901957476	0.888386042	0.895171759
6	0.900251755	0.898260398	0.899256077
Average for each plant	**0.776001814**	**0.773743573**	**0.774872693**

5.2 MLPRegressor

MLPRegressor models for hourly power output forecasting were developed for two solar PV farms under study using the datum as discussed in earlier section. The models were designed to predict power output at various time horizons, and the inclusion of weather data from the respective previous hours depended on the specific forecast horizon. Similar to the multiple linear regression models discussed earlier, each forecasting model for a particular wind farm and time horizon underwent training with multiple test-train splits, including scenarios of

70/30, 80/20, 65/35, and 50/50. The performance results were averaged across these splits to determine the overall model performance.

The overall averaged r^2 values for power output forecasting using MLPRegressor models for the three wind farms at various time horizons are presented in Figure 46. Notably, the MLPRegressor models exhibit enhanced predictivity when forecasting the power output for the current hour by utilizing weather data from the same hour. This results in a notably improved average r^2 value of 0.75 compared to the multiple linear regression model discussed earlier. In adherence, there is also increase in r^2 values for the MLPRegressor models when predicting power output at solar PV farms for different time horizons, specifically in the range of 2 to 6 hours ahead. The average r^2 value for this range is 0.92, as depicted in Figure 46. The higher values of r^2 must be further examined to determine if there is any overfitting of data or the model's capacity in capturing the data accurately.

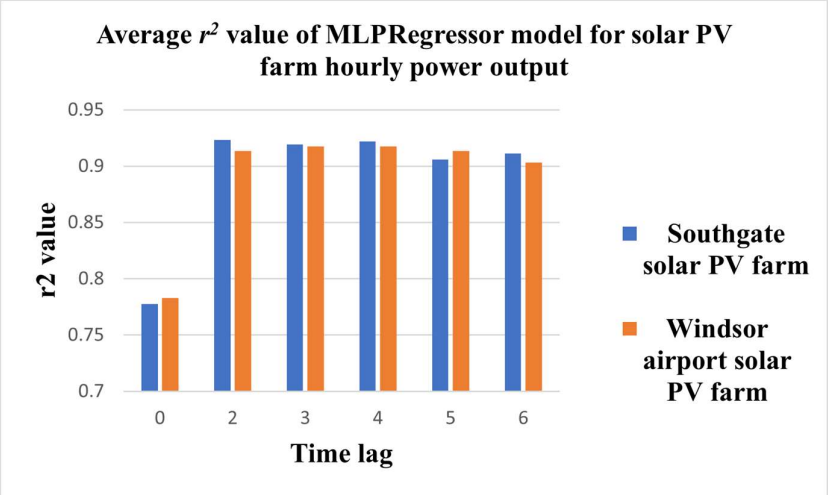

Figure 46. MLPRegressor model's average r^2 value for solar PV power output forecasting

The scatterplot in Figure 47, illustrates the actual power output of the Southgate solar PV farm alongside the three-hour ahead power output forecasts generated using the Multi-Layer Perceptron Regressor. The data points on the plot closely align with the line of best fit, with some points positioned above and below the line, reflecting prediction errors. The presence of

outliers on both sides of the line of best fit suggests that the model excels in some instances while performing less effectively in others.

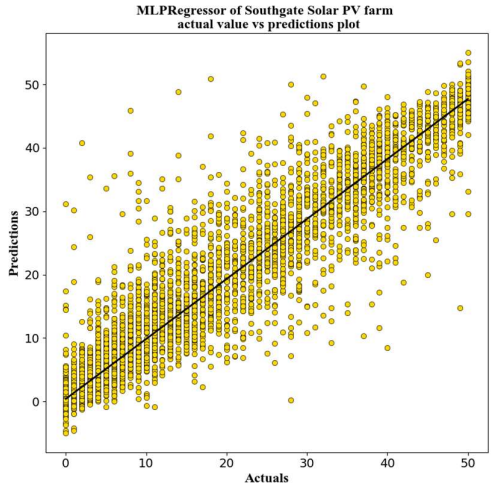

Figure 47. MLPRegressor actuals vs predictions scatterplot for solar PV farm at Southgate

The residuals distribution plot of the MLPRegressor model for forecasting the Southgate solar PV farm power output three hours ahead is shown in Figure 48. The distribution plot fits a normal curve with almost zero mean and standard deviation 1. The plot reveals minimal skewness in the residuals which suggest the good predictivity of the model. However, the predictions vs residuals plot must be accessed to determine the variance of the residuals with respect to predictions before concluding about the overall model performance.

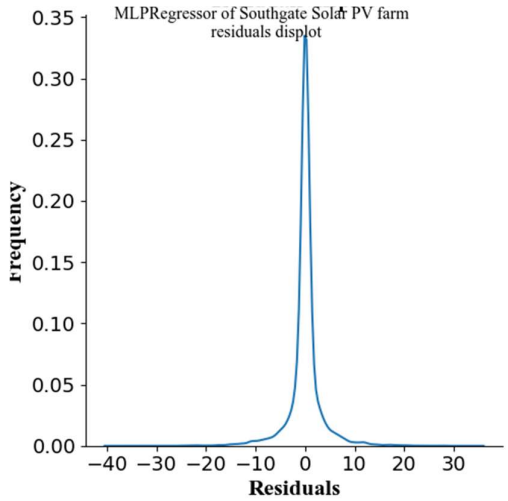

Figure 48. Residual plot of Solar PV power output forecast using MLPRegressor

The examination of the prediction and residual plot of the Southgate solar PV farm power output forecasting model using MLPRegressor model shown in Figure 49, reveals a downward tunnel pattern as observed in multiple linear regression's, prediction and residual plot. The downward tunnel pattern with most points clustered around 0, shows most points are exactly predicted, with some points above in the left side with predictions having more residuals (prediction higher than actual) and some points lower in the right side with predictions having less residuals (predictions less than actuals). This higher values of residuals and lower value of residuals in Figure 40 , reveals higher variance in residuals and the residuals are skewed. Therefore, the higher value of r^2 is due to overfitting of the data rather than the model capturing the actual behaviour of the pattern of the variables.

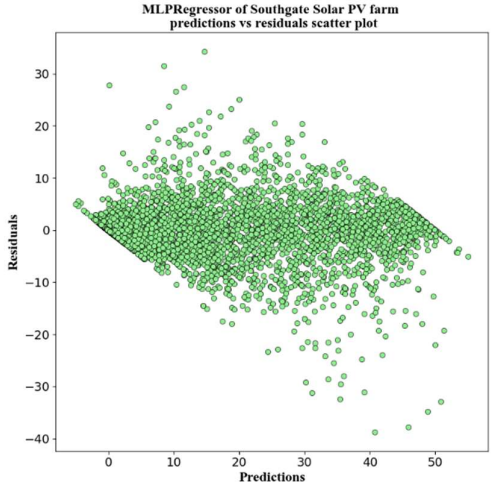

Figure 49. MLPRegressor predictions vs residuals plot

The r^2 values of the hourly power output forecasting models for two solar PV farms using MLPRegressor for the time lags of 0, 2, 3, 4, 5, 6 hours is shown in Table 12.

Table 12. Average r^2 values of solar PV farms power output forecast models using multiple MLPRegressor

Time lag	Southgate solar PV farm	Windsor airport solar PV farm	Average for all plants
0	0.777759777	0.782792406	0.780276092
2	0.923136194	0.913530461	0.918333328
3	0.919131843	0.917303789	0.918217816
4	0.921901925	0.917420457	0.919661191
5	0.906035983	0.913628152	0.909832068
6	0.911136495	0.903436528	0.907286512
Average for each plant	**0.893183703**	**0.891351966**	**0.892267834**

5.3 Random Forest regressor

The hourly power output forecasting models of solar PV farms across different time horizons were trained utilizing the ensemble learning method - random forest regressor. The averaged r^2 values for these forecasts, covering timeframes from 0 to 6 hours ahead, are graphically depicted in Figure 50. Notably, the ensemble learning technique demonstrates a substantial increase in predictive accuracy, with an r^2 value of 0.82, compared to both the multiple linear regression and MLPRegressor models when forecasting the current-hour power output. These models excel in effectively predicting the energy output of the solar PV farms, achieving an impressive average r^2 value of 0.92, as visualized in Figure 50. Further scrutiny is applied to investigate the presence of overfitting within the model's output.

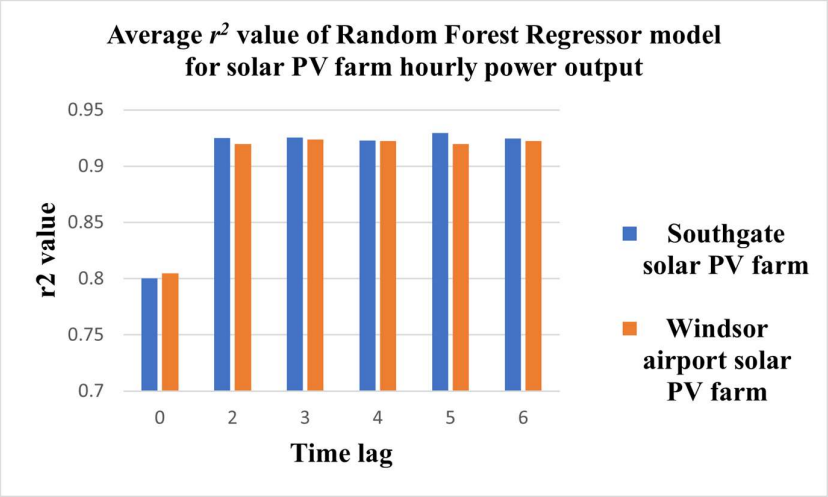

Figure 50. Random Forest Regressor model's average r^2 value for solar PV power output forecasting

The scatterplot of the hourly solar PV energy output and the predictions for the random forest regressor model is shown in Figure 51. The scatter plot shows a linear trend with the predictions almost equalling the actual values. The point aligns themselves toward the line of best fit, with some points above and below the line. The Figure shows that all the predictions were between 0 to 50, which is well within the range of the wind farm capacity (0 to 50 MW capacity). That is either the solar PV farm produced no output nor the energy output was maximum not exceeding the maximum.

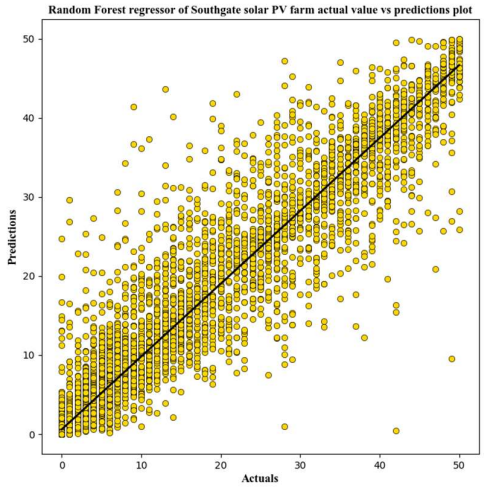

Figure 51. Random Forest Regressor actuals vs predictions scatterplot for solar PV farm at Southgate

These difference in residuals is examined in the residual plot shown in Figure 52. The normal cure, 0 mean and 1 standard deviation of the residual distribution curve shows that the residuals are normally distributed and the model predictivity is good.

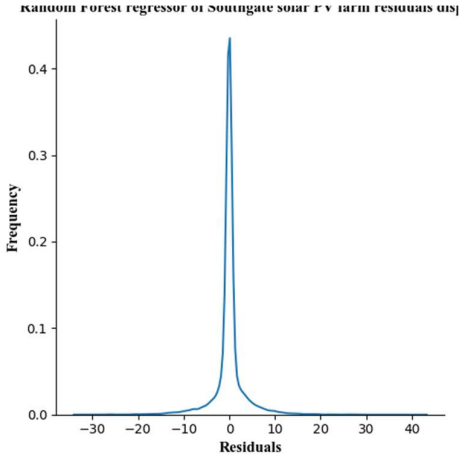

Figure 52. Residual plot of Solar PV power output forecast using Random Forest Regressor

The predictions vs the residuals plot is illustrated in Figure 53, examining the plot helps in determining the presence of heteroscedastic, outliers and patterns in the model. The points are distributed in a random pattern centred around 0 rather than like a tunnel pattern observed in multiple linear regression and MLPRegressor. This conforms that the model is free of heteroscedastic, i.e., variance of residuals is low. But they are still outliers in the model with points well above and below the graph. This conforms that Random Forest Regressor is one of the good fits for predicting the solar PV farms energy output with an average r^2 value of 0.91.

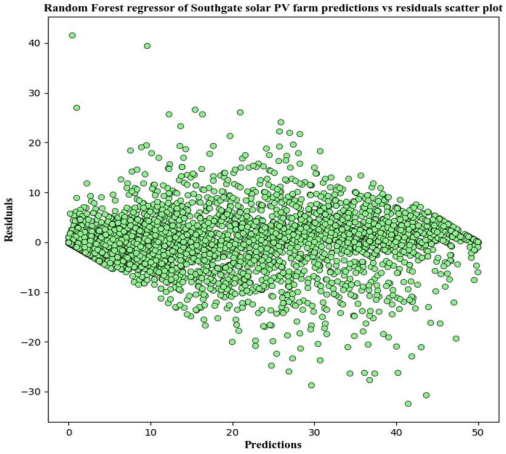

Figure 53. Random Forest Regressor predictions vs residuals plot

The r^2 values of the two solar PV farms power output forecasting using random forest regressor for the time lags of 0, 2, 3, 4, 5, 6 hours is shown in Table 13.

Table 13. Average r^2 values of solar PV farms power output forecast models using Random Forest Regressor

Time lag	Southgate solar PV farm	Windsor airport solar PV farm	Average for all plants
0	0.800431727	0.804802018	0.802616873
2	0.924902613	0.919719666	0.92231114
3	0.925492699	0.923705702	0.924599201
4	0.922824977	0.922188504	0.922506741
5	0.929504283	0.919581857	0.92454307
6	0.924500585	0.922588545	0.923544565
Average for each plant	**0.904609481**	**0.902097715**	**0.903353598**

5.4 Support Vector Regression

The hourly solar PV energy output forecasting model for various time horizons ahead was developed using Support vector regression (SVR) algorithm. The average r^2 value of the hourly power output SVR model for the two solar PV farms is shown in Figure 55. The overall model r^2 value is around .89. The scatter plots of actuals and predictions, distribution plot of residuals, residuals and predictions was examined as shown in Figure 55, 56, 57, respectively.

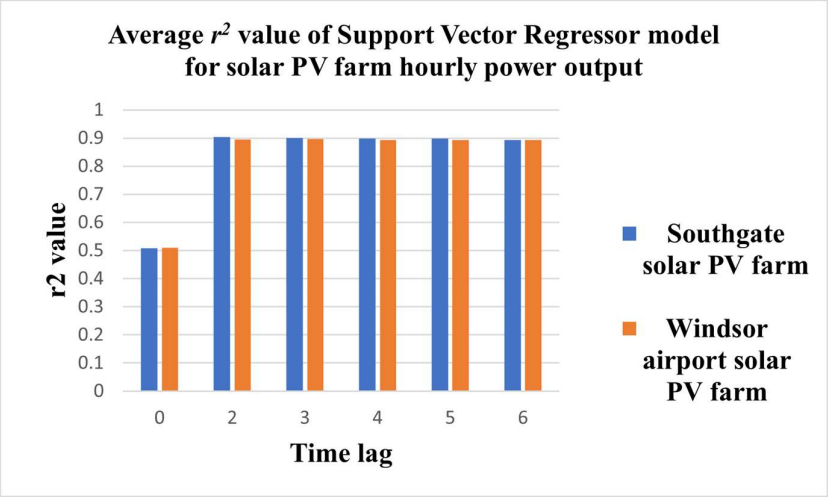

Figure 54. Support Vector Regression model's average r^2 value for solar PV power output forecasting

The scatterplot of actual hourly energy output of Southgate solar PV farm and the SVR model predicted power output 3 hours ahead is shown in Figure 55. The points align themselves toward the line of best fit, but all the points are slightly away from the line of best fit, i.e., for an actual value predicted value is not accurate and is subjected to error.

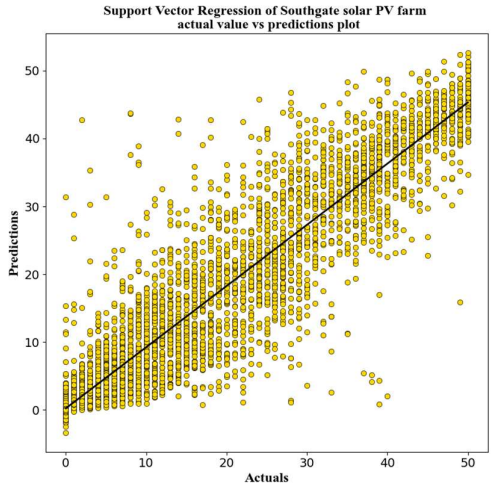

Figure 55. Support Vector regression actuals vs predictions scatterplot for solar PV farm at Southgate

The residual distribution plot (a way to visualize the error in predictions) shown in Figure 56, exhibits a normal curve with wide area under the curve, signifying greater error in predictions, which the higher MAE value confirms.

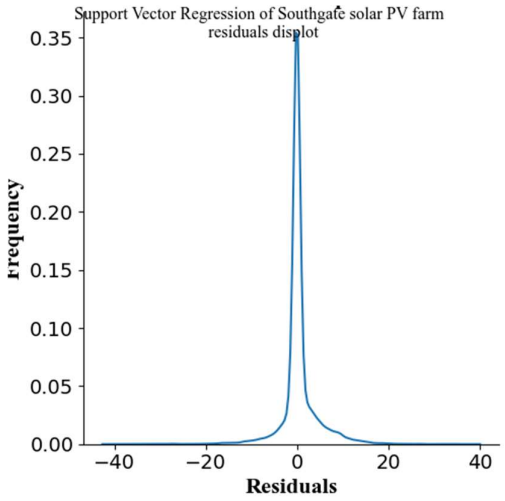

Figure 56. Residual plot of Solar PV power output forecast using Support Vector Regression

The predictions vs residuals plot shown in Figure 57, exhibits a tunnel shape, which confirms heteroscedastic in the predictions. The points clustered around 0 with outliers in both the sides shows that the residuals are more skewed and the model doesn't capture the actual behaviour of the input datum.

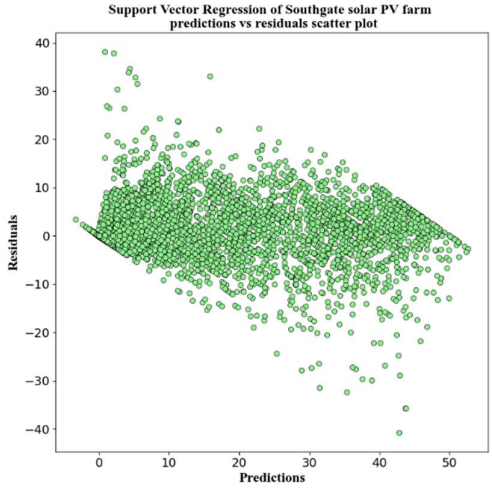

Figure 57. Support Vector Regression predictions vs residuals plot

The r^2 values of the two solar PV farms power output forecasting using support vector regression for the time lags of 0, 2, 3, 4, 5, 6 hours are shown in Table 14.

Table 14 . Average r^2 values of solar PV farms power output forecast models using Support Vector regression model

Time lag	Southgate solar PV farm	Windsor airport solar PV farm	Average for all plants
0	0.507238145	0.510725614	0.50898188
2	0.903209397	0.894988056	0.899098727
3	0.900660399	0.896781748	0.898721074
4	0.898265933	0.892360593	0.895313263
5	0.899245637	0.892257854	0.895751746
6	0.893038741	0.893970528	0.893504635
Average for each plant	**0.833609709**	**0.830180732**	**0.83189522**

5.5 Long-Short Term Memory

The recurrent neural network model LSTM was used in forecasting the hourly power output of solar PV farms for various time horizons ahead. The average r^2 values of deep learning LSTM model trained for forecasting the power output of two solar PV farms on various time horizons ahead is shown in Figure 58. The figure illustrates that the LSTM model has an average r^2 value of 0.92 for predicting the output of all wind farms and for all the time horizons ahead. The analysis of scatterplots of actual values and predictions, residuals distribution plots, predictions and residuals scatterplot are Figures respectively, provides a deeper intuition towards the model's performance.

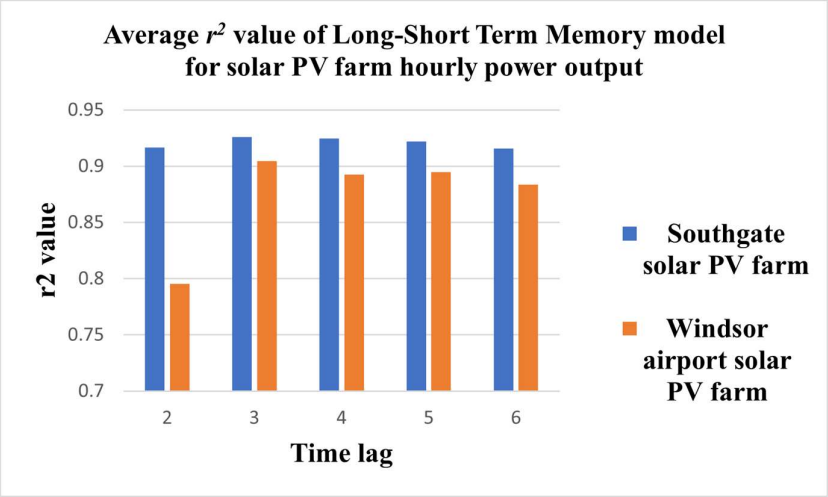

Figure 58. LSTM model's average r^2 value for solar PV power output forecasting

The scatterplot of the actual power output and the LSTM model predicted power output for three hours ahead time lag is showing a linear trend, with a smaller number of outliers as shown in Figure 59. The residual plot distribution plot shown in Figure 60, demonstrates a normal curve with zero mean and 1 standard deviation shows that the residuals are not skewed and properly distributed. The residual distribution plot show that the good predictability of the model.

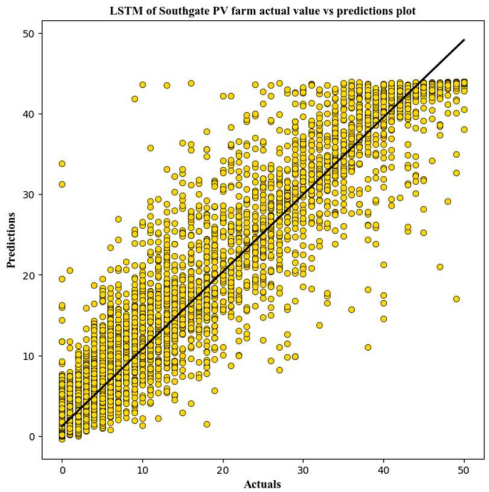

Figure 59. LSTM actuals vs predictions scatterplot for solar PV farm at Southgate

The scatter plot of predictions and the residuals of the LSTM model shown in Figure 61, depicts a random pattern rather than a tunnel pattern as observed in other models. The random pattern with most points clustered above and below zero shows that the LSTM models were free of heteroscedastic and the residuals was not skewed and the variance of the residuals was

low. The model has a very good predictability with higher r^2 value and lower MAE. The predictions and residuals plots show that the model captures the behaviour of the input datum rather than overfitting the input data.

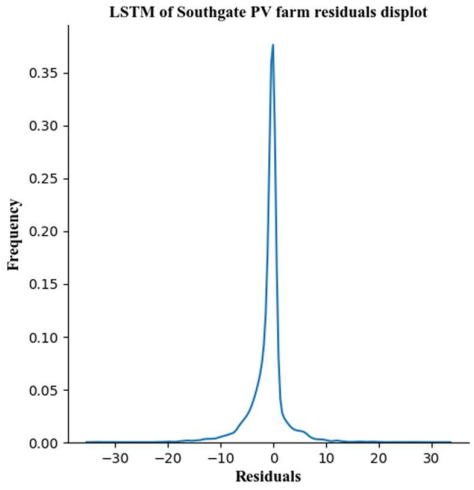

Figure 60. Residual plot of Solar PV power output forecast using LSTM

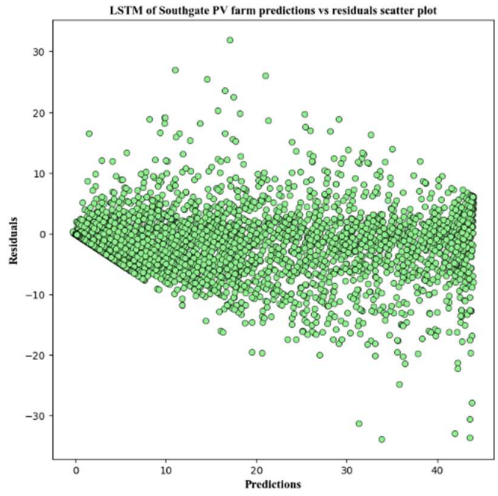

Figure 61. LSTM predictions vs residuals plot

The r^2 values of the two solar PV farms power output forecasting using LSTM for the time lags of 2, 3, 4, 5, 6 hours is shown in Table 15.

Table 15. Average r^2 values of solar PV farms power output forecast models using LSTM

Time lag	Southgate solar PV farm	Windsor airport solar PV farm	Average for all plants
2	0.916705173	0.795206238	0.855955706
3	0.925800748	0.904525485	0.915163117
4	0.924432903	0.892609338	0.908521121
5	0.921828055	0.894874289	0.908351172
6	0.915780615	0.883629667	0.899705141
Average for each plant	**0.920909499**	**0.874169003**	**0.897539251**

6. Result and Discussion

The average execution time of each algorithm for predicting hourly wind energy output at various time horizons (2-6 hours ahead) and the average R^2 value for the corresponding model are shown in Figure 53. The execution time exhibits the same trend regardless of the underlying algorithm used for training. The execution time increases from the 2-hour-ahead prediction model to the 6-hour-ahead wind energy prediction models using all algorithms. The anomalies in the execution time curve of all the algorithms are due to other system processes. However, the overall pattern shows that execution time increases from 2 to 6 hours ahead in forecasting as the input size increases.

On the other hand, the average R^2 curve of each algorithm demonstrates different patterns. The R^2 curve of LSTM constantly increases from 2 hours ahead to 6 hours ahead, as shown in Figure 62. The LSTM model is a Recurrent Neural Network (RNN) that stores information from previous data points, and therefore, the greater the number of hours in the input, the higher the accuracy of the model.

On the other hand, the flat SVR and MLPRegressor have a higher R^2 value for lower time lags ahead (e.g., 2 hours ahead), and the R^2 value gradually decreases with the increase in the number of hours ahead. This trend in SVR and MLPRegressor is due to the incapability of the model to store the previous hour's information. Simply adding the previous hour's data to the current hour's data increases noise in the input feature space, resulting in a lower R^2 value.

The Random Forest regressor as discussed in the former section, has higher r^2 value for all time lags because of overfitting of the data. Therefore, the LSTM model is best model with respect to the execution time.

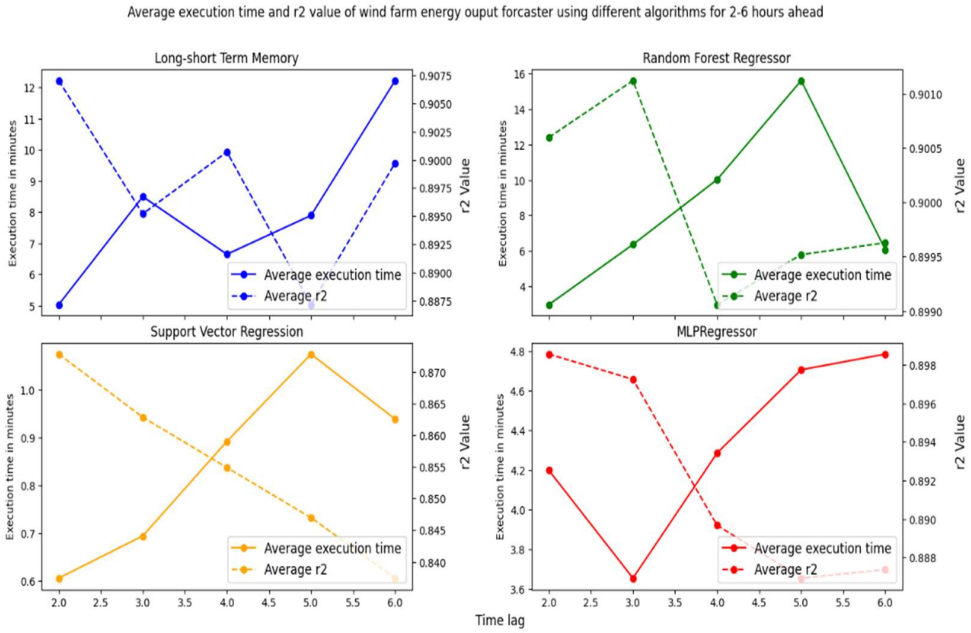

Figure 62. Average execution time and r^2 value of algorithms for 2-6 hours ahead prediction

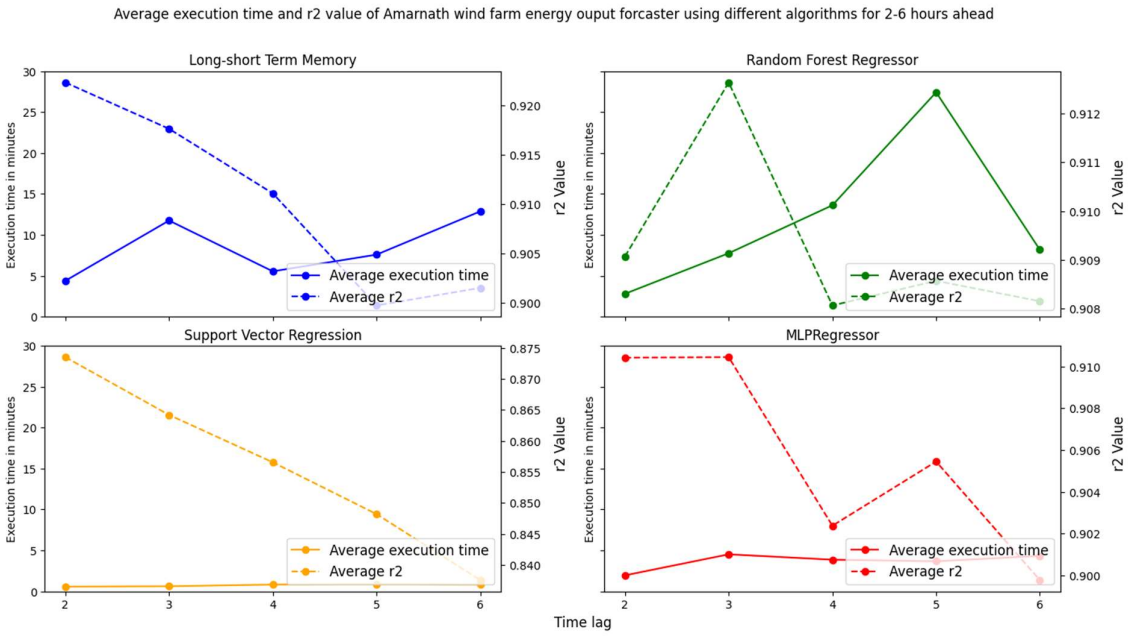

Figure 63. Average execution time and r2 value of algorithms for 2-6 hours ahead prediction

The average execution time of each algorithm for predicting hourly solar PV energy output at various time horizons (2-6 hours ahead) and the average r^2 value for the corresponding model are shown in Figure 64. The execution time exhibits the same trend regardless of the underlying algorithm used for training. The execution time increases from the 2-hour-ahead prediction model to the 6-hour-ahead wind energy prediction models using all algorithms. The anomalies in the execution time curve of all the algorithms are due to other system processes. However, the overall pattern shows that execution time increases from 2 to 6 hours ahead in forecasting as the input size increases.

On the other hand, the average R^2 curve of each algorithm demonstrates different patterns. The r^2 curve of LSTM constantly increases from 2 hours ahead to 6 hours ahead, as shown in Figure 64. The LSTM model is a Recurrent Neural Network (RNN) that stores information from previous data points, and therefore, the greater the number of hours in the input, the higher the accuracy of the model.

On the other hand, the flat SVR and MLPRegressor have a higher R^2 value for lower time lags ahead (e.g., 2 hours ahead), and the r^2 value gradually decreases with the increase in the number of hours ahead. This trend in SVR and MLPRegressor is due to the incapability of the model to store the previous hour's information. Simply adding the previous hour's data to the current hour's data increases noise in the input feature space, resulting in a lower r^2 value.

The Random Forest regressor as discussed in the former section, has higher r^2 value for all time lags because of overfitting of the data. Therefore, the LSTM model is best model with respect to the execution time.

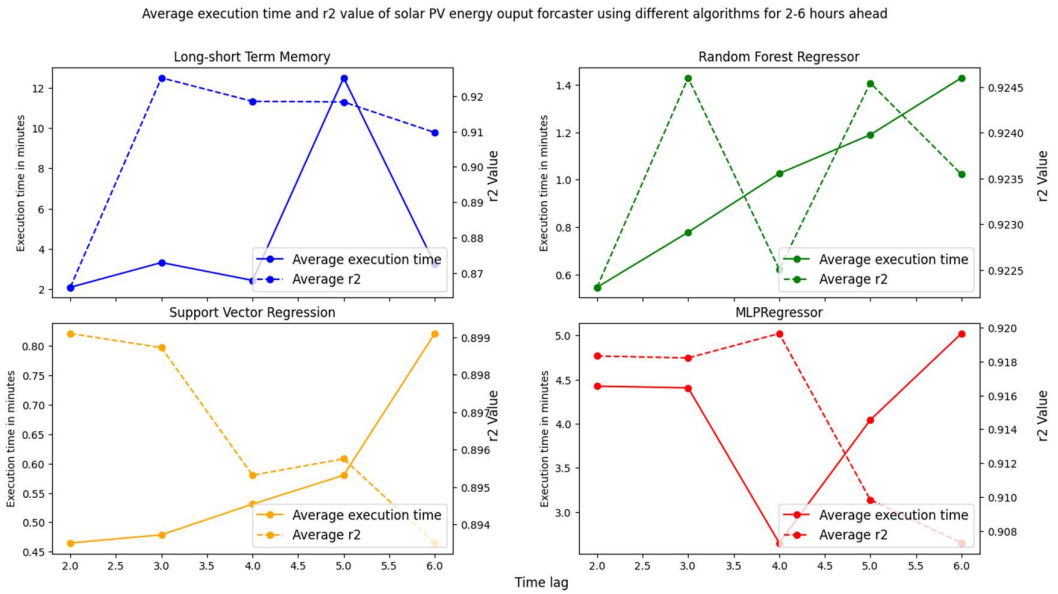

Figure 64. Average execution time and r^2 value of algorithms for 2-6 hours ahead prediction of solar PV farms

In the same way the anomalies in the execution time and the average r2 value is further investigated by using the result plots from individual solar PV farm at South gate as shown in Figure 65. The results of single power plant shows that the accuracy decreases with the increase in time lag and the execution time exhibits reverse trend. In that also there are slight anomalies in the execution time which is due to the system performance at the time of training the model

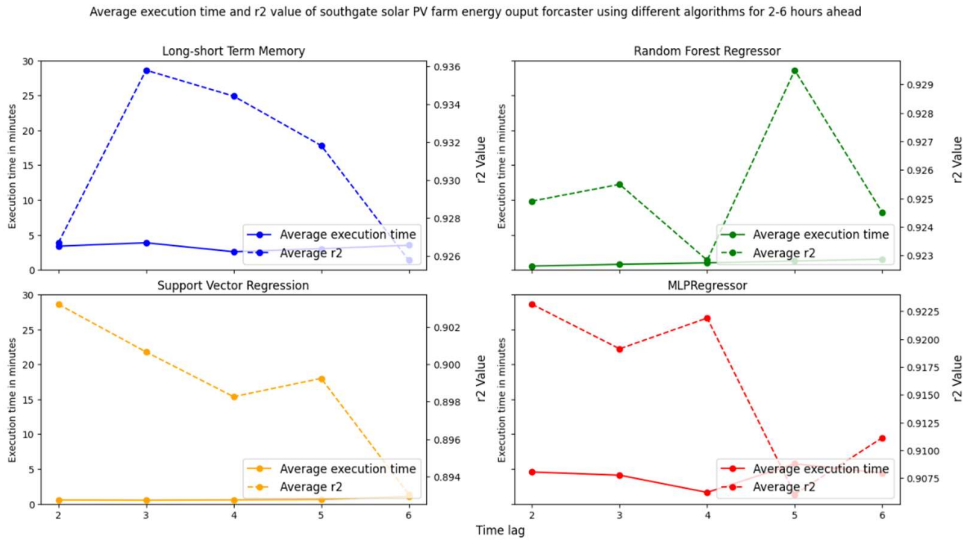

Figure 65. Average execution time and r2 value of algorithms for 2-6 hours ahead prediction of solar PV farms

6.2 Anomalies in execution time and accuracy

The results of the execution time and accuracy of wind farm and solar PV power prediction models discussed in previous section has some anomalies as shown in Figure 64, 65. The actual expected behaviour of the models were when the hours ahead prediction increases from 1 hour ahead to 6 hours ahead the accuracy decreases and the execution time increases as the number of fields increases. But for all the algorithms there were few anomalies i.e for smaller time lag say 2 hours ahead prediction the execution time was low when comparing with the 1 hour ahead forecasting and vice-versa. On further examination these anomalies were due to the Eriaue wind farm load curtailment, due to which the load output could not be predicted exactly. The output of predictions from individual power plant at Amarnath wind farm, Southgate solar PV farm is shown in Figure 63, 65. The results of single power plant shows that the accuracy decreases with the increase in time lag and the execution time exhibits reverse trend. In that also there are slight anomalies in the execution time which is due to the system performance at the time of training the model.

7. Conclusion

In conclusion, the hourly forecasting of wind farm and solar PV farm energy output was more accurately and reliably predicted using the model trained using deep learning recurrent neural network long-short term memory model. The forecasting model using LSTM algorithm provided good accuracy in predicting power output between 2 and 6 hours ahead with an average r^2 value of 0.90 and average execution time of 6 minutes. The models trained using random forest regressor exhibited good r^2 value, but it was due to overfitting of the data. The SVR and MLPRegressor predicts the power output accurately if it's forecasting for current hour, when it's predicted between 2 to 6 hours the r^2 value gradually decreases. The execution time could be decreased by incorporation parallel processing algorithm which will reduce the load on single machine and distribute the processing load to multiple systems when required thereby resulting in near real time prediction with atmost accuracy.

8. Future work and limitations

The current machine learning model will be deployed on a cloud server running on Apache Spark system. Historical hourly power data will be obtained for all power plant locations in Ontario through IESO and hourly weather data will be obtained from NASA API. Once the model developed and trained for a particular location the model will be deployed to production system. On real time hourly weather forecast will be obtained from OpenWeather map API.

The real time data will be used to predict the real time energy output. The architecture of th e of the proposed solution is shown in Figure 66.

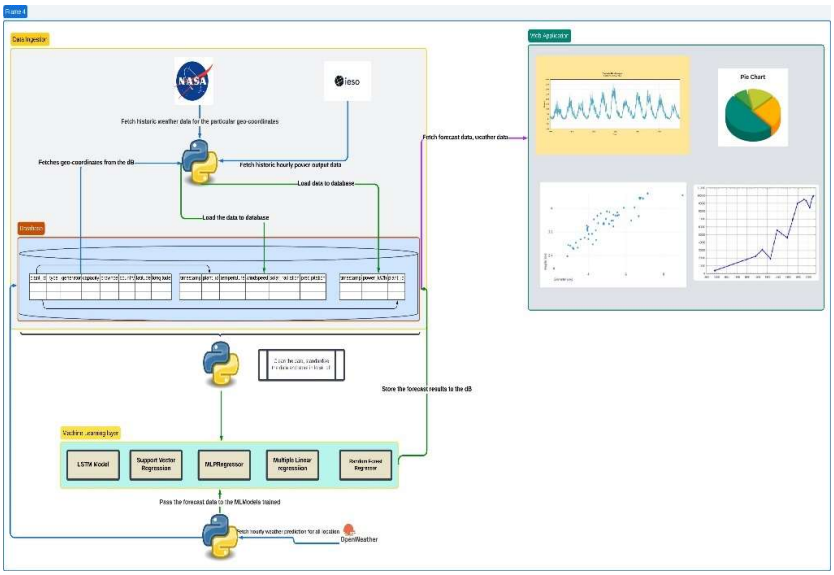

Figure 66. ML power output prediction application architecture

8. APPENDIX – A | GITHUB LINK

https://github.com/Jayasuriya4372175999/OneEnergyPredictor

9. References

[1] G. Demirezen and A. S. Fung, "Modelling and assessment of cloud based smart dual fuel switching system (sdfss) of residential hybrid hvac system for simultaneous reduction of energy cost and greenhouse 2 gas emission under smart grid infrastructure 3," eSimPapers, IBPSA.

http://www.ibpsa.org/proceedings/eSimPapers/2021/Contribution_1175_final_a.pdf

[2] G. Demirezen, K. Ullah, A. Rokn, and A. Fung, "Economical and environmental data analysis of hybrid hvac system of air source heat pump and natural gas furnace for cold climate - Canada," 13th IEA Heat Pump Conference 2020, 2020. (April 26-29, 2021 Jeju, Korea)

[3] D. Yu, K. Y. Tung, N. Ekrami, G. Demirezen, A. Fung, F. Mohammadi, and K. Raahemifar, "Proof of concept of a cloud-based smart dual-fuel switching system to control the operation of a hybrid residential hvac system," in 2019 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), pp. 1–5, 2019 (1-4 Dec. 2019, Macao ,China).

[4] A. M. Brockway and P. Delforge, "Emissions reduction potential from electric heat pumps in california homes," The Electricity Journal, vol. 31, no. 9, pp. 44–53, 2018.

[5] G. Demirezen, N. Ekrami, and A. Fung, "Monitoring and evaluation of nearly-zero energy house (nzeh) with hybrid hvac for cold climate – canada," IOP Conference Series: Materials Science and Engineering, vol. 609, p. 062001, Sep 2019.

[6] G. Demirezen, A. S. Fung, and M. Deprez, "Development and optimization of artificial neural network algorithms for the prediction of building specific local temperature for hvac control," International Journal of Energy Research, vol. 44, no. 11, pp. 8513–8531, 2020.

[7] K. Ye, G. Demirezen, A. Fung, and E. Janssen, “The use of artificial neural networks (ann) in the prediction of energy consumption of air-source heat pump in retrofit residential housing,” IOP Conference Series: Earth and Environmental Science, vol. 463, p. 012165, mar 2020.

[8] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, “Machine learning methods for solar radiation forecasting: A review,” Renewable energy, vol. 105, pp. 569–582, 2017.

[9] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, “Forecasting power output of photovoltaic systems based on weather classification and support vector machines,” IEEE Transactions on Industry Applications, vol. 48, no. 3, pp. 1064–1069, 2012.

[10] I. Gherboudj and H. Ghedira, “Assessment of solar energy potential over the united arab emirates using remote sensing and weather forecast data,” Renewable and Sustainable Energy Reviews, vol. 55, pp. 1210–1224, 2016.

[11] C. Chen, S. Duan, T. Cai, and B. Liu, “Online 24-h solar power forecasting based on weather type classification using artificial neural network,” Solar energy, vol. 85, no. 11, pp. 2856–2870, 2011.

[12] X. Qing and Y. Niu, “Hourly day-ahead solar irradiance prediction using weather forecasts by lstm,” Energy, vol. 148, pp. 461–468, 2018.

[13] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, “Predicting solar generation from weather forecasts using machine learning,” in 2011 IEEE international conference on smart grid communications (SmartGridComm), pp. 528–533, IEEE, 2011 (17-20 October 2011, Brussels, Belgium).

[14] H. Wang, Y. Li, M. Xiong, and H. Chen, “A combined wind speed prediction model based on data processing, multi-objective optimization and machine learning,” Energy Reports, vol. 9, pp. 413–421, 2023. 2022 The 3rd International Conference on Power Engineering (December 09–11, Sanya, Hainan Province, China).

[15] J. Heinermann and O. Kramer, “Machine learning ensembles for wind power prediction,” Renewable Energy, vol. 89, pp. 671–679, 2016.

[16] C. Cakiroglu, S. Demir, M. Hakan Ozdemir, B. Latif Aylak, G. Sariisik, and L. Abualigah, “Data-driven interpretable ensemble learning methods for the prediction of wind turbine power incorporating shap analysis,” Expert Systems with Applications, vol. 237, p. 121464, 2024.

[17] S. T. Ayele, M. B. Ageze, M. A. Zeleke, and T. A. Miliket, “Adama ii wind farm long-term power generation forecasting based on machine learning models,” Scientific African, vol. 21, p. e01831, 2023

[18] Independent Electricity System Operator (IESO), "Data Directory," [Online]. Available: https://www.ieso.ca/en/Power-Data/Data-Directory [Accessed: 12 10, 2022].

[19]"NASA POWER," NASA Langley Research Center, [Online]. Available: https://power.larc.nasa.gov/ [Accessed: 12 10, 2022].

[20] TransAlta Corporation, "Melancthon Wind Farm," TransAlta, [Online]. Available: https://transalta.com/about-us/our-operations/facilities/melancthon/ [Accessed: 12 10, 2022].

[21] "Wind Energy," ENGIE North America, [Online]. Available: https://www.engiena.com/wind/ [Accessed: 12 10, 2022].

[22] "Windsor Project," Samsung Renewable Energy, https://www.samsungrenewableenergy.ca/our-projects/windsor/ .