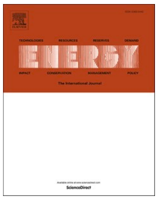
Energy 278 (2023) 127807

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy



Machine learning for forecasting a photovoltaic (PV) generation system

Connor Scott [a], Mominul Ahsan [b,*], Alhussein Albarbar [a]

[a] Department of Engineering, Manchester Metropolitan University, Chester St, Manchester, M1 5GD, UK
[b] Department of Computer Science, University of York, Deramore Lane, Heslington, York, YO10 5GH, UK

ARTICLE INFO

Handling Editor: X Zhao

Keywords:
Photovoltaic (PV)
Renewable energy (RE)
Machine learning
Random forest (RF)
Neural networks (NN)
Support vector machines (SVM)
And linear regression (LR)

ABSTRACT

To mitigate the carbon print of buildings, they should have on-site renewable energy generation systems to supply energy for the buildings without relying on the national grid. Renewable generation sources rely on weather conditions and are therefore difficult to rely on as the only source of energy. Photovoltaic (PV) is forecasted through machine learning algorithms (MLA), but different methods have varied accuracy and have different training requirements such as more inputs or more data in general. No previous research has concluded an optimal MLA but to better apply them to PV systems, this must be established. To conclude an optimal MLA for a particular application, the dataset and required outputs must be determined, and how they affect the performance of the algorithm must be evaluated. The aim of this work is to compare benchmark MLA's through accuracy and usability for an operational University campus located in central Manchester, in the north of England. The MLA's including random forest (RF), neural networks (NN), support vector machines (SVM), and linear regression (LR) have been employed to forecast the PV system. If the power output of the renewables is accurately forecasted, a building management system (BMS) can be equipped to optimise on-site renewable energy generation. To accomplish this, sixty-four MLA models are created in total for forecasting at multiple horizons and dataset sizes which are validated against real-time data. Results in this work revealed that the RF algorithms have the lowest average error of the multiple tests at 32 root mean squared error (RMSE), whereas SVM, LR, and NN showed at 32.3 RMSE, 36.5 RMSE, and 38.9 RMSE respectively. Errors between forecasted and actual results are recorded in RMSE whereas changes in error are shown in mean actual percentage error (MAPE) to show the changes with respect to the original value. No MLA outperforms all others for accuracy and for requiring less data. No previous research is conducted to evaluate the performance of various MLA PV forecasting models through various sized data sets with critical analysis on the results. The comparison of benchmark algorithms when forecasting the PV generation of a local system allows the critical analysis of the models' accuracy and surrounding characteristics.

1. Introduction

Large business's consume 45% of non-domestic building energy [1] and the UK has plans to become the world leader in green energy, stating that by 2050, buildings will need to be almost completely decarbonised [2]. Renewable energy (RE) cannot be considered a reliable source [3] but can provide a more reliable energy mix [4], with promotion of RE by the world leaders improving application [5]. Energy generation in the UK connected at the distribution level accounts for 28% of all generation [6] which shows there is clear room for increasing decentralised RE. RE generation contributes 17.18 GW which is 43% of the UK's total energy demand [7], with 4% solar and 31% wind due to the 6500 wind turbines installed across the country. This shows usability of various RE sources to contribute to the whole demand.

Building management systems (BMS)'s are an increasingly researched topic that should consider all aspects of a building including physical models, environmental conditions, comfort, safety, occupants' preferences, thermal, and visual specifications [8]. A BMS is designed to improve the efficiency of all building functions, mostly through optimisation of energy actuation and use [9]. The accuracy of prediction techniques and local climate conditions are critical for an optimisation system [10] with BMS's reducing annual energy consumption by an average of 16% [11] and saving costs in the range of 11% [12]. Combination of a BMS and smart-design features can reduce energy consumption by up to 49% [13,14]. A control strategy is developed to forecast energy generation and on-site energy storage for a reduction of

[*] Corresponding author.
E-mail address: md.ahsan2@mail.dcu.ie (M. Ahsan).

https://doi.org/10.1016/j.energy.2023.127807
Received 15 December 2022; Received in revised form 12 April 2023; Accepted 11 May 2023
Available online 11 May 2023

energy costs of up to 84% [15]. On site-storage is forecasted with the intention of reducing the size of required battery systems [16]. The MLA's applied to BMS's are used to forecast on-site energy generation among other energy characteristics. Multiple MLA's can be applied to forecast PV generation [17] with different results of accuracy. Supervised machine learning relies on historical data of inputs and PV generation data to train the algorithm, but with various algorithms, it cannot be undisputed which algorithms provide the most accurate forecasts.

Recent research contains many single uses of machine learning for their respective datasets with the aim of optimisation towards their respective dataset [18–21]. The MLA cannot be accurately compared against a different MLA with a different dataset as there are many variations in the methodology. This leaves a problem in which there cannot be a justifiable conclusion of a more suitable method of machine learning algorithm for PV forecasting.

This research compares ML techniques for the forecasting of a localised PV system in a cool climate with various datasets to resemble real scenarios. For the successful application of ML to a buildings' PV system, the right ML technique must be used, in which the comparison of actual results in this research are crucial. Furthermore, the use of various datasets authenticates the method of comparing the MLAs for real scenarios.

The contributions are given as below.

1) A comprehensive analysis of MLA performances for the application of photovoltaic forecasting into a building management system.
2) Computational power, training and prediction speed, and accuracy over various sized datasets and horizons are analysed.
3) Feature importance of a UK roof-mounted PV system is evaluated through the maximum relevance minimum redundance algorithm, providing critical information on what data is necessary for an accurate algorithm. This may lead any future applications of MLA for forecasting PV as only the necessary data can be collected, and effective algorithms can be developed by considering the results in this research.

The novelty of this research is the comparison between benchmark MLA's for the forecasting of a local PV system for a non-domestic building in a cooler climate (UK). The forecasting of on-site renewables is crucial towards the improvement of energy efficiency in buildings and the effect on the national grid [22]. Previous research shows the independent accuracies of MLA's for the forecasting of PV generation plants, and localised systems, with various data and ML methods. Each algorithm performs better for different PV systems, meaning no decision can be made on the best algorithm for the purpose of PV forecasting as no investigation has occurred under the same conditions. The proposed work analyses the available MLA's and critically analyses them for real-world utilisation through varied training and forecasting datasets on a working localised PV system. The motivation for this research is to reduce carbon emissions of that buildings through optimisation of renewable energy generation. This is performed through MLA forecasting of the energy generation, but it is not certain which algorithm provides the best results and how much data needs to be collected. For buildings with less or incorrect collected data, the application of MLA's can be difficult as they require cleaned data and depending on the algorithm, a defined time of collected data. This work computes the necessary data for an accurate forecast, with the aim of showing the algorithms forecasting shorter and longer horizons with varied datasets.

The paper is organized as below. The literature review in section 2 summarises previous work on the applications of MLA's for buildings' energy characteristics. The novelty and contribution to knowledge are reiterated in this section. The proposed research methodology in section 3 describes the data collection and processing, feature importance and selection, the datasets used in each algorithm, an explanation of how each algorithm works, and how the machine learning models are validated and tested. The results and analysis in section 4 show the accuracy of each algorithm depending on the size of the dataset it is trained with, the effect less inputs have on the accuracy, and the training and forecasting speeds of the algorithms. Critical analysis and discussion in section 5 contain a deeper description of the results, limitations of the study, and comparison with existing works. The conclusion in section 6 summarises the works completed in the study. 2. Literature Survey.

1.1. *Photovoltaic energy generation*

Two types of solar energy including thermal and photovoltaic (PV) are available in the market. There are varied efficiencies of PV energy generation depending on the climate [23], with the integration of photovoltaic and thermal methods showing greater energy generation [24,25]. Application of solar technologies is just as important as generation as [26] shows various uses to aid building systems reducing energy demand. PV generation doesn't rely on heat and instead generates energy almost purely through sunlight. It is the preferred method for this research in a cooler climate because there is less heat energy but there is still enough light for a PV system to produce enough energy to be useful towards the building. Renewable generation can be stored within the building through 11 different methods of recent research with 6 of them being consistently improved [27]. The application of renewable generation energy storage relies greatly on accurate forecasting of the buildings' energy demand and renewable generation [28]. Renewable energy contains many forms such as solar PV, solar thermal, hydro, wind, geothermal, and more, but the accurate forecasting of any RE provides necessary information for an optimised BMS. Demand and on-site generation are the most combined forecasting objectives for home and building energy management systems [29,30]. PV forecasting is vital towards the optimisation of economical operation of the grid [31].

1.2. *Machine learning algorithm methods and applications*

A machine learning algorithm (MLA) can form mathematical bonds from inputs to outputs of a given dataset. When new data is added, the bonds remain, and the outputs can be predicted [32]. More training data increases the accuracy of MLA's with 97 unknown datasets [33] but more data isn't always necessary for an accurate forecast as 95% accuracy is achieved with only 200 out of 5000 samples [34]. Neural network (NN) is used to forecast PV generation in Refs. [18–21,35] with an average MAPE of 15% over various horizons no more than a day. Random forest (RF), support vector machine (SVM) and linear regression have a mean actual error (MAE) of 1.47, 5.92, and 9.3 J/m^2 respectively, with RF providing higher accuracy [36]. A comparison of MLA's for PV forecasting shows high accuracy from kernel ridge but takes an extremely long training time and huge memory. The NN has the second highest accuracy with a much lower training time and computational power [37]. An MLA is recommended for forecasting the energy generation of a solar plant.

Previous models are trained and tested on a fixed dataset whereas for real applications, buildings have been collecting data for various time periods. Various accuracies have been achieved in recent research between 1.7% error [38] and 10% error [39] and forecasting horizons are between 1 min [21] and 36 h [39]. A single model is developed for optimisation of a single dataset [21,38], whereas multiple datasets are used which also cannot be used to compare the performance of multiple algorithms [39]. The size of the PV systems vary between the largest solar plant in the world [36] and various localised systems [20]. The 1.7% error was achieved through a SVM for a 3 kWp PV system in Australia with 6672 training points with an unspecified forecast horizon. The 10% error was achieved through an analogue ensemble algorithm which is very similar to a K-nearest neighbour. The system was tested on 3 plants with a peak power output of 11,994 kW, 2,649 kW, and 6,876 kW, in south-west France. The training data is in 30-min intervals from

January 2014–September 2018 having at least 80,352 training points over 7 inputs and 5 years. The forecast horizons varied from 30 min to 36 h in advance. The comparison of the two algorithms with 1.7% and 10% error isn't a fair comparison as they are being used with many different variables. Average temperatures across previous research vary from 24.66^{O}C in central China with 4% error [40] to 11.73^{O}C in Belgium with 11.89% error [20]. The correlation between 6 case study errors and average temperatures of the locations is 0.338 [18,20,38,39,41,42], showing warmer temperatures have slightly higher accuracies. The variation between these studies' PV system sizes, amounts of data, and method of forecasting vary drastically, making the methods very difficult to compare. The sizes of PV systems' forecasted through MLA's vary from 75 MW [18] to 700W [43] with outdoor air temperatures ranging from 23.12 °C [44] in Peru to 8.04 °C [41] in America. The errors of the forecasts of the PV generation range from 1.64% [45] to 13.04% [46] across the 4 algorithms and variety of developed models. The numerous methods of PV forecasting through MLA's provide varied results of accuracy, different forecast horizons, and can be applied for different sized systems. As each of these results are collected from various algorithms, using different sizes and types of data, from a different sized PV system, in different climates, no conclusion can be made on which algorithm is most suitable for the desired forecast. The research in this paper focuses on the use of varied ML methods with different datasets and forecasting horizons for a local PV system in the UK. This is a novel method of comparing the effectiveness of each ML forecasting method through an actual PV system in a cool climate while maintaining the same conditions for each test.

2. Proposed research methodology

The proposed method employing MLA's are used to forecast installed PV generation with validation of the model against actual data collected from an operational university campus. The MLA's are trained using previous collected weather and solar generation data which corresponds with that horizon. The MLA's consist of neural networks, random forests, linear regression and support vector machines. Data collection and processing algorithms are explained in this section. Comparison of developed methods are determined through accuracy of forecast, through training data of weekly, monthly, and yearly, for forecast horizons of 15-min, hourly, and daily. The models are evaluated using mean actual percentage error (MAPE) when compared against actual data.

2.1. Data collection and processing

The raw PV generation data contains 10 months of data and 30,842 iterations. The predictor variables are cloud coverage (%), humidity (%), rainfall (mm), air pressure (mb), temperature (^{o}C), and wind speed (mph). The entire original dataset contains 215,818 data points across the 6 inputs and the output (PV generation) for 10 months of data collection (see Fig. 1).

The initial dataset is split into multiple smaller datasets for each algorithm to be trained and tested on. These are aggregated to include the original dataset of 15-min iterations, hourly, and daily measurements. These can be used to train the algorithms with a varied time lag to

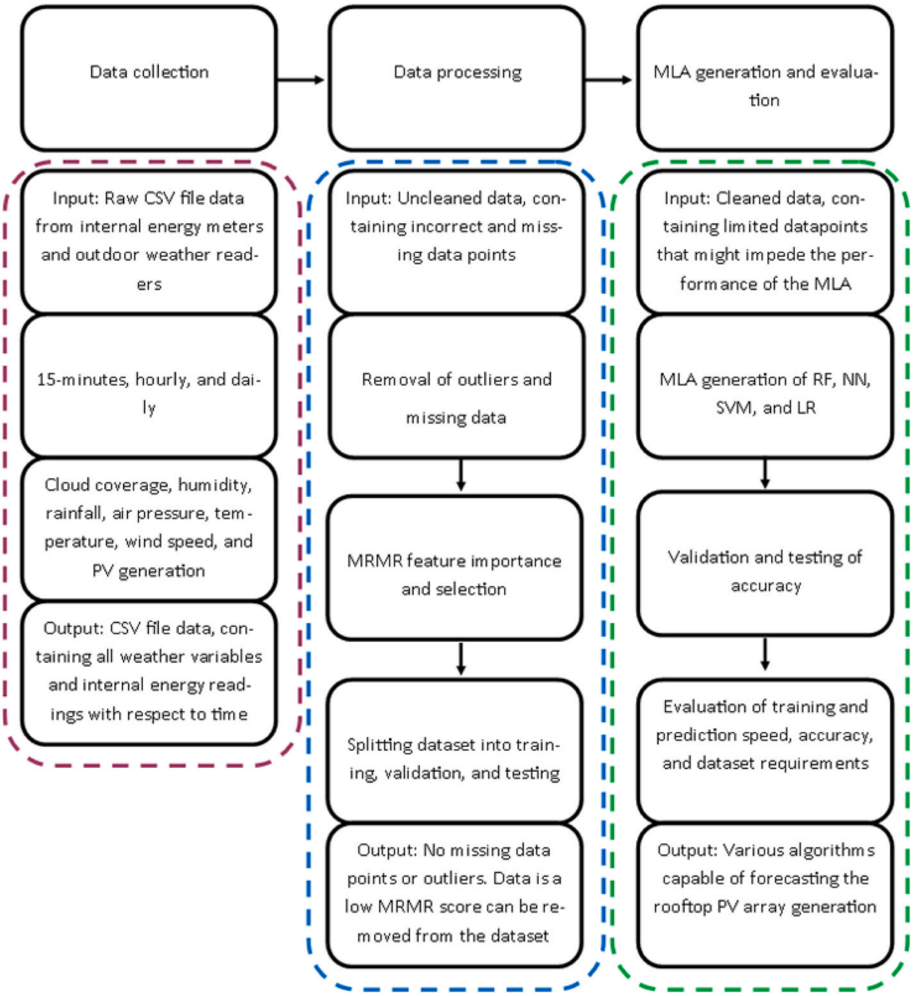

Data collection
Data processing
MLA generation and evaluation
Input: Raw CSV file data from internal energy meters and outdoor weather readers
15-minutes, hourly, and daily
Cloud coverage, humidity, rainfall, air pressure, temperature, wind speed, and PV generation
Output: CSV file data, containing all weather variables and internal energy readings with respect to time
Input: Uncleaned data, containing incorrect and missing data points
Removal of outliers and missing data
MRMR feature importance and selection
Splitting dataset into training, validation, and testing
Output: No missing data points or outliers. Data is a low MRMR score can be removed from the dataset
Input: Cleaned data, containing limited datapoints that might impede the performance of the MLA
MLA generation of RF, NN, SVM, and LR
Validation and testing of accuracy
Evaluation of training and prediction speed, accuracy, and dataset requirements
Output: Various algorithms capable of forecasting the rooftop PV array generation

Fig. 1. The methodology from data collection to PV forecasting for multiple algorithms and datasets.

forecast a specified horizon. The hourly dataset contains 7281 iterations and the daily dataset contains 302 iterations. For each smaller dataset, another predictor variable is added for the time of day. This ranges from 1 to 96 for 15-min iterations, 0–24 for hourly, and 0–7 for daily. Daily forecasting has larger errors when used with a ‘time of day’ input, so it isn't used as an input for training. To show the effects of processing the data, the training and testing hourly dataset is shown in Fig. 2.

Major outliers and missing data are emphasised, where the data is much too high, too low, or completely missing. This is due to the sensors malfunctioning and restarting, which aggregates all the missing data collected over a period of time into a high value at a single timescale. The data after processing is shown in Fig. 3.

Major outliers and missing data are corrected and removed to provide the algorithms with higher quality training and testing data. Although there are clearly data points that don't follow the trend of the data, they are not outlying by 90% or more. For data that doesn't follow the trend but also isn't outlying, it can often be advantageous to include it in the training set to increase the algorithms' ability to learn a variety of data. The missing data points are filled through linear interpolation or removed, and outlying data points are corrected through a rolling average of the surrounding day, week, or month, depending on the size of iteration. The linear interpolation and rolling average methods are shown in equations (1) and (2) respectively.

$$y = y_1 + (x - x_1)\frac{(y_2 - y_1)}{(x_2 - y_1)} \tag{1}$$

Where the previous and next points in the dataset are ‘y_1’, ‘y_2’, ‘x_1’, and ‘x_2’ respectively for each variable. Current points are represented with ‘x’ and ‘y’.

$$\mu = \frac{1}{N}\sum_{-n}^{n} A_i \tag{2}$$

The rolling average is ‘μ’ and is the sum of the selected values, from ‘ − n’ to ‘n’ divided by the number of variables between those points ‘N’. The values of ‘n’ influence the calculated value of the rolling average, depending on the data. If the data is volatile, the values of ‘n’ should be smaller, and thus will cover less data around the missing dataset. This is because there will be more known data in the form of ‘n’ that is varied from the missing data that is being calculated. The value of ‘n’ may be larger for less volatile data as known data in the form of ‘n’ will have less of an effect on the calculation. This is because the known values are closer in value to the unknown values that are being calculated. In the case of the missing data shown in Fig. 2, no missing data consists of more than a single month and so the values of ‘n’ are equal to the 30 days either side of the missing data, totalling at 60 days, or two months of daily data. This value has been chosen as it still gives a full month of correct data to use even when the missing data is in the middle of a month and there is 15 days of missing data either side of it.

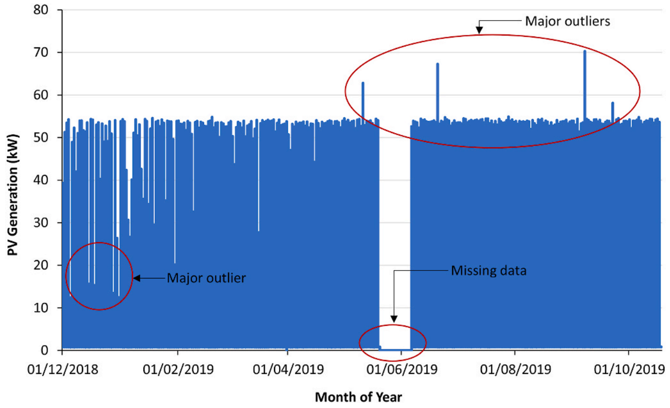


Fig. 2. The hourly dataset, spanning 11 months; 10 months of training and 1 month of testing.

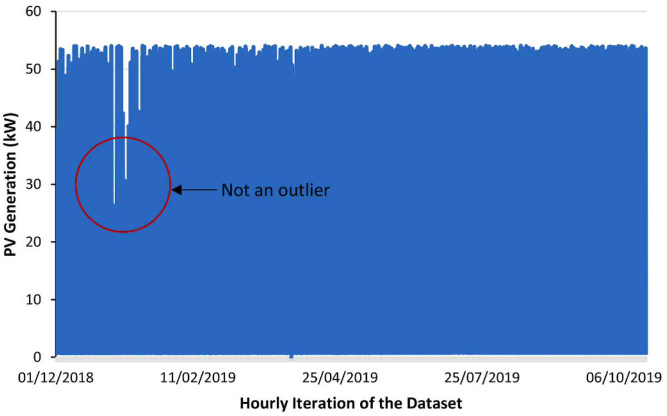


Fig. 3. The hourly dataset after processing.

The three datasets that the MLA's are trained with from 10 months of data are shown in Table 1.

The three iterations are used to train the models on 2 different horizons each. 15-minute iterations are forecasted for 15-min in advance and a full day in advance. Hourly iterations are forecasted for 1 h and 1-day horizons, and daily iterations are forecasted for 1 day and 1 week horizons, so each horizon has a different sized dataset. The total datapoints are the sum of iterations multiplied by the sum of the input variables. The three datasets that the MLA's are trained with from 1 month of data are shown in Table 2.

2.2. Machine learning algorithms

2.2.1. Random forest (RF)

Random forest is an ensemble algorithm that use an aggregated result of multiple decision trees to determine the outcome. The data is recursively split to classify the target data when given a set of predictor data. The size of the random forests can be optimised for the dataset which range from 1 leaf per tree to 50, and between 30 trees and 50 for each algorithm. Gini impurity is used to decide whether to continue splitting the data.

$$Gini = 1 - \sum_{i=1}^{n}(P_i)^2 \tag{3}$$

It can be defined as the deduction of squared probabilities of each class from one, where ‘P_i’ is the probability of an element being classified for in a certain class.

Once the Gini impurity reaches the minimum value, it can be considered a ‘pure’ split, meaning it no longer must be split. This means that the tree cannot split the data to a better degree and the algorithm has finished training. The data splits and values of the target variables are remembered, and once new data is added, the target variables can be

Table 1
The average size of the datasets used for training when forecasting various horizons with 10 months of data.

Iteration	Inputs	Total Datapoints	Training Features
15 min	8	220,000	Cloud coverage, humidity, rainfall, air pressure, temperature, wind speed, PV generation, and time of day
Hourly	8	58,228	Cloud coverage, humidity, rainfall, air pressure, temperature, wind speed, PV generation, and time of day
Daily	7	2285	Cloud coverage, humidity, rainfall, air pressure, temperature, wind speed, and PV generation

Table 2
The average size of the datasets used for training when forecasting various horizons with 1 month of data.

Iteration	Inputs	Total Datapoints	Training Features
15 min	8	23,087	Cloud coverage, humidity, rainfall, air pressure, temperature, wind speed, PV generation, and time of day
Hourly	8	5764	Cloud coverage, humidity, rainfall, air pressure, temperature, wind speed, PV generation, and time of day
Daily	7	210	Cloud coverage, humidity, rainfall, air pressure, temperature, wind speed, and PV generation

calculated through splitting the data like it did while training.

2.2.2. Neural network (NN)

Neural networks build non-linear relationships between predictor and target variables. These relationships are determined through weights assigned to each input depending on the importance of the input towards the prediction of the target variable. In this case, weights are selected at random, and the error of the target variable is measured with comparison to the actual target variable. The weights can then be altered. If the error decreases, the weights will be continuously altered until the error reaches a minimum. This provides the algorithm with the weights for the inputs that provide the highest accuracy. To do this, the Levenberg-Marquardt training algorithm is used. This is a combination of the two simpler algorithms including gradient descent and Gauss-Newton. It combines the better features from each algorithm to find the minimum error quickly. Firstly, it uses the gradient descent method to find the weights with the lowest error with large steps in the weights. Then it uses the Gauss-Newton method to find the value of the weights more accurately with smaller steps. Once the weights are calculated that provide the minimum error, new data can be added, and the weights remain, allowing the network to forecast the target variable. The NN's applied in this work vary from 10 neurons to 25 neurons that are single or double layer with an ReLU activation function and a limit of 1000 iterations. 10 neurons with a single layer are used for the daily models, 20 neurons with a single layer are used for hourly models, and 25 neurons for two layers are used for 15-min resolutions. As data is collected in smaller iterations, the relationship between the training and target data becomes more complex. To calculate more complex relationships, more neurons and layers are added to the network.

2.2.3. Support vector machines (SVM)

Support vector machines use kernel functions to separate and transform the data through a hyperplane so the data can be categorised. The most used function is the radial basis function kernel (RBF). The SVM's used range from linear to medium gaussian. Once the training data is categorised, new data can be used within the same categories and relationship between training and target data. The RBF kernel is explained in equation (4) [47].

$$K(X_1, X_2) = \exp\left(-\frac{\|X_1 - X_2\|^2}{2\partial^2}\right) \tag{4}$$

In equation (4), '$K(X_1,X_2)$' is the function of the input variable(s) to the target, '∂' is the variance, and '$\|X_1, X_2\|$' is the Euclidian distance between two points. The Euclidian distance is calculated through equation (5).

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{5}$$

In equation (5), it can be described as the sum of the distances between the number of inputs selected to the target variable. 'k' is the amount of input variable distances from the target variable are calculated and summed, 'x' is the input variable(s) and 'y' is the target variable.

2.2.4. Linear regression (LR)

Linear regression is the simplest in terms of mathematical computation as the slope of the line of the training data is calculated and is used when provided with new information. It doesn't have to update the model which often results in quicker training and forecasting when compared to other benchmark models. It is used to estimate the relationship between two variables including the input and the target. This can be computed through equation (6).

$$y = B_o + B_1 X + \varepsilon \tag{6}$$

In equation (6), 'y' is the predicted output value when an input value is specified, 'B_o' is the predicted value of the output when the input is 0, 'B_1' is the relationship between the input and the output, 'X' is the input variable, and 'ε' is the error between the estimated value of the output and the actual value [48].

2.3. Feature importance and selection

The maximum relevance minimum redundance (MRMR) algorithm is used to calculate which features have the best impact on the forecasting algorithms accuracy and efficiency. It calculates the relevance of the predictor feature to the target feature while also calculating how closely linked that predictor feature is to all other predictor features. This negates the scenario where two features have a high correlation that one of them is not improving the performance of the forecasting algorithm. The maximum relevance, minimum redundance, and information gain calculations are provided in equations (7)–(9) respectively. The maximum relevance equation is given in equation (7).

$$Vs = \frac{1}{|S|}\sum_{x \in S} I(x, y) \tag{7}$$

In the above equation (7), the importance of the feature with respect to the target is calculated. Where 'S' is the set of features, 'x' is the predictor, 'y' is the target, and 'I' is information gain. The information gain of the given feature to the target is summed for each iteration to give maximum relevance of the feature to the target [49]. The minimum redundance equation is shown in equation (8).

$$Ws = \frac{1}{|S|^2}\sum_{x,z \in S} I(x, z) \tag{8}$$

In equation (8), 'Z' symbolises another feature and not the target. It is calculated the same way as maximum relevance, but instead of with respect to the target, it is with respect to a different feature, giving minimum redundance of a feature [49]. Information gain is shown in equation (9).

$$I(X,Z) = \sum_{i,j} P\left(X = x_j, Z = z_j\right) log \frac{P\left(X = x_j, Z = z_j\right)}{P(X = x_i)P(Z = z_i)} \tag{9}$$

In the above equation (9), information gain shows the mutual information between two variables, where the uncertainty of one variable can be reduced by knowing the other variable. 'I' is information gain, '(X, Z)' are variables and 'P' is the probability of that event occurring. It can be described as each variable's probability of occurring simultaneously divided by the probability of them occurring independently. The iterations are summed, and it is multiplied by log to give an answer between 0 and 1, where the variables are independent at 0 and are completely dependent at 1 [49]. Each algorithm requires varied pruning of predictor features, so the performance of a pruned model can be compared to that of a benchmark model to dictate the optimal MRMR feature selection. The accuracy of the developed methods is validated against actual data collected from the case study public building. This is evaluated in mean-actual-percentage-error (MAPE)due to it allows a fair comparison between various sized PV systems as it scales as a percentage.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{10}$$

The errors between the forecasted and actual value are multiplied by 1% of the actual value in equation (10). All the values are averaged to give an average MAPE for the forecast horizon.

3. Results and analysis

Each ML technique is applied for minimum and maximum horizon forecasting when trained with 1 month and 10 months of data. Minimum horizons are just 1 step ahead of the time iteration. The maximum forecasting horizons for daily, hourly, and 15-min are 1 week, 1 day, and 1 day in advance, respectively. Results of RF, NN, LR, and SVM forecasting are displayed in this section.

3.1. Forecasting obtained by RF

Results for RF forecasting errors based on the size of data are shown in Fig. 4. The errors range from 123.53 RMSE to 1.98 RMSE, with smaller forecast horizons providing a higher accuracy. The RF algorithm has an average error of 32.02 RMSE throughout all tests. An RF algorithm is trained with 10 months and 1 month of data and is used to forecast daily, hourly, and 15-min iterations. Each iteration is forecasted with a small-time horizon into the future and for a larger time horizon. Daily minimum and maximum forecasting horizons are 1 day and 1 week. Hourly minimum and maximum forecasting horizons are 1 h and 1 day. 15-minute minimum and maximum forecasting horizons are 15-min and 1 day in advance. There are three-time iterations each for when the algorithm is trained with 10 months and 1 month of data considering daily, hourly, and 15-min.

Each of these algorithms are then tested for minimum and maximum time iterations as is shown in Fig. 4. These variables are the same for each ML technique. The daily error is much higher when trained with 10 months of data for the RF technique compared to just 1 month of training. This involves much less data for training, for 1 month, but the RF algorithm still only creates the same number of decision tress, regardless of how much data it is given. The RF performing poorly when trained with more data but performing better when trained with less data shows how the algorithm has been overfitting the data. The algorithm has been trained with too much data, showing that it has calculated the relationship between the training data and the target data to a high degree. This provides a high accuracy when training the algorithm, but this sometimes does not translate to new data. If new data varies from the training data, which in this case it does, the algorithm is not able to provide accurate forecasts. This is not demonstrated well for a minimum forecasting horizon but is clearly shown for the maximum forecasting horizon where the forecasting error has increased. 50 DT's are used for each algorithm to allow accurate comparisons of using different sized datasets. To eliminate overfitting, the algorithm can be validated against new data, and less DT's or less data can be used. More DT's and data result in more overfitting and thus, less accuracy when provided with new data. More trees within an algorithm can provide a more stable prediction, but the computation time increases linearly [50].

3.2. Forecasting obtained by NN

Results for NN forecasting errors are presented in Fig. 5. The NN forecasting errors range from 170.86 to 1.76 RMSE. The maximum error is when the algorithm is trained with only 1 month of data and is forecasting daily PV generation with a 1-day horizon. In 4 out of 6 forecasts, the maximum horizon forecast provides a higher error than when the minimum horizon is forecasted. The NN algorithm has an average error of 38.88 RMSE for all tests. It is more affected by lower quality data as in Fig. 9 when the lowest importance data is removed, and the error is decreased.

3.3. Forecasting obtained by LR

Results for LR forecasting errors are shown in Fig. 6. The LR forecasting errors range from 99.019 to 5.3187 RMSE. Minimum horizon forecasting consistently provides lower errors except from when the algorithm is trained with 10 months of data for hourly and 15-minite forecasts. In this case, the minimum horizons provide a lower accuracy. The LR algorithm has an average error of 36.47 RMSE for all tests. The algorithm provides lower accurate forecasts when forecasting for hourly and 15-min iterations than other algorithms, whereas it outperforms the RF algorithm when forecasting for daily iterations. Even though the LR algorithm is the simplest model, it can still provide a higher accuracy than RF and SVM when trained with 10 months for a daily forecast.

When trained with only 1 month of data, it provides less accuracy than RF and SVM due to the need for more data. As it calculates the linear relationship between the inputs and target data, it requires enough data to do so. When it is given a limited amount of data, it cannot calculate the relationship accurately. The only other algorithm that requires more data than the LR is the NN when it is forecasting the PV generation in daily horizons. This is due to the limited amount of data in daily iterations.

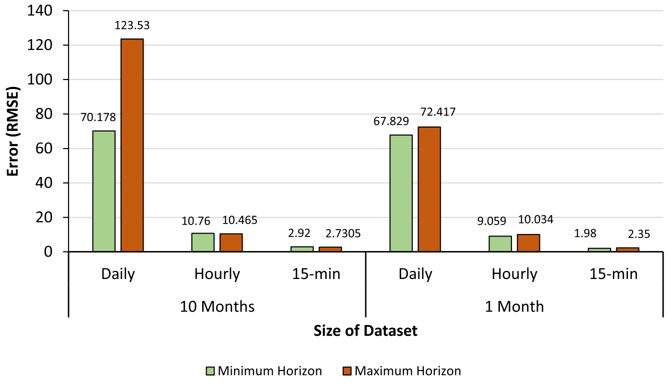


Fig. 4. The RF forecasting results when trained with 1 month and 10 months of data.

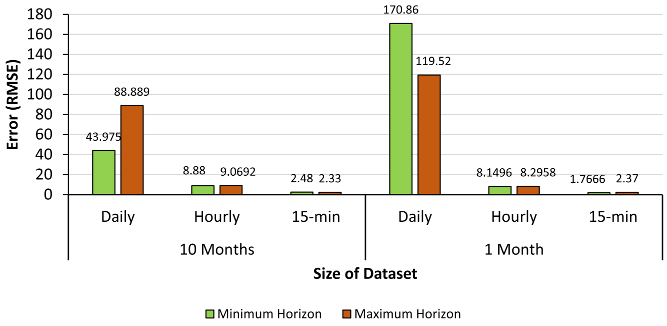


Fig. 5. The NN forecasting results when trained with 1 month and 10 months of data.

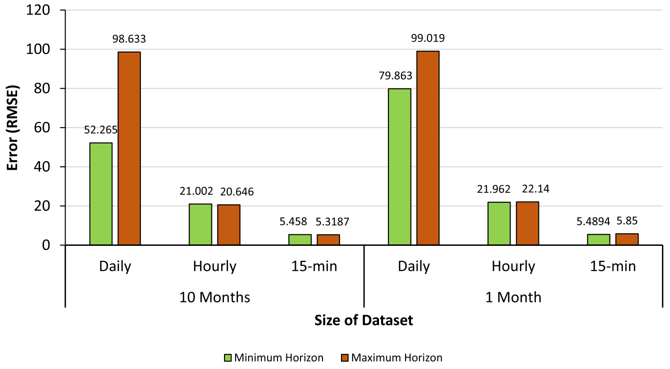


Fig. 6. The LR forecasting results when trained with 1 month and 10 months of data.

3.4. Forecasting obtained by SVM

The SVM algorithm has a maximum and minimum error of 84.85 and 2.61 RMSE respectively. Results for SVM forecasting errors are shown in Fig. 7. Maximum horizons have a higher error in 4 out of 6 tests but the RMSE error difference between the minimum and maximum horizons are much smaller than other algorithms. The average RMSE for the SVM algorithm across all tests is 32.34. The variation of the SVM's performance is limited when less data is given. When compared against the other algorithms, the SVM is better when forecasting all larger horizons than LR. Overall, the SVM models have higher accuracies than 16 other models.

The limited error increase from 10 months of training to 1 month shows that SVM's require less data to be optimised, but they cannot compete with the other techniques in this research when given 10 months of data apart from while forecasting the daily model. The daily model has the smallest dataset for 10 months of data. The MRMR feature importance findings are presented in Fig. 8.

As the MRMR algorithm compares importance of features with respect to the target, it also compares importance with respect to other training features. If multiple features have a high correlation between each other, they may slow down the MLA if they are both used for training as they may have the same effect with the target feature. To eliminate unnecessary training features, the MRMR algorithm calculates the feature importance with dependence on correlation with other features too. Although the rainfall may have high importance when used to train an MLA independently, it has less importance when used with other training features and thus, the MRMR score is represented in Fig. 8.

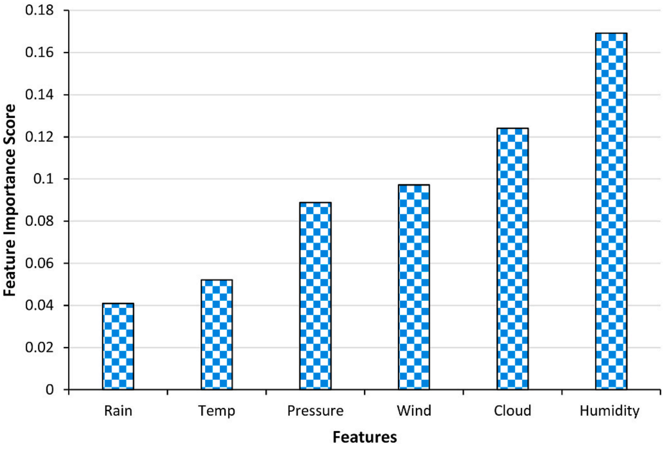
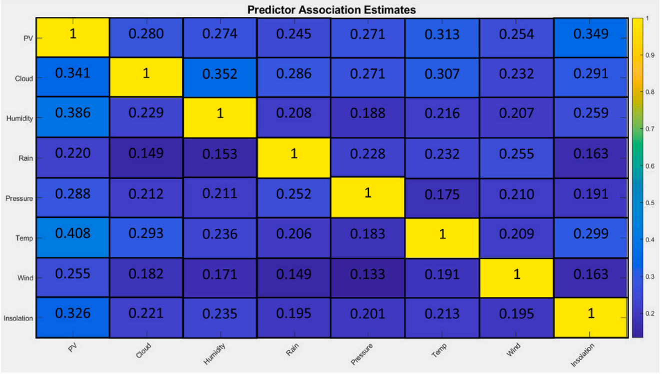


Fig. 8. The MRMR feature importance algorithm, showing the optimal selection methodology. Humidity has greatest effect on PV generation whereas rainfall has the least effect.



	PV	Cloud	Humidity	Rain	Pressure	Temp	Wind	Insolation
PV	1	0.280	0.274	0.245	0.271	0.313	0.254	0.349
Cloud	0.341	1	0.352	0.286	0.271	0.307	0.232	0.291
Humidity	0.386	0.229	1	0.208	0.188	0.216	0.207	0.259
Rain	0.220	0.149	0.153	1	0.228	0.232	0.255	0.163
Pressure	0.288	0.212	0.211	0.252	1	0.175	0.210	0.191
Temp	0.408	0.293	0.236	0.206	0.183	1	0.209	0.299
Wind	0.255	0.182	0.171	0.149	0.133	0.191	1	0.163
Insolation	0.326	0.221	0.235	0.195	0.201	0.213	0.195	1

Fig. 9. A confusion matrix showing the association between the collected variables used for training the algorithms.

Table 3
The average training and forecasting speeds across all tests.

Model	Training Speed	Forecasting Speed
RF	61.9 s	2368/s
NN	270.3 s	29821.9/s
LR	41.7 s	22802.4/s
SVM	172.4 s	8087.7/s

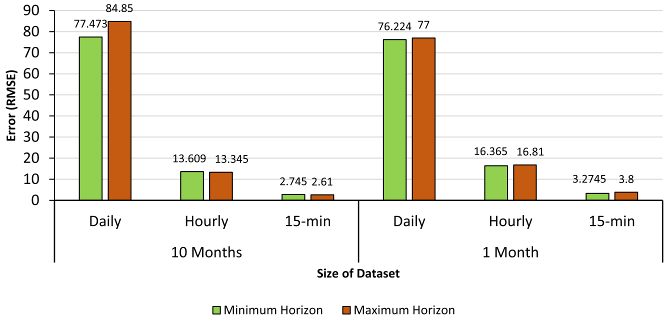


Fig. 7. The SVM forecasting results when trained with 1 month and 10 months of data.

The previous forecasting of a 10 kW PV system includes 13 inputs and 131,616 5-min samples [46]. This equates to 1.7 million datapoints. The training speeds for the RF, NN, and SVM algorithms were 1–3 h, 2–24 h, and 30–50 h respectively (see Table 3). This is due to the references research having optimised models. The models can be optimised by selecting different parameters and evaluating the results of the algorithm. For the RF, the number of trees vary from 400 to 3000 whereas in the proposed work the RF only contains 50 decision trees. The NN ran for a longer amount of time to determine the value of the weights when a minimum error is reached. The SVM contains a higher 'C' value, where the algorithm doesn't misclassify any data, requiring more computational power and time, while developing a more complex algorithm, increasing the accuracy.

LR is the simplest and thus is the quickest to train, but NN is the quickest algorithm to use once it is trained. The LR algorithm only calculates a linear relationship between the input and output variables, providing the simplest algorithm capable of forecasting the PV generation. It does not have to train like a NN does where it must optimise weights, bias, and activation functions. Instead, the relationship between the input and output variables are linearly updated at each iteration through equation (6). The NN takes the longest to train due to the complexity and each input has a weight and bias with the activation function summing the inputs together. This increase in complexity provides the capability to accurately learning the relationships between non-linear data, often providing more accurate forecasts than LR. Once the NN is trained, it retains the values of the weights, bias, and activation functions, and can forecast the PV generation output quicker than the other algorithms as it is not required to calculate anything else. The NN's used in this research vary from a 1 to 2 layers, meaning they do not need to be as complex as they have the potential to be. If more layers are added, more weights, bias, and activation functions must be calculated, and the algorithm becomes slower to train and forecast.

Among previous research, there are limited results of the time taken to train the algorithms and the forecasting speed of the algorithms once they have been trained. Evaluation of the developed algorithms for long term and short-term dataset training is shown in Fig. 9.

Interestingly, the insolation of the area was not associated highly with the PV generation. Instead, the highest value was obtained from the outdoor air temperature. This does not indicate the importance of each variable to the performance of the algorithms; however, it shows the association. This association is different from the correlation, showing any relationship between variables, linear or non-linear. The data from this is necessary to collect as this can allow accurate manipulation of the training data. For example, if the temperature and the humidity had a high association close to 1, the algorithms would probably benefit from the removal of one of the variables. This is due to them having a linear relationship, and therefore, a similar effect on the accuracy of the forecast is seen. In this case, none of the variables have a high association due to them being weather variables as the temperature can often be low while the cloud coverage is high and vice versa. The association results show that the cloud coverage and humidity have the highest association, although it is only 0.352. These have the highest MRMR feature importance scores, so those have not been removed from the algorithm.

Although the calculated association have not shown the causation of the variables from one to another, variables with less association with the target variable may be removed for a faster and potentially more accurate algorithm. The variable with the smallest association score with the PV output is the rain, and thus it may be removed. The association can be compared to the calculated feature importance scores, showing large differences. This is because the MRMR feature importance algorithm does not measure association, but actual importance of the input variable to the accuracy of the output variables' forecast, requiring more computational power. The MRMR algorithm calculates the correlation of the selected variable to the output variable and also calculates the correlation from that variable to other variables in the training dataset. This is performed for each iteration of the dataset and the results are summed to achieve a single value of correlation towards the output and a single value for correlation towards other input variables. The correlation of the variable to the output variable is then divided by the correlation between the variable and other variables of the training dataset to give a feature importance score. The multitude of calculations to achieve the feature importance score requires more computational power than calculating the association score between two features. The average RMSE forecasting horizons using the 10 months and 1 month data is presented in Fig. 10.

The NN, LR, and SVM algorithms all performed better with more data with an average RMSE increase of 10.6% when the training set was reduced. The RF performed better when given less data by 20.3%. In this research, the algorithms are not changed with dependence on the number of inputs, but they may become more complex with input data while requiring the same computational power. There are no cases in the selected previous work where the same algorithms are compared with different amounts of data when forecasting the same target. It can be concluded from this research that more data aids the algorithms' forecasting accuracy, as overall, 10 months of training data provided an error of 774.16 kW whereas 1 month of training data provided an error of 902.42 kW. Removal of predictor data does not improve any of the developed algorithms when they are trained with 1 month of data, but the performance of them is decreased when more data is removed (Fig. 11).

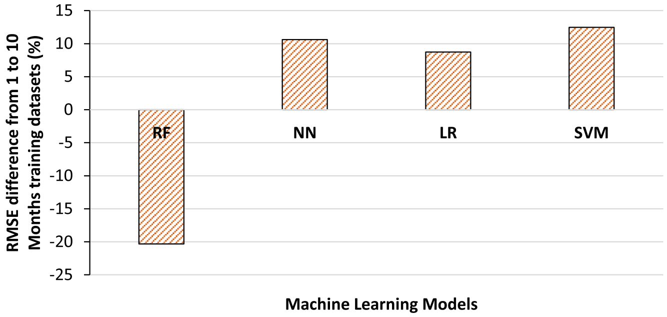


Fig. 10. The average RMSE for daily, hourly, and 15-min forecasting horizons when trained with 10 months and 1 month of data.

The average change for RF, NN, LR, and SVM are +10.94%, −5.65%, −8.77%, and −11.18% respectively. The largest decrease in error is for the NN during daily iterations at −37.25% and the largest increase in error is for the RF during 15-min iterations at +33.19%. 10 months of data is used and forecasting horizons are kept to a minimum to determine the change in accuracy. This allows there to be enough data to still train the algorithms to provide an accurate comparison on how the features affect the performance of the algorithm.

The process of calculating feature importance and optimising the MLA through the removal of features with less importance is often not included in previous literature. Although this is an important method towards improving the performance of MLA's, it is not the main achievement of much literature. The results of this study show that by measuring the importance of the input features towards the target feature(s), the algorithms can be improved, especially where there is an abundance of data.

4. Critical analysis and discussion

4.1. Strength of this study

The size of the datasets, forecast horizon, feature importance and selection, training, and prediction speed, and forecast accuracy are all evaluated and discussed comprehensively.

As each building will have various datasets, the forecasting achieved by the techniques are evaluated with different amounts of data. When provided with more data, the ML techniques performed better as a whole. This is not followed by the RF algorithm due to the method of generating the same number of decision trees regardless of the amount of data it is given. The remaining algorithms employed in this study rely more on the volume of training data, meaning was a new build and had

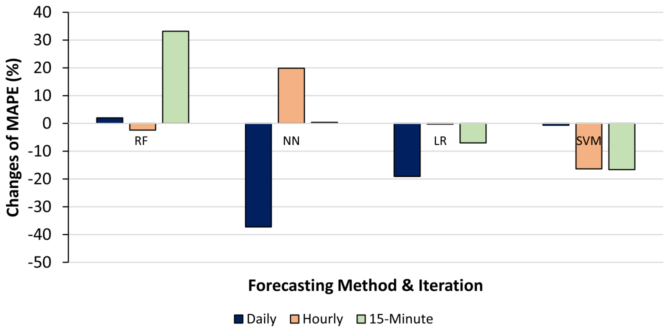


Fig. 11. The change in MAPE when the three features with the lowest MRMR are removed, showing the results for daily, hourly, and 15-min iteration averages for each method.

only collected a week or a month of climate and PV generation data, it would have to choose RF ahead of other methods. This is due to the RF model providing a higher forecasting accuracy when trained with less data. The NN, LR, and SVM's required more data to reach optimum accuracy but even when given less data, the RF had higher accuracy than LR and SVM's on average. If the building had collected upwards of 1 month of data then a NN outperforms the other methods.

Each BMS's forecast horizon requirements are different. The various forecasting horizons show that the error increased through LR and RF techniques the most when the forecasting horizon was increased. The error achieved by SVM did not change enough to warrant an absolute conclusion on the optimum amount of data. Shorter forecasting horizons provided higher accuracies in 4 of the 6 comparisons. The NN error increased with a larger horizon when trained with 10 months of data but decreased with a larger horizon when trained with 1 month of data. This shows that the NN may be more advantageous if a BMS requires a larger forecasting horizon and has a month of data to use to train the algorithm. In contrast, if the NN has 10 months of data to use for training, then a smaller forecasting horizon provides larger accuracies.

The removal of predictor features with an increase in accuracy means that the predictor no longer needs to be collected, which saves time. The NN's error decreased the most when rain, temperature, and air pressure was removed. This was unexpected as these are important climate variables for the other algorithms, but the NN was benefited from it. The error of RF increased the most from the removal of data due to the removal of a lot of data as it was for 15-min iterations. If the data is collected for less time, the RF performs the best however, the NN has the highest accuracy for more time and less features. If this method was applied to a PV system in a different climate, such as with low cloud coverage, the MRMR algorithm might provide information that cloud coverage is less important to the algorithms. In this case, input features with less importance may still be removed from the algorithms and features with more importance may be used instead, such as wind speed or any feature that may have a higher importance. The affecting factors of the PV power output remain the same. The algorithms will have similar performances with regards to the development, training, and accuracy, while the MRMR algorithm is used, and the target variable is the power output of a PV system.

The training and prediction speeds are evaluated for each algorithm to evaluate the useability of them. They were all trained in under 5 min but there was only 10 months of data used for training. If an algorithm required more data to improve accuracy, the NN would take much longer to train. LR algorithms have the simplest method and thus are trained the fastest, making them the most appealing if there's less time available or more data for training. All the forecasting speeds are above 2000 per second and BMS is not required that much data that means all forecasting speeds of the algorithms are sufficient. For 15-min forecasting for a full week, there are 672 samples, meaning it can still be trained in under half a second.

4.2. Limitation

The limitations of this study exist within the data collection and applications. A BMS could potentially employ solar PV, solar thermal, and wind, whereas this research only focuses on solar PV. Although the methodology of this research is applicable towards other types of renewable generation such as wind, solar thermal, kinetic, hydrogen, etc, this research only addresses solar PV.

Previous research shows the variability of accuracies for various sized PV forecasting in various locations. As previous research is independent and have many variables, it can be difficult to determine the cause of the different accuracies. This research only focuses on the PV forecasting in a UK climate, showing the MLA forecasting accuracy of a local PV systems' energy generation. This doesn't provide information on how the forecasting methods change with different weather conditions or sized systems.

4.3. Performance comparison with existing works

The comparison of this research against previous ML techniques applied on PV system forecasting through NN methods is shown in Table 4. The results in this research are shown in the final row. The other results show that most research is executed for forecasting a PV solar farm and not for localised systems. The accuracy of the forecasts is difficult to compare due to the different sized systems and climate they are in. The largest system is in South Africa with 75 MW peak, and the smallest is in this research for the local PV system at 30 kWh peak.

The coolest climate is in the proposed work with a temperature of 10.43 °C was still able to forecast with an error of 1.76 kW or 5.86% of the peak output power. The warmest climate is in South Africa with 19.76 °C was able to forecast with an error of 3.42% of the peak output power. Most of the available research have used NN's when forecasting a solar farm. A comparison of this research against SVM PV forecasting system forecasting is shown in Table 5.

The first reference uses many PV systems across Germany. The size of the systems is not disclosed, but the method of using SVM to forecast the PV generation have an average RMSE of 10% of the total peak power output. The systems are all rooftop, but no actual size is given. From the previous research, SVM's are used more when forecasting smaller PV systems. This may be due to the way SVM's work, as they can produce higher accuracies when they are given less input variables. Buildings may not prioritise the collection of data as they have other functions. The comparison of this research against previous MLA PV system forecasting through RF methods are shown in Table 6.

The second reference uses various sized systems across the Netherlands, hence the measurement of error that is in W/kW peak. This allows the error to be divided by the size of the system to give a universal error to compare the different sized systems. Most of the previous work that uses RF models apply it to various different systems under the same weather conditions to try to produce a general model. This means that can be used by anybody within a defined distance, under the same weather conditions. The first reference in Table 6 has an error of 11.77% whereas the proposed systems' RF error is 6.6%. As the size of the systems in the second reference from Table 6 vary, the error can be difficult to compare, but the error can be converted to a percentage to compare against the proposed work. This equates to 1.64% error on average. The outdoor air temperature, system size, and algorithm are very similar, leading to the probability that the first reference has less or poorer quality data for training. The second reference has 17 input variables to train the algorithm with, allowing an accurate forecast. The performance comparison of this research against previous MLA PV forecasting systems through LR methods are shown in Table 7.

Gaussian Regression is a more complex variation of LR, capable of producing higher accuracies at the expense of more computational power. The second and third reference in Table 7 have multiple PV systems that are all stated to be rooftop installed. Although the size isn't stated, it can be assumed that they are of a similar size to the system in

Table 4
A comparison of current research against previous research when forecasting PV energy generation through MLA methods [31,34,35,41].

Location and Average Outdoor Air Temperature °C	System Size	Performance	Method
South Africa. 19.76 [18].	75 MW Peak	3.42% MAPE	Neural Network
Belgium. 11.73 [20].	368 kWh Peak	11.89% RMSE	Neural Network
Italy. 15.04 [21].	327 kWh Peak	3.41% RMSE	Neural Network
Peru. 23.12 [43].	700W Peak	10.57% MAPE	Neural Network
Proposed work. 10.43	30 kWh Peak	1.76 kW RMSE	Neural Network

Table 5
A comparison of current research against previous research when forecasting PV energy generation through SVM MLA's.

Location and Average Outdoor Air Temperature °C	System Size	Performance	Method
Germany. 9.88 [51]	NA	7.5–12.5% RMSE	SVM
Peru. 23.12 [43].	700W Peak	8.81% MAPE	SVM
Denmark. 9.57 [46].	10 kW Peak	13.04% RMSE	SVM
Mongolia. 14.49 [40].	30 MW Peak	3.86% RMSE	SVM
Proposed work. 10.43	30 kWh Peak	2.61 kW RMSE	SVM

Table 6
A comparison of current research against previous research when forecasting PV energy generation through RF MLA's.

Location and Average Outdoor Air Temperature °C	System Size	Performance	Method
Denmark.9.57 [46].	10 kW Peak	11.77% RMSE	RF
Netherlands.11.3 [45]	0.5–6.8 kW Peak	0.06kW/kWp	RF
America. 8.04 [44].	NA	8.32% RMSE	RF
Portugal. 21 [52].	2 kWh Peak	26.04% RMSE	RF
Proposed work.	30 kWh Peak	1.98 kWh RMSE	RF

Table 7
A comparison of current research against previous research when forecasting PV energy generation through LR MLA's.

Location and Average Outdoor Air Temperature °C	System Size	Performance	Method
America. 8.04 [41].	30 MW Peak	1.23 MW RMSE	Gaussian Regression
Netherlands. 11.3 [45]	NA	0.075kW/kWp	LR
America. 8.04 [44].	NA	8.95% RMSE	Gaussian Regression
Chile. 17.3 [45].	3 kWh Peak	1.87% RMSE	Multiple Linear Regression
Proposed work	30 kWh Peak	5.31 kWh RMSE	LR

the proposed work. The second reference has a RMSE% of 8.95% whereas the LR model in this research has a RMSE% of 17.7%. This is due to the model used, as the size of the system and the average outdoor temperature are similar.

The correlation between the temperatures and the performance of previous research has a value of 0.137, showing there is little correlation. The correlation between the size of the systems and the performance is −0.4, showing there is a moderate negative correlation between the size of the system and the accuracy. Bigger PV systems have higher accuracies when forecasting the PV generation. The main differences between sizes include the number of PV panels and the amount of collected data, where larger systems have more data, and can train algorithms with higher accuracies. Another consideration is faults within the PV system, as a building has other priorities to solve, whereas a PV farm may repair any faults quicker, and thus, maintain the quality of the training data. There is no definitive algorithm that can forecast with the highest accuracy while requiring the least data for both PV farms and local systems. As previous algorithms are trained with data for different sized and located PV systems, there cannot be an accurate comparison between methods.

5. Conclusions

The proposed methodology developed in this work has compared benchmark within machine learning algorithms for forecasting a localised photovoltaic system and to demonstrate the benefits of each method. Data collection and processing is important for the performance of the algorithms; however, it is not crucial for all of them. The quality of data improves the accuracy of the algorithms more than quantity of data. Rain, temperature, and air pressure have the least effect on the performance of the algorithms when forecasting PV generation in a UK climate. This demonstrates the effectiveness of selecting the correct algorithm for the amount and type of data within the dataset. In this research, 64 models are created with 4 algorithms and tested with a variety of sizes for training data and horizons. Random forest algorithms produced the lowest error with an average RMSE of 32 and it required less data to successfully train the algorithms. This is due to the algorithm generating a set amount of decision trees regardless of the amount of data it is given. This means that it may not be able to include all the data it is provided with for training. The incorrect datapoints that the processing algorithm cannot remove are still included in the training set which can reduce the accuracy of the algorithm. These include data that isn't an outlier or missing. The neural networks showed the highest RMSE at 38.8 however it had a lower error on 7 of the 12 datasets it was trained and validated with compared to the RF with the lowest error on only 3 of the 12 datasets. The linear regression algorithms trained the fastest on average at 41.7 s per model. Overall, RF provided the highest accuracy among the tested algorithms in this research work while requiring the least amount of training data. As more buildings are aiming to reduce the carbon emissions, RE is becoming more popular and thus, the forecasting of RE systems is necessary. The forecasting of these systems can be implemented into BMS's and can be used to reduce the carbon emissions of the building. Currently, there is no comparison of MLA's when forecasting the PV generation of a rooftop system. This research provides information on how they can be improved through the MRMR algorithm and how much data they require for an optimal model. The information provided in this research can be used as a guide to how much data and what variables are needed for training, and what algorithms to use when forecasting the PV generation of a rooftop system.

Credit author statement

Connor Scott: Conceptualization, Methodology, Validation, Formal analysis, Data Curation, Writing - Original Draft, Preparation, Writing - Review & Editing, **Mominul Ahsan:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Original Draft, Preparation, Writing - Review & Editing, Supervision. **Alhuessien Albarbar**: Conceptualization, Validation, Formal analysis, Investigation, Writing - Original Draft, Preparation, Writing - Review & Editing, Visualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Business energy statistical summary [Online] Available, https://www.gov.uk/government/publications/business-energy-statistical-summary; 2018.
[2] Net zero strategy: build back greener [Online] Available, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1033990/net-zero-strategy-beis.pdf; 2021.

[3] Ahmed SU, Ali A, Memon A. Renewable energy's reliability issue and possible solutions: a meta-analytic review. J Inf Commun Technol 2018;2(3):170–5.

[4] Gao C, Zhu S, An N, Na H, You H, Gao C. Comprehensive comparison of multiple renewable power generation methods: a combination analysis of life cycle assessment and ecological footprint. Renew Sustain Energy Rev 2021/09/01/2021; 147:111255. https://doi.org/10.1016/j.rser.2021.111255.

[5] Lu Y, Khan ZA, Alvarez-Alvarado MS, Zhang Y, Huang Z, Imran M. A critical review of sustainable energy policies for the promotion of renewable energy sources. Sustainability 2020;12(12). https://doi.org/10.3390/su12125078.

[6] Grid N. "Annual report and accounts 2020/2021,". National Grid 2021. 2021.

[7] Stolworthy M. "GB Fuel type power generation production.". GridWatch 2021. https://gridwatch.co.uk/. [Accessed 26 November 2021].

[8] Eini R, Linkous L, Zohrabi N, Abdelwahed S. "Smart building management system: performance specifications and design requirements,". J Build Eng 2021;39: 102222. https://doi.org/10.1016/j.jobe.2021.102222. 2021/07/01.

[9] Elnour M, et al. "Performance and energy optimization of building automation and management systems: towards smart sustainable carbon-neutral sports facilities,". Renew Sustain Energy Rev 2022;162:112401. https://doi.org/10.1016/j.rser.2022.112401. 2022/07/01.

[10] Mariano-Hernández D, Hernández-Callejo L, Zorita-Lamadrid A, Duque-Pérez O, Santos García F. A review of strategies for building energy management system: model predictive control, demand side management, optimization, and fault detect & diagnosis. J Build Eng 2021;33:101692. https://doi.org/10.1016/j.jobe.2020.101692. 01/01/2021.

[11] Chaouch H, Çeken C, Arı S. "Energy management of HVAC systems in smart buildings by using fuzzy logic and M2M communication,". J Build Eng 2021;44: 102606. https://doi.org/10.1016/j.jobe.2021.102606. 2021/12/01.

[12] Farinis GK, Kanellos FD. "Integrated energy management system for Microgrids of building prosumers,". Elec Power Syst Res 2021;198:107357. https://doi.org/10.1016/j.epsr.2021.107357. 2021/09/01.

[13] Salerno I, Anjos MF, McKinnon K, Gómez-Herrera JA. "Adaptable energy management system for smart buildings,". J Build Eng 2021;44:102748. https://doi.org/10.1016/j.jobe.2021.102748. 2021/12/01.

[14] Alanne K, Sierla S. An overview of machine learning applications for smart buildings. Sustain Cities Soc 2022;76:103445. https://doi.org/10.1016/j.scs.2021.103445. 2022/01/01.

[15] Brandi S, Gallo A, Capozzoli A. "A predictive and adaptive control strategy to optimize the management of integrated energy systems in buildings,". Energy Rep 2022;8:1550–67. https://doi.org/10.1016/j.egyr.2021.12.058. 2022/11/01.

[16] Yang Y, Bremner S, Menictas C, Kay M. "Forecasting error processing techniques and frequency domain decomposition for forecasting error compensation and renewable energy firming in hybrid systems,". Appl Energy 2022;313:118748. https://doi.org/10.1016/j.apenergy.2022.118748. 2022/05/01.

[17] Gaviria JF, Narváez G, Guillen C, Giraldo LF, Bressan M. "Machine learning in photovoltaic systems: a review,". Renewable Energy; 2022. https://doi.org/10.1016/j.renene.2022.06.105. 2022/07/01.

[18] du Plessis AA, Strauss JM, Rix AJ. "Short-term solar power forecasting: investigating the ability of deep learning models to capture low-level utility-scale Photovoltaic system behaviour,". Appl Energy 2021;285:116395. https://doi.org/10.1016/j.apenergy.2020.116395. 2021/03/01.

[19] Huang Q, Wei S. "Improved quantile convolutional neural network with two-stage training for daily-ahead probabilistic forecasting of photovoltaic power,". Energy Convers Manag 2020;220:113085. https://doi.org/10.1016/j.enconman.2020.113085. 2020/09/15.

[20] Kaffash M, Bruninx K, Deconinck G. Data-driven forecasting of local PV generation for stochastic PV-battery system management. Int J Energy Res 2021;45(11): 15962–79. https://doi.org/10.1002/er.6826.

[21] Polimeni S, Nespoli A, Leva S, Valenti G, Manzolini G. "Implementation of different PV forecast approaches in a MultiGood MicroGrid: modeling and experimental results,". Processes 2021;9(2). https://doi.org/10.3390/pr9020323.

[22] Perera M, De Hoog J, Bandara K, Halgamuge S. "Multi-resolution, multi-horizon distributed solar PV power forecasting with forecast combinations,". Expert Syst Appl 2022;205:117690. https://doi.org/10.1016/j.eswa.2022.117690. 2022/11/01.

[23] Li Z, et al. "Investigation on the all-day electrical/thermal and antifreeze performance of a new vacuum double-glazing PV/T collector in typical climates — compared with single-glazing PV/T,". Energy 2021;235:121230. https://doi.org/10.1016/j.energy.2021.121230. 2021/11/15.

[24] Joshi SS, Dhoble AS. "Photovoltaic -Thermal systems (PVT): technology review and future trends,". Renew Sustain Energy Rev 2018;92:848–82. https://doi.org/10.1016/j.rser.2018.04.067. 2018/09/01.

[25] Reddy S, Mallick T, Chemisana D. "Solar power generation,". Int J Photoenergy 2013;2013. https://doi.org/10.1155/2013/950564. 01/01.

[26] Tzivanidis C, Bellos E. "Solar energy utilization in buildings,". 2018. p. 119–65.

[27] Wang W, Yuan B, Sun Q, Wennersten R. "Application of energy storage in integrated energy systems — a solution to fluctuation and uncertainty of renewable energy,". J Energy Storage 2022;52:104812. https://doi.org/10.1016/j.est.2022.104812. 2022/08/01.

[28] Dreher A, et al. "AI agents envisioning the future: forecast-based operation of renewable energy storage systems using hydrogen with Deep Reinforcement Learning,". Energy Convers Manag 2022;258:115401. https://doi.org/10.1016/j.enconman.2022.115401. 2022/04/15.

[29] Sierla S, Pourakbari-Kasmaei M, Vyatkin V. "A taxonomy of machine learning applications for virtual power plants and home/building energy management systems,". Autom ConStruct 2022;136:104174. https://doi.org/10.1016/j.autcon.2022.104174. 2022/04/01.

[30] Bao-ying W, Yu X, Majd JK. "A novel forecasting method based on the economic and demand response for FC/WT/PV unit and a 3 in 1 TES energy storage,". Int J Hydrogen Energy 2021;46(65):32995–3009. https://doi.org/10.1016/j.ijhydene.2021.07.074. 2021/09/21.

[31] Zsiborács H, Pintér G, Vincze A, Birkner Z, Baranyai NH. "Grid balancing challenges illustrated by two European examples: interactions of electric grids, photovoltaic power generation, energy storage and power generation forecasting,". Energy Rep 2021;7:3805–18. https://doi.org/10.1016/j.egyr.2021.06.007. 2021/11/01.

[32] Georga EI, Fotiadis DI, Tigas SK. "6 - nonlinear models of glucose concentration,". In: Georga EI, Fotiadis DI, Tigas SK, editors. Personalized predictive modeling in type 1 diabetes. Academic Press; 2018. p. 131–51.

[33] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. "Deep learning for time series classification: a review,". Data Min Knowl Discov 2019;33(4):917–63. https://doi.org/10.1007/s10618-019-00619-1. 2019/07/01.

[34] Bongiorno V, Gibbon S, Michailidou E, Curioni M. "Exploring the use of machine learning for interpreting electrochemical impedance spectroscopy data: evaluation of the training dataset size,". Corrosion Sci 2022;198:110119. https://doi.org/10.1016/j.corsci.2022.110119. 2022/04/15.

[35] Wen Y, et al. Performance evaluation of probabilistic methods based on bootstrap and quantile regression to quantify PV power point forecast uncertainty. IEEE Transact Neural Networks Learn Syst 2020;31(4):1134–44. https://doi.org/10.1109/tnnls.2019.2918795.

[36] Jebli I, Belouadha F-Z, Kabbaj MI, Tilioua A. "Prediction of solar energy guided by pearson correlation using machine learning,". Energy 2021;224:120109. https://doi.org/10.1016/j.energy.2021.120109. 2021/06/01.

[37] Markovics D, Mayer MJ. "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction,". Renew Sustain Energy Rev 2022;161:112364. https://doi.org/10.1016/j.rser.2022.112364. 2022/06/01.

[38] VanDeventer W, et al. "Short-term PV power forecasting using hybrid GASVM technique,". Renew Energy 2019;140:367–79. https://doi.org/10.1016/j.renene.2019.02.087. 2019/09/01.

[39] Carriere T, Vernay C, Pitaval S, Kariniotakis G. "A novel approach for seamless probabilistic photovoltaic power forecasting covering multiple time frames,". IEEE Trans Smart Grid 2020;11(3):2281–92. https://doi.org/10.1109/tsg.2019.2951288.

[40] Wang F, Zhen Z, Wang B, Mi Z. "Comparative study on KNN and SVM based weather classification models for day ahead short term solar PV power forecasting,". Appl Sci 2018;8(1):28 [Online]. Available: https://www.mdpi.com/2076-3417/8/1/28.

[41] Najibi F, Apostolopoulou D, Alonso E. "Enhanced performance Gaussian process regression for probabilistic short-term solar output forecast,". Int J Electr Power Energy Syst 2021;130:106916. https://doi.org/10.1016/j.ijepes.2021.106916. 2021/09/01.

[42] Moradzadeh A, Mansour-Saatloo A, Mohammadi-Ivatloo B, Anvari-Moghaddam A. "Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings,". Appl Sci 2020;10(11):3829 [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3829.

[43] Arias Velásquez RM. "A case study of NeuralProphet and nonlinear evaluation for high accuracy prediction in short-term forecasting in PV solar plant,". Heliyon 2022:e10639. https://doi.org/10.1016/j.heliyon.2022.e10639. 2022/09/16.

[44] Liu D, Sun K. "Random forest solar power forecast based on classification optimization,". Energy 2019;187:115940. https://doi.org/10.1016/j.energy.2019.115940. 2019/11/15.

[45] Visser L, AlSkaif T, van Sark W. "Operational day-ahead solar power forecasting for aggregated PV systems with a varying spatial distribution,". Renew Energy 2022; 183:267–82. https://doi.org/10.1016/j.renene.2021.10.102. 2022/01/01.

[46] Pombo DV, Bacher P, Ziras C, Bindner HW, Spataru SV, Sørensen PE. "Benchmarking physics-informed machine learning-based short term PV-power forecasting tools,". Energy Rep 2022;8:6512–20. https://doi.org/10.1016/j.egyr.2022.05.006. 2022/11/01.

[47] Voyant C, et al. "Machine learning methods for solar radiation forecasting: a review,". Renew Energy 2017;105:569–82. https://doi.org/10.1016/j.renene.2016.12.095. 2017/05/01.

[48] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2021.

[49] Ding C, Peng H. "Minimum redundancy feature selection from microarray gene expression data," (in eng). J Bioinf Comput Biol 2005;3(2):185–205. https://doi.org/10.1142/s0219720005001004.

[50] Probst P, Wright MN, Boulesteix A-L. "Hyperparameters and tuning strategies for random forest,". WIREs Data Mining and Knowledge Discovery 2019;9(3):e1301. https://doi.org/10.1002/widm.1301.

[51] Wolff B, Kühnert J, Lorenz E, Kramer O, Heinemann D. Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. Sol Energy 2016/10/01/2016;135:197–208. https://doi.org/10.1016/j.solener.2016.05.051.

[52] Huertas Tato J, Centeno Brito M. "Using smart persistence and random forests to predict photovoltaic energy production,". Energies 2019;12(1):100 [Online]. Available: https://www.mdpi.com/1996-1073/12/1/100.