This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/UPEC61344.2024.10892390, 2024 59th International Universities Power Engineering Conference (UPEC)

Performance Analysis of Photovoltaic Installations Based on Machine Learning Techniques

Daniil Hulak
Department of Electronic and Electrical Engineering
Brunel University London
London, UK
Daniil.Hulak@brunel.ac.uk

Yang Xie
School of Cyber Science and Engineering
Southeast University
Nanjing China
Yang.Xie@brunel.ac.uk

Gareth Taylor
Department of Electronic and Electrical Engineering
Brunel University London
London, UK
Gareth.Taylor@brunel.ac.uk

Abstract—**This paper investigates the novel performance analysis of photovoltaic (PV) installations by applying machine learning techniques (ML). The data used for the research is a mixed dataset of data from an experimental PV installation located at Brunel University London with correspondingly available weather data. Firstly, the analysis aims to establish various sensitivity relationships between PV power generation and weather conditions using techniques based on Random Forest ML. Secondly, the processing stage is implemented to assess the fitness of the different ML techniques through cross-validation. The results highlight the differing effectiveness of the applied approaches in achieving accurate and reliable results for the PV installations. The best techniques offer valuable insights for optimizing renewable energy usage in diverse environmental conditions.**

Index Terms—**Generation Forecasting, Machine Learning, Photovoltaic Installation, Solar Irradiance**

I. Introduction

In recent years, the global energy profile has changed rapidly, shifting towards the higher penetration of renewable energy sources (RES). Among these, photovoltaic (PV) installations have emerged as one of the key players, contributing significantly to the power generation mix. Increasing RES penetration aligns with the efforts to address climate change, reaching Net Zero and transitioning towards sustainable energy practices. PV power generation globally increased by a record 270 TWh, up 26% in 2022, reaching almost 1 300 TWh. It demonstrated the largest absolute power generation growth of all renewable technologies in 2022, surpassing wind for the first time in history [1].

The development of RES, particularly PV installations, creates a significant challenge to the reliable and efficient energy supply for the existing power systems. PV are heavily dependent on meteorological conditions, and predicting their power generation is a complex and dynamic task. As solar power production is linked to weather patterns, accurate forecasting becomes essential to ensure grid stability, optimal resource utilization, and effective energy management.

The rise of distributed generation in the form of residential PV installations also started to be a growing challenge for system operators. Unlike centralised power generators, equipped with advanced monitoring systems, individual households typically do not install additional weather monitors simultaneously with their PV installations. As a result, they concentrate data on their power generation, but not on the weather conditions. This data gap creates significant challenges to accurate forecasting of weather-dependent output from these distributed sources. The absence of real-time weather data complicates integrating these decentralised power resources into the overall power system, highlighting the need for innovative approaches.

In response to these challenges, this paper investigates the application of machine learning (ML) models for predicting power generation of distributed PV installations. Using the capabilities of ML, we aim to enhance the accuracy and reliability of generation forecasts, considering the complicated relationships between PV power output and meteorological parameters. ML`s ability to consider complex patterns and dependencies within huge datasets makes it a promising candidate for dealing with volatile patterns of RES generation behaviour. Accordingly, this paper analyses the performance of different ML algorithms for obtaining a suitable model for accurate prediction of the PV installation, demonstrating the differing effectiveness of the applied approaches.

II. Data Description

A. PV installation data and general characteristics

In this study, the dataset used for analysis originates from the power generation of an experimental PV installation situated on the top of the Joseph Lowe building, located at 51.5308°N, 0.4740°W at Brunel University`s London campus. The PV installation consists of 22 solar panels with a total nominal power rating of 7.5 kilowatts peak and serves as a case study for the investigations as presented in this research paper. The PV installation is located on a flat roof and occupies an area of around 300 m^2, 15x20 m. A photographic image of the PV installation and related configuration diagram are presented in Figures 1 and 2, respectively.

The system`s power generation data is received through advanced monitoring and data management tools, specifically the EnnexOS Data Manager M and Sunny Portal powered by EnnexOS [2]. The combination of this device and software provides both comprehensive real-time and historical insights

British Academy, CARA, NSFC

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/UPEC61344.2024.10892390, 2024 59th International Universities Power Engineering Conference (UPEC)

into the performance of the PV installation. The power generation information has 5-minute records that allow a detailed exploration of the PV installation power generation dynamics.

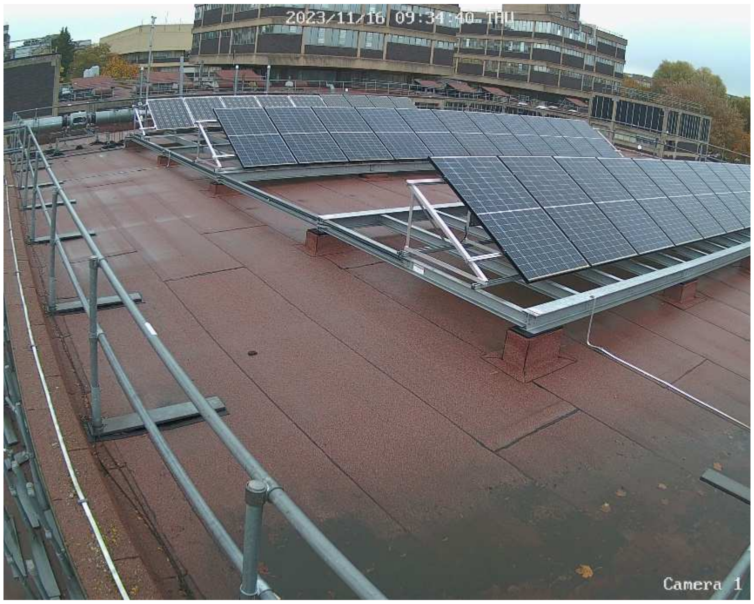

Figure 1. Photographic image of the PV installation located at Brunel University London

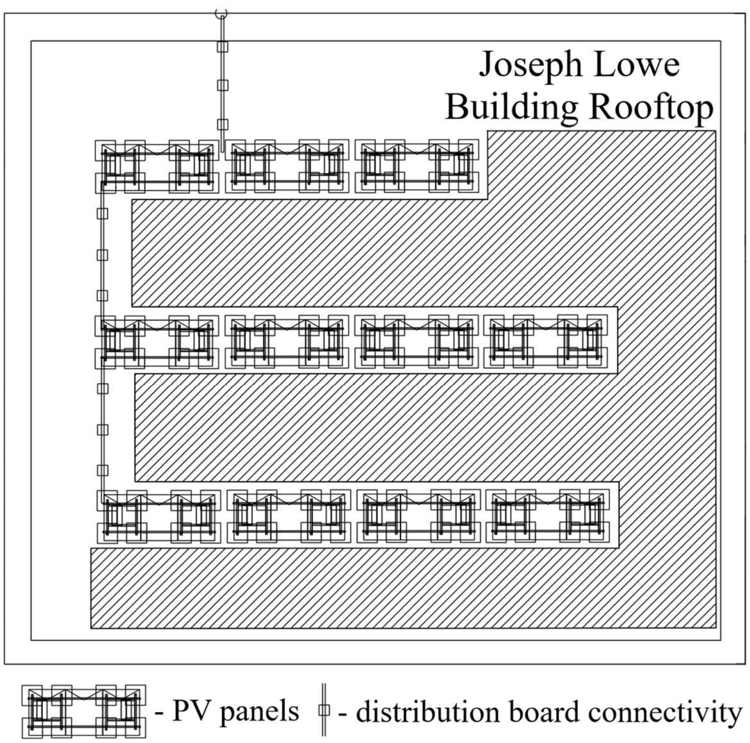

Figure 2. Configuration diagram of the PV installation located at Brunel University London

By focusing on a specific PV installation with defined characteristics and using monitoring technologies, our research aims to provide insights that can be then extrapolated on predictive capabilities of distributed PV power generation models across diverse settings and installed in different regions. By relying on generation data from the PV installation located at Brunel University London, the installation`s characteristics become representative of an average household customer`s system, and the data flow is considered equivalent across the scenario.

B. *Weather dataset*

The weather data correspondingly combined with the PV power generation dataset was obtained through the Weather Query Builder of the Visual Crossing website. Visual Crossing is a leading platform [3] containing weather data and enterprise data analysis tools for researchers, business analysts, professionals, and academics. The platform is also accessible to customers or distributed PV installation stakeholders as it has open access, offering extensive data packages on a complimentary basis. The company`s Weather Data API facilitates access to a comprehensive range of historical weather information, including specialised weather measures such as solar radiation, degree days, and weather alerts.

The Visual Crossing platform provides essential meteorological variables crucial for understanding the complex interplay between weather conditions and PV installation performance. The weather factors analysed in this paper using the dataset and measurements look as follows: cloud cover %, dew point °C, humidity %, sea level pressure mb, solar energy MJ/m², solar radiation W/m², temperature °C, the temperature feels like °C, UV index, visibility km, wind direction degrees, and wind speed kph. Solar radiation measures the power at the instantaneous moment of the observation and solar energy indicates the total energy from the sun that builds up over an hour or day Information about all these weather factors is reflected on the Visual Crossing site and allows us to perform a comprehensive analysis defining the main influencers on the PV power generation. The ability to integrate such specialised weather measures into our analysis makes this research useful for diverse datasets for enhanced prediction accuracy.

Moreover, Visual Crossing`s Weather Data API opens possibilities for a wide range of projects beyond the scope of our current study. The availability of historical weather data, coupled with specialised measures, makes these datasets suitable for applications in RES power generation forecasting, climate research, and other data-driven investigations. However, due to the platform limitations, it provides only the data on a daily and hourly basis.

For the research described in this paper, the weather dataset comes from the platform mentioned above and uses the Hillingdon Borough sensor as the main location. Hillingdon Borough is in London, UK and represents an area of 42 square miles [4] which is equal to around 109 km². The weather tracking device for the Visual Crossing platform is set up at the estimated distance of 2 km from the PV installation applied for the research. That leads to insignificant errors in interpreting the weather for the PV installation located at Brunel University. The mapping representation of the objects located in the Hillingdon borough is described in Figure 3. The maps utilized in this study adhere to the terms outlined by the Google Maps service [5].

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/UPEC61344.2024.10892390, 2024 59th International Universities Power Engineering Conference (UPEC)

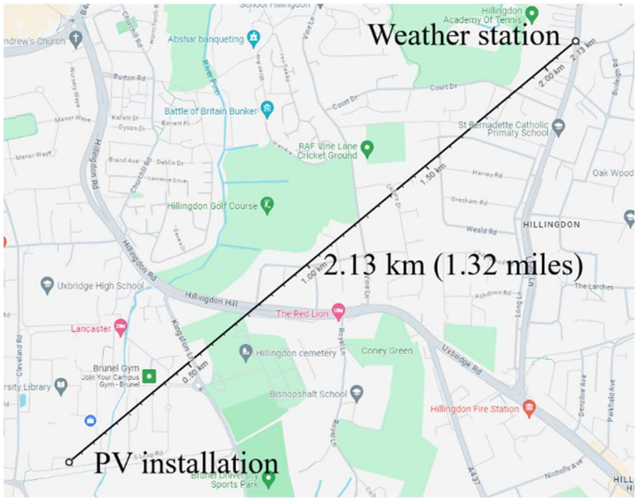

Figure 3. Representation of the PV installation and weather station locations

C. Overall dataset shaping

This research dataset is blended from two primary datasets mentioned – PV power generation and meteorological data. To get a balance between accessibility and computational efficiency and take into account the dataset`s limitations we aggregated these data based on an hourly period. Accordingly, the power generation dataset was enlarged from the 5-minute periods into the hourly values.

For the scope of our analysis, we chose the data from the year 2022. This selection aligns our investigation with actual and relevant information available. By managing this data from 2022, our research reflects the current energy landscape, capturing up-to-date features in PV installation performance. The year-long perspective offers insights into the PV installation`s behaviour across the dynamic spectrum of seasonal changes.

D. Dataset preprocessing

To obtain the best quality research results, preprocessing approaches were applied to the data. Columns containing overall date and time details were split into more detailed components, specifically day, month, year, and hour. After that, categorical variables representing specific weather conditions were transformed through the application of one-hot encoding. This technique effectively expanded categorical columns into binary vectors, each representing a unique weather category [6]. Table 1 represents the structure of the processed dataset.

TABLE I. Representation of the Final Dataset Used in the Paper

Day	Month	Year	Hour	PV generation	Temperature	Humidity	…
1	1	2022	0	0	12,8	92,77	…
1	1	2022	1	0	13,1	93,14	…
…	…	…	…	…	…	…	…
31	12	2022	22	0	12,1	81,8	…
31	12	2022	23	0	12,0	81,3	…

III. Methodology Proposed

A. Benefits of using ML for forecasting

In general ML techniques appear as a powerful tool for tackling forecasting. Utilizing ML for RES power generation offers sophisticated approaches to reveal the patterns and relationships between the data in complex datasets. Another benefit of the ML models is that historical data serves as the foundation for them and accordingly exists a continuous learning cycle, where models can be updated with new data. As a result, using more detailed datasets or expanding the overall amount of data can lead to an increase in accuracy [7].

Applied to PV installation power generation, ML shows superior quality in capturing the interplay between environmental factors and solar power output. Unlike traditional methods, ML models autonomously learn from historical data, enabling the development of accurate and adaptive forecasting models. By highlighting the most influential weather variables ML techniques allow power generation operators, owners, researchers, investors, and other stakeholders to make informed decisions for PV installation work optimization, and impact solar power generation. This insight contributes to the development of targeted strategies for enhancing PV installation efficiency [8].

B. Boruta as a feature extraction method

Boruta is a feature selection algorithm commonly used in ML to identify and retain the most relevant variables for predictive modelling. This technique uses a Random Forest (RF) classifier, to assess the importance of each feature. The Boruta involves mixing the original sample feature matrix M to create a shadow feature P, forming a hybrid feature set N = [M, P]. The mixed feature set undergoes disordering to eliminate feature-response variable correlations. This process ensures that the assessment of feature importance is unbiased and not influenced by the original data structure, allowing for a fair comparison between real and shadow features.

An RF model is established based on the mixed feature set, and features are evaluated for importance using Z values. Features with Z values higher than Z_{max}, along with significant bilateral equality test results, are retained, while those below Z_{max} are considered unimportant and deleted, iteratively repeated until confirmation or rejection of all features or reaching the maximum iteration limit [9].

Boruta, as a feature selection method, offers several advantages over other techniques like Minimum Redundancy Maximum Relevance (MRMR), RF, and Decision Tree. One of the key strengths of Boruta is its focus on all-relevant feature selection, unlike other methods that concentrate on finding a small set of predictive features, thus providing a more comprehensive insight into the data. It also shows robustness against overfitting due to its RF-based approach, which gives it an edge over single decision trees that can lead to overfitting. Boruta also can handle complex interactions and non-linear relations between features, a limitation for linear methods like MRMR [10].

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/UPEC61344.2024.10892390, 2024 59th International Universities Power Engineering Conference (UPEC)

C. ML models used for the research

Our research paper uses multiple ML models to comprehensively analyse the dataset and investigate the accuracy and reliability of the ML techniques. The model selection process aimed to capture the differing effectiveness of algorithmic approaches, each bringing unique strengths and characteristics to the predictive modelling task. The set of models used includes Linear Models (LM), Generalized Linear Models (GLM), k-Nearest Neighbors (k-NN), Support Vector Machines (SVM) with linear, polynomial, and Radial Basis Function (RBF) kernels, Generalized Boosted Models (GBM), RF, Classification and Regression Trees (CART), Stacking.

Investigating the models, LM provides simplicity and interpretability of linear relationships within the data. GLM extended this framework to accommodate non-linear relationships and heteroscedasticity in the dataset. The k-NN algorithm leverages the proximity of data points to make predictions, adapting well to local patterns [11].

SVM is a powerful class of models, employed with various kernels and widely applied to classification tasks. However, these models could be effectively used for forecasting techniques. The linear kernel represents the forecasting of linearly separable data, while the polynomial kernel allows for the exploration of non-linear decision boundaries. The RBF kernel enabled the SVM to capture complex relationships between features, enhancing its capacity to manage advanced datasets [12].

Stacking represents a combination of predictions from multiple models to create a meta-model that aggregates their strengths and compensates for individual weaknesses. This ensemble approach aimed to improve overall predictive performance and generalization by involving the complementary nature of different mathematical algorithms [13].

D. Methods and measurements for result assessment

Cross-validation is a technique that assesses ML model performance by splitting the dataset into subsets for training and testing. K-fold cross-validation, a common approach, involves training the model *k* times on different subsets, ensuring robustness and reducing dependency on specific data subsets. This method provides a reliable estimate of the model`s performance on unseen data, aiding in identifying issues like overfitting or underfitting.

For assessing the models representing PV installation power generation, Root Mean Square Error (RMSE) [14] and R-squared (R^2) are the metrics that play important roles in evaluating the accuracy and reliability of predictive models.

RMSE is a value describing the average difference of the errors between predicted and observed values. It calculates the square root of the average squared differences between predicted and actual values. In the context of power generation prediction, RMSE provides a clear and interpretable representation of the model`s predictive performance. A lower RMSE value signals that the model`s predictions are closer to the observed values, indicating better accuracy [15]. The mathematical description of the RMSE is located further:

$$RMSE = \sqrt{\sum_{t=T+1}^{T+h} (\hat{y}_t - y_t)^2 / h} \quad (1)$$

where $\hat{y}_t$ is modelled value, y_t is the real value, h, in turn, is the number of forecasting steps ahead and T is the time step [14].

R^2 is a statistical measure that indicates a perfect fit [16]. R^2 is valuable for understanding how well the model captures the underlying patterns. The values of R^2 which are close to 1 represent the better-quality models. However, in certain situations, especially when the model has a poor fit to the data, R^2 can turn negative. The mathematical representation of the R^2 can be found further:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (2)$$

where RSS is the sum of residual squares and TSS is the total sum of squares.

E. Research algorithm used for this paper

The investigation algorithm used for this paper is structured as a step-by-step process and shown in Figure 4. It runs from the collection of both historically distributed PV installation power generation data and relevant weather data. The subsequent integration of this dataset allows the production of analysis of the weather impact on energy generation, forming the base for applying the ML models and their validation.

The overall algorithm is constructed using a flowchart and follows general approaches to building structured algorithms.

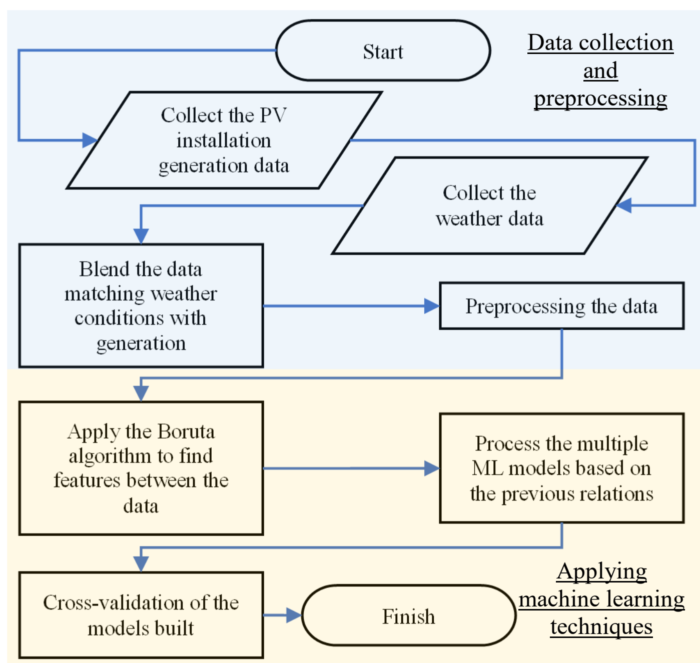

Figure 4. Algorithm used for the investigations in this paper

IV. Review of the Research Papers on Similar Topic

This chapter is dedicated to the review of current research publications in the field of ML applied to forecasting PV installation power generation to show the topicality of this

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/UPEC61344.2024.10892390, 2024 59th International Universities Power Engineering Conference (UPEC)

research. The authors investigated the recent papers exploring the Scopus database using the ‘ML’, ‘forecasting’ and ‘PV power generation’ keywords.

Accordingly, paper [17] explores the influence of Artificial Intelligence and specifically ML in facing various aspects of solar systems. This paper focuses on different applications to study the performance of state-of-the-art ML models. The paper discusses diverse ML applications in PV installation systems, covering topics from generation profile and anomaly detection to PV installation systems combined with storage, offering a comprehensive overview of the field`s challenges.

Authors in the paper [18] provide a review of recent advancements in PV installation power generation forecasting, with a specific focus on the application of ML techniques. They confirm the crucial role of accurate PV power generation forecasting for optimal grid management. Their study covers both direct and indirect forecasting pathways, incorporating solar irradiance forecast models, plane of array irradiance estimation models, and PV installation performance models. The review highlights the effectiveness of ML methods, particularly deep neural networks and ensemble or hybrid models, in outperforming traditional approaches for short-term PV installation power generation forecasting, emphasizing the significance of intelligent optimization and data preparation techniques for enhanced accuracy.

Paper [19] investigates the correlation between various input parameters and PV power generation through the application of ML models, specifically SVM and Gaussian process regression. The input parameters, encompassing solar PV installation temperature, ambient temperature, solar flux, time of day, and relative humidity, were used to predict solar PV power generation.

In conclusion, the findings of this chapter highlight and confirm the significance and wideness of using ML for PV installation power generation and closely related fields.

V. Empirical Results

Applying the Boruta feature selection method determines the influence of weather factors on PV installation power generation. Figure 5 reflects the obtained results. The key factors influencing generation are the hour of the day, solar radiation, cloud cover, solar energy, UV index and the month of the year. Consistent with earlier assumptions in this paper, solar radiation is confirmed as one of the critical factors directly influencing the power generation of the PV installation applied for this study.

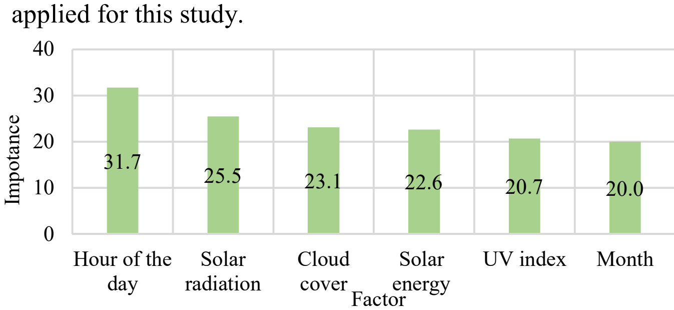

Figure 5. Importance of the factors influencing the power generation of the PV installation

After identifying these key factors, we constructed ML models accordingly. Further, an evaluation framework using 10-fold cross-validation was used for the ML models assessment. The results of the calculations are presented in Figure 6. Models Stacking represents the calculation results for the bootstrapping of the RF, k-NN, GBM and CART models.

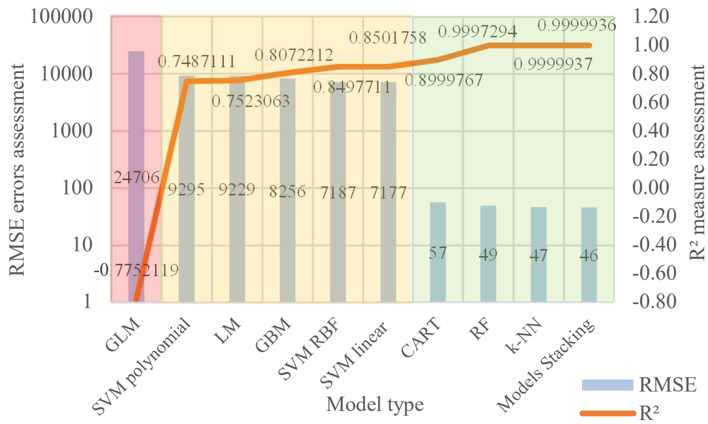

Figure 6. Performance of the ML models used for the investigations

The analysis includes the models previously discussed. Accordingly, the performance of these models can be split into 3 groups with different effectiveness. First as presented in Figure 6 and highlighted green on the graph, provide the best results for forecasting, and include models stacking, k-NN, RF and CART. The results reveal that stacking models, a combination of different models, exhibited the most promising performance. This stacking approach includes the strengths of each model, resulting in a more powerful and generalized predictive capability. The key strength of stacking lies in its ability to leverage the unique strengths of each individual model, resulting in a more powerful and generalized predictive capability. By integrating various algorithms, stacking can handle diverse types of data and relationships, providing a more comprehensive solution. The ensemble approach of stacking helps mitigate the risk of overfitting, a common issue in standalone models.

Furthermore, other standalone models such as RF, k-NN, and CART demonstrated commendable performance, showcasing their efficiency in predicting PV installation power generation based on weather factors. Despite the different approaches to these methods, they offer reliable experience for this case.

The second group, highlighted with yellow on the graph, contains all the built SVM models with different kernels, LM and GBM. These models are diverse in their approaches and characteristics. In comparison they have achieved average results, however, and still can be used for forecasting with not high requirements for accuracy.

The last group covered with the red color on the graph, contains only the GLM model and has non-satisfactory results. The highest level of RMSE in combination with a negative value of R^2 means poor fitness of the model and impossibility of further use for predictions of the PV installations.

The detailed analysis of the ML models` nature showed that k-NN, RF, and CART models are strong in their ability to handle non-linear patterns in data, which are inherent in PV installation power generation forecasting. On the other hand, models such as LM, GLM, SVM, and GBM are widely showing effectiveness in predicting linear behaviour. While these models are powerful, their assumption of linearity limits their effectiveness in scenarios where the underlying data patterns are non-linear, such as in PV power generation forecasting.

VI. Conclusions and Future Research

This study emphasizes the need for appropriate ML model selection based on the nature of the data and the specific task at hand. Moreover, the paper highlights the importance of selecting appropriate valuable features as the initial stage for future ML models to be built.

Integrating Boruta feature selection with 10-fold cross-validation has enhanced the model assessment process. The research contrasted different effectiveness of the applied ML techniques. The Stacking models, particularly when combining RF, k-NN, GBM, and CART, emerged as the top performers. However, other individual models like RF, k-NN, and CART also demonstrated notable efficacy. The models with linear mathematical backgrounds reveal unsatisfactory results.

Given the effectiveness of ML models in PV installation power generation forecasting, the authors` future research will delve deeper, combining techniques and utilizing more detailed datasets to enhance predictive accuracy. By investigating extended datasets, we aim to uncover the influence of additional factors contributing to forecasting results. Also, future research could focus on using the closer weather sensor to improve results for the PV installation.

References

[1] International Energy Agency. (2023) *Solar PV*. [Online]. Available: www.iea.org/energy-system/renewables/solar-pv

[2] Sunny Portal. (2023) *PV System Overview Brunel Uni Solar*. [Online]. Available: www.sunnyportal.com

[3] Visual Crossing. (2023) *About Page*. [Online]. Available: www.visualcrossing.com/about

[4] Hillingdon borough. (2024) *Facts and statistics about the borough of Hillingdon*. [Online]. Available: https://www.hillingdon.gov.uk/article/8942/Facts-and-statistics-about-the-borough-of-Hillingdon#:~:text=Hillingdon%20is%20the%20second%20largest,%2C%20rivers%2C%20parks%20and%20woodland.

[5] Google. (2024) *Google Maps/Google Earth Additional Terms of Service*. [Online]. Available: https://www.google.com/help/terms_maps/

[6] T. Al-Shehari and R. A. Alsowail, 'An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques', Entropy, vol. 23, no. 10, 2021, doi: 10.3390/e23101258.

[7] S. Salcedo-Sanz, L. Cornejo-Bueno, L. Prieto, D. Paredes, and R. García-Herrera, 'Feature selection in machine learning prediction systems for renewable energy applications', Renewable and Sustainable Energy Reviews, vol. 90, pp. 728–741, 2018, doi: https://doi.org/10.1016/j.rser.2018.04.008.

[8] H. Ghoddusi, G. G. Creamer, and N. Rafizadeh, 'Machine learning in energy economics and finance: A review', Energy Econ, vol. 81, pp. 709–727, Jun. 2019, doi: 10.1016/J.ENECO.2019.05.006.

[9] R. Tang and X. Zhang, 'CART Decision Tree Combined with Boruta Feature Selection for Medical Data Classification', in 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), 2020, pp. 80–84. doi: 10.1109/ICBDA49040.2020.9101199.

[10] M. Sharma and D. K. Sharma, 'Optimising Feature Selection: A Comparative Study of mRMR-Boruta/RFE Hybrid Approach', in 2023 6th International Conference on Information Systems and Computer Networks (ISCON), 2023, pp. 1–6. doi: 10.1109/ISCON57294.2023.10112125.

[11] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.

[12] Y. Liu, H. Chen, L. Zhang, X. Wu, and X. Wang, 'Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China', J Clean Prod, vol. 272, p. 122542, 2020, doi: https://doi.org/10.1016/j.jclepro.2020.122542.

[13] B. Pavlyshenko, 'Using Stacking Approaches for Machine Learning Models', in 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018, pp. 255–258. doi: 10.1109/DSMP.2018.8478522.

[14] D. Hulak and G. Taylor, 'Investigating an Ensemble of ARIMA Models for Accurate Short-Term Electricity Demand Forecasting', in 2023 58th International Universities Power Engineering Conference (UPEC), 2023, pp. 1–6. doi: 10.1109/UPEC57427.2023.10294946.

[15] M. Ćalasan, S. H. E. Abdel Aleem, and A. F. Zobaa, 'On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function', Energy Convers Manag, vol. 210, p. 112716, 2020, doi: https://doi.org/10.1016/j.enconman.2020.112716.

[16] P. K. Ozili, 'The acceptable R-square in empirical modelling for social science research', in Social research methodology and publishing results: A guide to non-native english speakers, IGI Global, 2023, pp. 134–143.

[17] G. M. Tina, C. Ventura, S. Ferlito, and S. De Vito, 'A State-of-Art-Review on Machine-Learning Based Methods for PV', Applied Sciences, vol. 11, no. 16, 2021, doi: 10.3390/app11167550.

[18] P. Gupta and R. Singh, 'PV power forecasting based on data-driven models: a review', International Journal of Sustainable Engineering, vol. 14, no. 6, pp. 1733–1755, Nov. 2021, doi: 10.1080/19397038.2021.1986590.

[19] B. Zazoum, 'Solar photovoltaic power prediction using different machine learning methods', Energy Reports, vol. 8, pp. 19–25, 2022, doi: https://doi.org/10.1016/j.egyr.2021.11.183.