

Research Paper



Forecasting capacitor banks for improving efficiency of grid-integrated PV plants: A machine learning approach

Saurabh Kumar Rajput^a, Deepansh Kulshrestha^b, Nikhil Paliwal^b, Vivek Saxena^c,
Saibal Manna^c, Mohammed H. Alsharif^{d,*}, Mun-Kyeom Kim^{e,*}

^a Centre for Internet of Things, MITS Gwalior- Deemed University, Gwalior, India

^b Electrical Engineering Department, MITS Gwalior-Deemed University, Gwalior, India

^c Department of Electrical & Computer Engineering, ABES Engineering College, Ghaziabad, UP, India

^d Department of Electrical Engineering, College of Electronics and Information Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, South Korea

^e School of Energy System Engineering, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea

ARTICLE INFO

Keywords:

Grid-connected PV systems
Power factor
Capacitor bank
Predictive modeling
Machine learning

ABSTRACT

Grid-connected rooftop PV systems are becoming more popular to promote renewable energy. The rooftop PV may diminish the system's energy efficiency by lowering the power factor (PF) on the grid side. The current work provides a machine learning approach that estimates the necessary capacitor banks to boost the PF to unity, enabling proactive remedial action for energy savings. Various machine learning models, such as linear regression, ridge regression, lasso regression, random forest, decision tree, XGBoost, Adaboost, and gradient boosting, are evaluated to improve the system's efficiency. The best model is Lasso Regression, which produces a high R^2 score of 0.89 with low MSE and MAPE values. The model is based on real-time data collected from a 100 kWp PV plant connected to an 11 kV grid supply and an institutional building load. The model undergoes validation by implementing the forecasted capacitor banks. According to the findings, a 10.60 kVAR-rated shunt capacitor is required to maintain the PF at unity and save an average of 1673.52 kWh of energy per month. This work highlights the necessity of implementing Lasso regression in energy management systems to improve PF, decrease electricity costs, and reduce environmental impacts.

1. Introduction

Photovoltaic (PV) plants are clean and environmentally friendly energy sources that generate electricity from sunlight (Gusman et al., 2020a). PV plants may function in autonomous mode, supplying energy to individual residences, and they can also integrate with the grid (Alsharif et al., 2021). The integration of PV plants with electrical grids

has several advantages and challenges (Cortes-Vega and Alazki, 2021). The effects of PV on distribution networks, such as power loss, reverse power flow, and voltage changes, have been carefully studied, and the findings highlight the need for advanced optimization methods to address the challenges associated with expanding PV-grid integration (Alsharif et al., 2024, 2016). For increasing load with rooftop PV systems, a multi-objective strategy using the Gravitational Search

Abbreviations: ADC, Analytics-driven Control; AC, Alternating Current; AI, Artificial Intelligence; APFC, Automatic Power Factor Correction; BESS, Battery Energy Storage Systems; BTC, Trivial Battery Control; DC, Direct Current; DG, Distributed Generation; DT, Decision Tree; DU, Deemed University; GSA, Gravitational Search Algorithm; GB, Gradient Boosting; IoT, Internet of Things; ISC, Short Circuit Current; KVA, Kilovolt-Ampere; KVAh, Kilovolt-Ampere Hour; kWp, Kilowatt Peak the maximum power output in kilowatts; KWh, Kilowatt-hour; LV, Low Voltage; LR, Linear Regression; LSTM, Long Short-Term Memory; MAE, Mean Absolute Error; MAPE, Mean Absolute Percentage Error; MITS, Madhav Institute of Technology & Science; MSE, Mean Squared Error; ML, Machine Learning; MLC, Machine Learning-based Control; MPPT, Maximum Power Point Tracking; NWP, Numerical Weather Prediction; PF, Power Factor; PPC, Peak Power Curtailment; RF, Random Forest; RMSE, Root Mean Squared Error; RPC, Reactive Power Control; R^2 , R-squared Coefficient of Determination; RNN, Recurrent Neural Network; RSCAD, Research Condition and Disease Categorization; SASM, Smart Adaptive Switching Module; SCB, Shunt Capacitor Bank; SVM, Support Vector Machine; THD, Total Harmonic Distortion; VRE, Variable Renewable Energy.

* Corresponding authors.

E-mail addresses: saurabh9march@mitsgwalior.in (S.K. Rajput), nikhil7@mitsgwalior.in (N. Paliwal), malsharif@sejong.ac.kr (M.H. Alsharif), mkim@cau.ac.kr (M.-K. Kim).

<https://doi.org/10.1016/j.egy.2024.12.011>

Received 23 September 2024; Received in revised form 2 December 2024; Accepted 3 December 2024

Available online 9 December 2024

2352-4847/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Algorithm (GSA) may be employed to effectively manage seasonal changes in system performance (Roy et al., 2023). A wide range of methods for forecasting Variable Renewable Energy (VRE) generation, focusing primarily on solar PV systems, has been examined. The study combined the use of Numerical Weather Prediction (NWP) models, satellite image-based techniques, and machine learning methodologies to conduct a thorough assessment on both site-level and regional aggregate forecasts for India. The study has illustrated the need for better infrastructure and regulatory frameworks to provide reliable forecasting of VRE. The research studies underline the importance of improved forecasting for maintaining grid security and provide suggestions for infrastructure development to optimize gains from integrating renewable energy in India (Alsharif, 2017; Gámez Medina et al., 2022). A hybrid renewable energy system with a Smart Adaptive Switching Module (SASM) is investigated. The system is based on sensor and meteorological data and utilizes big data technology, with both Support Vector Machine (SVM) and Linear Regression (LR) techniques to handle the massive volume of textual input. Adding more parameters to Support Vector Machine (SVM) significantly enhances its performance in terms of Root Mean Squared Error (RMSE), as demonstrated by the study. The approach also showed how quickly and efficiently PV production and power demand could be balanced, demonstrating the system design's ability to manage renewable energy resources effectively (Preda et al., 2018). In a grid-connected system, multi-phase voltage-based inverters are studied for the purpose of reducing output current Total Harmonic Distortion (THD). The study used Research Condition and Disease Categorization (RSCAD) software and carried out several experiments based on simulations to investigate the grid-side effects of solar PV integration. This work focused on analyzing the flow of various harmonic distortions, voltage, and current imbalances for different levels of PV penetration (Mitra et al., 2022). An exhaustive survey was conducted on prediction techniques of renewable energy generation particularly emphasis is given to conventional Machine Learning (ML) and Deep Learning (DL) methodologies. The study analysed linear regression, random forests, Support Vector Machines (SVMs), and XGBoost (a gradient boosting algorithm) and time-series forecasting methods including autoregressive model, moving averages model and a specific type of neural network called an Recurrent Neural Network (RNN). Subsequently investigated hybrid models that combine time-series analysis with Machine Learning (ML) and Deep Learning (DL) techniques to improve prediction performance. Overall, the authors of this review find that ML and DL models provide generalizable methods for making accurate forecasts; yet key challenges such as data quality complexities or model interpretability questions arising from non-linearity in high-dimensional multimodal renewable energy sources still need to be addressed (Benti et al., 2023). Machine Learning-based Control (MLC) solution is developed for enhancing PV hosting capacity and mitigating congestion in low-voltage networks. The study employed Long Short-Term Memory (LSTM) algorithm for day-ahead PV forecasting to manage residential Battery Energy Storage Systems (BESS). The MLC approach was compared with various techniques, including Analytics-driven Control (ADC), Trivial Battery Control (BTC), Peak Power Curtailment (PPC), and Reactive Power Control (RPC). The study, conducted on a real-life Low Voltage (LV) distribution system with 15 households and different PV scenarios, found that the MLC approach reduced the overvoltage hours by 30 % compared to ADC and by 4 % compared to BTC under high PV installations. The techno-economic analysis revealed that MLC could lead to 8–22 % savings for customers compared to BTC and potentially lower energy costs from utilities (Shabbir et al., 2024).

Further research works are also conducted in the field of PV integration with Low Voltage (LV) distribution power systems that are environmentally sustainable and focus on improving energy efficiency (Palahalli et al., 2022). Optimizing Distributed Generation (DG) and Shunt Capacitor Banks (SCBs) is necessary to maximize power system usage and minimize the demand for additional infrastructure. The study

highlights the need for research on optimization techniques and practical deployment for techno-economic evaluation of simultaneous Distributed Generation (DG) and Shunt Capacitor Bank (SCB) integration with reduced power losses (Leghari et al., 2022). Smaili et al. examined the methods for combining Shunt Capacitor Banks (SCBs) with PV modules to increase voltage stability and lower energy-dissipation losses in distribution networks. Their investigation demonstrated a notable advancements in voltage profile augmentation and power loss reduction (Smaili et al., 2024). Understanding the larger context of PV integration and the function of capacitor banks is made possible by their insights into operational, technical, and financial issues (Shafiullah et al., 2022). The reliability-related concerns for PV systems, with a specific emphasis on the impact of heat stress on DC-link capacitors shows that the model-based reliability research is crucial to take heat loading into account when designing solar systems since doing so is required to guarantee long-term performance (Vernica et al., 2018). For PF management, multifunctional PV inverters and conventional capacitor banks are compared and research indicates that even when multifunctional inverters provide accurate reactive power management, they may still lessen system dependability. This highlights how important meticulous planning and thorough investment research are (Gusman et al., 2020b). A method of building capacitor banks in conjunction with PV systems to maintain voltage stability is proposed for improved system performance and decreased unpredictability, providing a feasible means of increasing grid-integrated PV systems' efficiency and reliability (Kalyuzhny et al., 2013).

One of the major obstacles to PV and grid integration is the impact of rooftop PV systems on grid supply PF. Low PF reduces the system's efficiency by increasing reactive power inputs and additional power losses. The primary goal of this study is to develop an advanced machine learning-based solution that forecasts the PF of grid-connected rooftop PV systems, enabling preventive action to ensure energy-efficient operation.

The work flow of present study is as follows-

1. Real-time data collection from a 100 kWp rooftop PV plant integrated with a 33 kV grid line under the composite climate of central India.
2. The development of various machine learning models to predict the reduction in PF and to identify the best-fitting, error-free model among all.
3. Utilizing the identified machine learning model to determine the necessary SCB for grid-connected PV plant, with the aim of reducing energy consumption.

2. System description

MITS-DU, Gwalior, uses a 100 kWp rooftop PV plant that has been developed to harness as much from solar electricity as possible while integrating with conventional backup systems for high utility levels. This setup can make the university to have state-of-the-art with regard environmental sustainability as well, without unexpected cost saving efforts. The 100 kWp PV plant have 308 PV panels, each rated to generate 325 Wp (Watt-peak) under standard test conditions. The PV panels converts sunlight using the photovoltaic effect into DC electricity. The output from the panels are then converted into AC power using two 50 kW inverters. The inverters ensure that the AC output is at correct voltage and frequency to be connected with your existing electricity infrastructure while adhering to grid requirements. The schematic diagram of university distribution power system and 100 kWp PV plant is shown in Figs. 1 and 2 respectively.

From the grid supply, the AC power is received at main substation. By using two transformers (500 kVA and 315 kV), the grid supply voltage is stepped down from 33 kV to 440 V, a safe and feasible level appropriate for campus distribution network. Subsequently, this power is continuously supplied to the campus buildings. For the backup, the

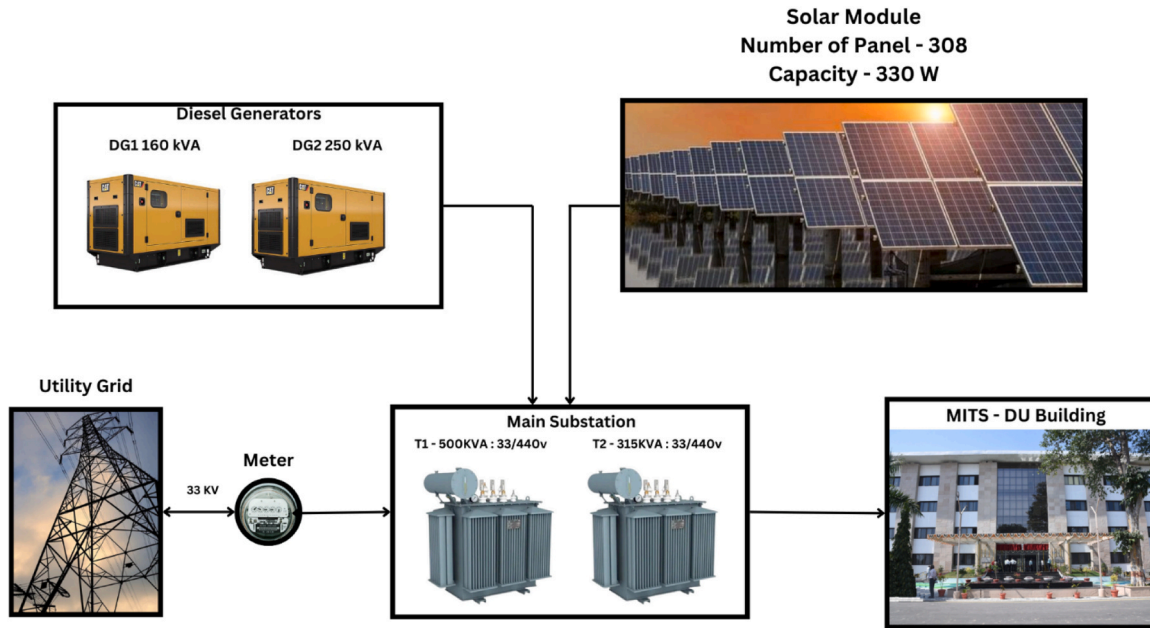


Fig. 1. Schematic diagram of hybrid system.

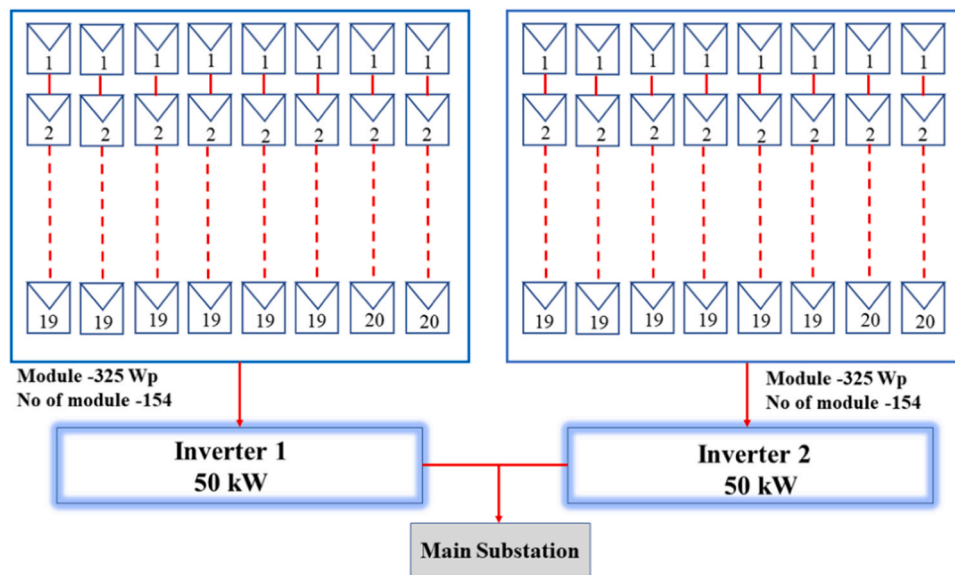


Fig. 2. Schematic representation of the 100 kWp rooftop solar power plant at MITS, Gwalior.

system also uses diesel generators (DG1 160 kVA & DG2 250 kVA) to address intermittency and grid discharge issues. As shown in Fig. 2, there are 2 PV arrays, each rated for 50 kWp. The total number of PV strings in each PV array is 8. There are 6 PV strings, each with 19 PV modules connected in series, and 2 PV strings, each with 20 PV modules connected in series. The installation company made these connections to meet the design requirements of the PV inverter and PV plant.

- **PV Inverter Specification:** As per the specification sheet of the PV inverter, there are 4 MPPT (Maximum Power Point Tracking) channels and each channel can carry 2 PV strings. Also, each MPPT channel is rated for a maximum 28 A input current and a maximum 950 V voltage (Appendix 1: Specification details of PV inverter).
- **PV Array Specification:** As shown in Fig. 2, there are total 154 PV modules in 50 kWp PV array and each module is rated for short

circuit current (I_{sc}) of 9.24 A and open circuit voltage of 46.3 V (Appendix 2: Specification details of PV module).

- **Electrical connection of PV inverter with PV array:** In order to meet the MPPT channel's connection requirement, the 154 modules are divided into 4 pairs. For maintaining the same voltage level in a parallel connection, each string of a pair needs to have an equal number of PV modules. Hence, there are 3 pairs (6 PV strings) having 19 PV modules each and 1 pair (2 PV strings) having 20 PV modules each. Here 1 PV string (having 19 modules) will have a total voltage of 897.7 V and current of 9.24 A, which matches the PV inverter's MPPT rating requirement of 950 V and 28 A. Similarly, the PV string (having 20 modules) will have a total voltage of 926.0 V and a current of 9.24 A, which also matches the PV inverter's MPPT rating requirement of 950 V and 28 A.

In this manner, the PV modules are arranged to form an array of 154

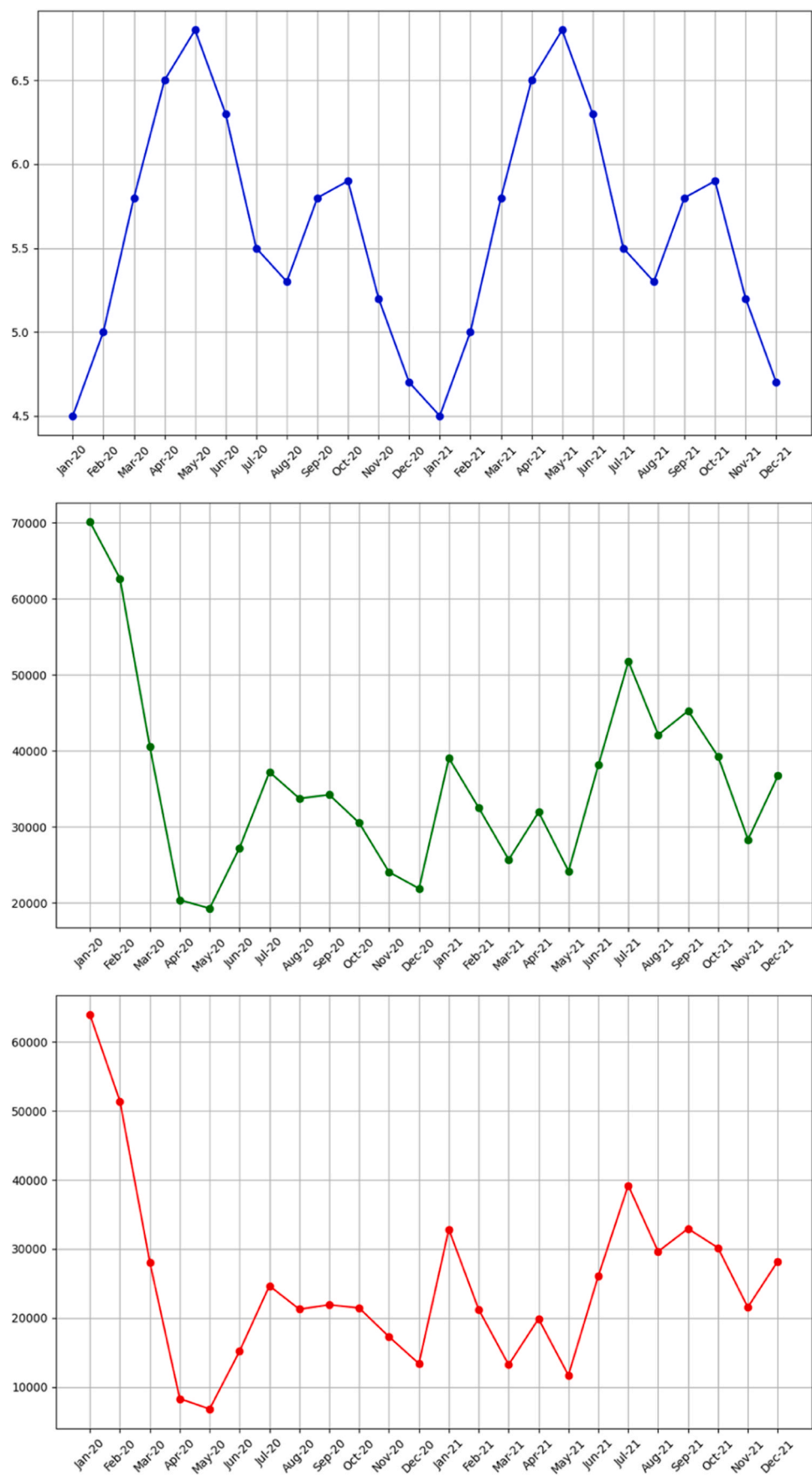


Fig. 3. Data collection process from the system design.

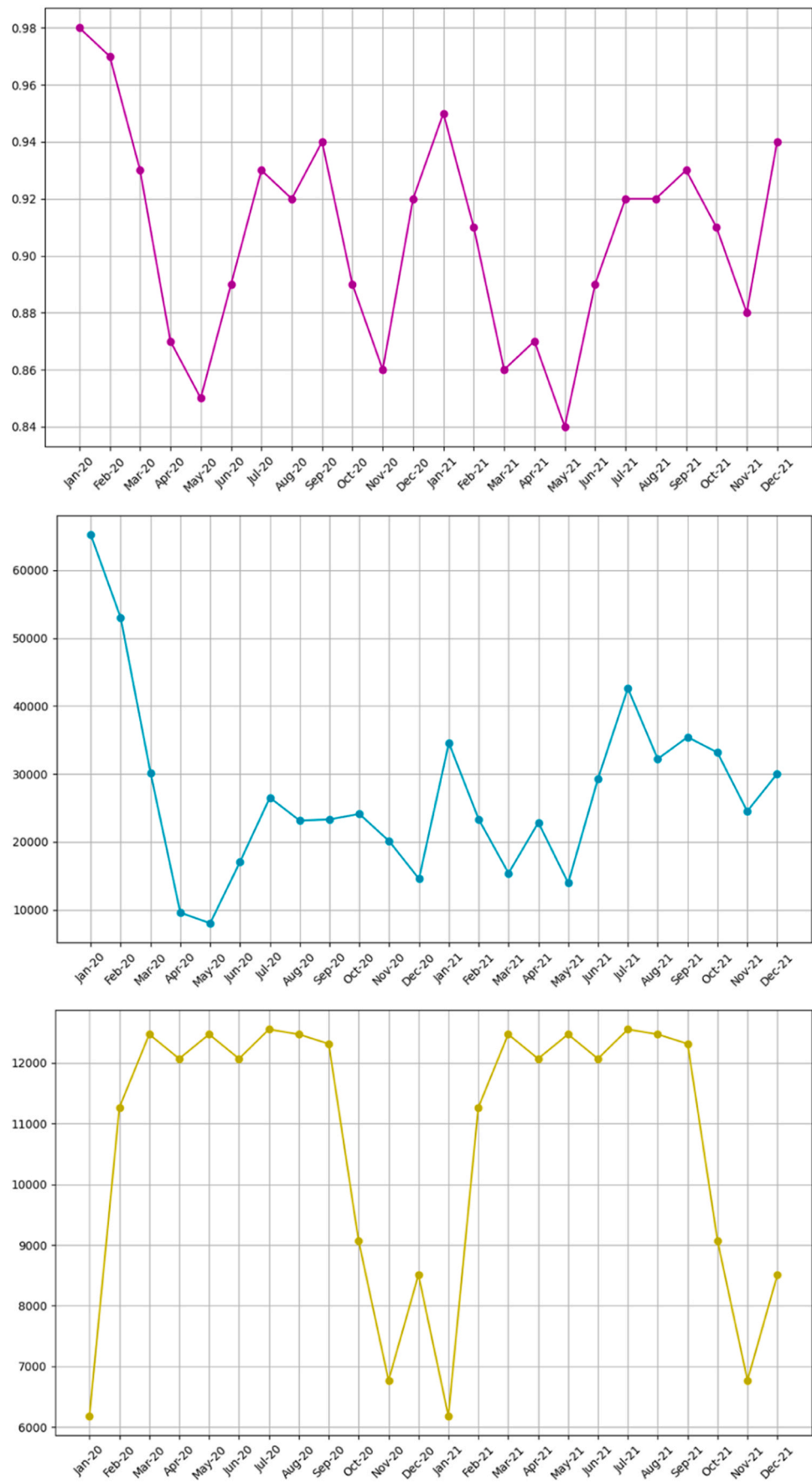


Fig. 3. (continued).

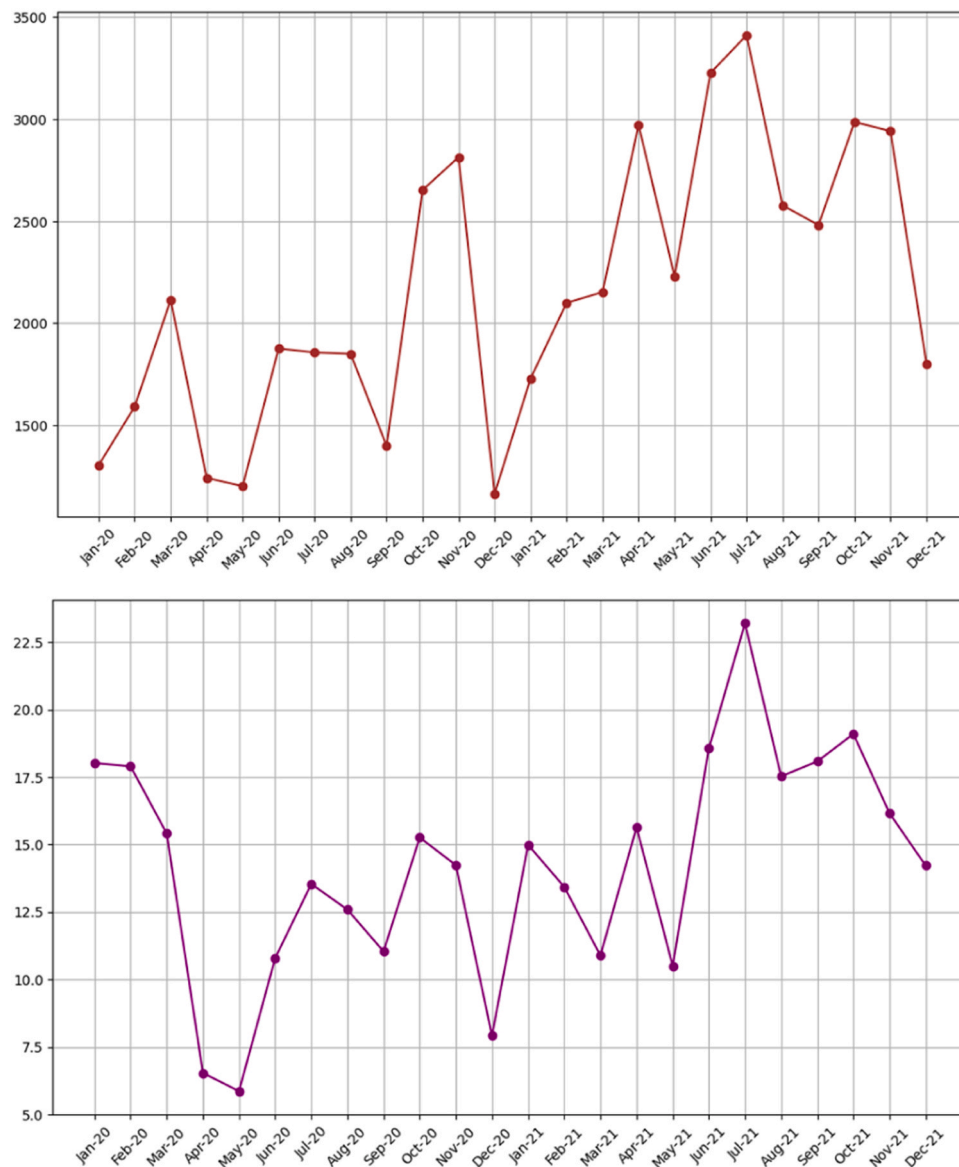


Fig. 3. (continued).

modules (rated at 50 kWp), which will also satisfy the PV inverter's specifications and electrical connection requirements.

An energy meter allows for monitoring the performance of a solar plant through power generation and consumption tracking. All these operations are managed through a complex control system that controls the solar panels, inverters and back-up generators for efficiency optimization as well as defect detection and real-time monitoring. For the present study the data is collected from the IoT based system installed with PV plant and energy meter at substation.

3. Methodology

3.1. Data collection

For the study, a large dataset from 100 kWp rooftop PV plant is utilized. The data is collected between January 2020 to December 2021. This dataset is shown in Fig. 3, and contains monthly average values for a selection of key variables- Solar Irradiance (kWh/m²/day), unit consumption by load (kWh), actual grid supply to a load (kWh), an average PF of the grid supply, net apparent energy supplied by grid to load (kVAh), energy saving (in kVAh), additional APFC required, and actual

additional required (kVAR). This data is used for building accurate forecasting models and thus increase efficiency of electrical systems.

The dataset analyzed consists of various parameters recorded from January 2020 to December 2021. Where Fig. 3(a) displays a distinct seasonal pattern of Solar Irradiance (kWh/m²/day). The maximum solar irradiance, recorded at 6.8 kWh/m²/day, occurs during May in both 2020 and 2021, while the minimum value of 4.5 kWh/m²/day is observed in January 2020 and 2021. The average solar irradiance across the two years is 5.6 kWh/m²/day. The data variation is consistent with expected seasonal changes, where longer daylight hours and higher sun angles during summer months lead to increased solar irradiance. Fig. 3 (b) displays Unit Consumed by Load (kWh) shows significant fluctuations throughout the analyzed period. The maximum consumption occurs in January 2020 with 70,094 kWh, whereas the minimum consumption of 19,278 kWh is recorded in May 2020. The average energy consumption is 33,403.83 kWh. The data reveals a sharp decline in energy consumption in April and May 2020, potentially due to external factors such as operational changes or reduced demand during these months. Overall, the data suggests a variable energy usage pattern influenced by both seasonal factors and operational demands. Grid Supply to Load (kWh) shows in Fig. 3(c) closely mirrors the pattern

observed in unit consumption. The grid supply peaks at 63,911 kWh in January 2020 and reaches its lowest point at 6807 kWh in May 2020. The average grid supply is 26,690.03 kWh. The significant dip in grid supply in April and May 2020 corresponds to the drop in load consumption, indicating a direct relationship between grid supply and energy demand. The data reflects the adaptability of grid supply to varying consumption needs. Fig. 3(d) shows Actual Grid Supply Average PF exhibits moderate fluctuations with a maximum value of 0.98 recorded in January 2020 and a minimum value of 0.84 in May 2021. The average PF across the period is 0.90. The power factor generally remains within a narrow range, indicating overall efficient energy usage. However, the lower PF values observed in May 2021 and April 2020 suggest periods of inefficiency, where the system may have required correction to optimize energy usage.

Net Apparent Energy Supplied by Grid to Load (kVAh) shown in Fig. 3(e) follows a pattern similar to grid supply and load consumption. The maximum net apparent energy is 65,215.31 kVAh in January 2020, while the minimum value of 8008.24 kVAh is recorded in May 2020. The average net apparent energy is 26,822.69 kVAh. The data variation reflects the impact of seasonal demand changes and operational adjustments, with notable decreases during May 2020, aligning with lower energy consumption and grid supply. In Fig. 3(f) shows PV Produced Energy (kWh) shows a consistent production pattern with higher values during the summer months and lower values during winter. The maximum PV energy produced is 12,552.00 kWh in July 2021, while the minimum production of 6182.95 kWh is observed in January 2020 and 2021. The average PV energy production is 10,913.29 kWh. The data aligns with solar irradiance trends, where longer daylight and higher irradiance during summer months result in increased PV production. Fig. 3(g) displays additional APFC Required to Maintain PF = 1 (kVAR) varies across the period, with a maximum requirement of 23.20 kVAR in July 2021 and a minimum of 5.86 kVAR in May 2020. The average additional APFC required is 14.42 kVAR. The data indicates that more APFC is necessary during months with lower power factor efficiency, particularly in July 2021, to maintain a PF of 1. This highlights the dynamic need for power factor correction depending on seasonal and operational factors. Fig. 3(h) displays Energy Savings (kVAh) exhibits moderate fluctuations, with the highest savings of 3409.39 kVAh occurring in July 2021 and the lowest savings of 1163.57 kVAh in December 2020. The average energy savings is 2178.60 kVAh. The data suggests a correlation between PV energy production and energy savings, with higher savings recorded during periods of increased PV production, particularly in the summer months. Fig. 3(h) displays additional APFC required to maintain PF = 1 (kVAR) varies across the period, with a maximum requirement of 23.20 kVAR in July 2021 and a minimum of 5.86 kVAR in May 2020. The average additional APFC required is 14.42 kVAR. The data indicates that more APFC is necessary during months with lower power factor efficiency, particularly in July 2021, to maintain a PF of 1. This highlights the dynamic need for power factor correction depending on seasonal and operational factors.

The analysis of the data represented in Fig. 3(a) to (h) provides insights into the seasonal and operational variations affecting energy production, consumption, and power factor management. The variations in solar irradiance, PV production, and energy savings highlight the influence of seasonal factors, while fluctuations in unit consumption, grid supply, and APFC requirements underscore the impact of operational demands. Understanding these variations is crucial for optimizing energy management and ensuring efficient power factor correction throughout the year.

1) Data preprocessing: One of the most important steps in ensuring the quality of the dataset prior to its application in machine learning models is data preprocessing. A dataset collected from a 100 kWp rooftop PV plant for the time period between January 2020 and December 2021 was subjected to extensive preprocessing for model training and evaluation.

- **Data cleaning:** The initial dataset underwent cleaning to eliminate inconsistency, duplication, and error values. Outliers were detected and handled by statistical methods involving Z-scores and the IQR technique. Where found, extreme values were either trimmed or transformed to fall within acceptable ranges to avoid affecting the performance of a model negatively.
 - **Removing data redundancy:** Correlation analysis identified highly correlated or redundant features and removed multicollinearity from the data set. This is an essential step because it ensures relevance-only features in the modeling process to prevent overfitting models and enhance the accuracy of predictions.
 - **Data transformation:** We use categorical variables such as month of the year through one-hot encoding to make it readily usable in the machine learning models. Continuous variables such as solar irradiance and energy consumption are normalized or standardized to ensure all features lie within similar scales. This transformation is critical to improve model stability and accuracy in predictions.
- 2) **Missing value analysis:** Handling missing data is an important preprocessing step, ensuring that it does not cause models to become biased or inaccurate because of incomplete data. For numerical feature variables that have missing values, mean imputation is used whereas for categorical features, mode imputation is utilized. This approach was based on the assumption of data missing at random. For time-series data such as those from solar irradiance and grid supply, time-based imputation techniques were used: interpolation and forward/backward filling. These are in order to ensure that logical imputation of missing values took place in confluence with the nature of data produced, hence ensuring the underlying patterns are maintained. After imputation, the missing values were handled effectively without missing entries in the complete dataset.
 - 3) **Feature engineering:** Feature engineering is a vital step in improving the predictive capabilities of machine learning models.
 - **Feature selection and creation:** Relevant features were selected based on their correlation with the target variable (such as energy consumption, power factor, and additional APFC). New features were also created to enhance model performance, including the interaction between solar irradiance and grid supply. These new features helped capture the relationship between these parameters, which are integral to the model's predictions.
 - **Temporal feature engineering:** Time-based features, such as hour, day, and month, were derived to account for daily, weekly, and seasonal variations in energy consumption and solar production. These temporal features enabled the models to understand the cyclical nature of energy consumption and solar output, which is crucial for accurate predictions in energy systems.
 - 4) **Data sampling and model splitting:** Proper sampling and splitting are crucial for model training and evaluation on unseen data.
 - **Training and testing split:** It split the dataset into training and testing sets with a 70:30 ratio, wherein 70 % of data was used for the training model, while the remaining 30 % formed the test set. It means that the created model is trained on a highly significant portion of data but tested on an independent subset, which would help to check the generality of this model.
 - **Data sampling:** Since the dataset is a time-series, monthly sampling was performed with the aim of capturing seasonal trends and temporal dependencies. This strategy ensures that the models can subsequently learn the time-dependent characteristics of solar irradiance, energy consumption, and grid supply.
 - **Cross-validation:** We utilized cross-validation to evaluate model performance and prevent overfitting. This was done by taking subsets of the data on which to test the model, in effect ensuring that the

model could generalize well to unseen data and should perform consistently.

After performing data cleaning, redundancy removal, feature engineering, and handling missing values, the final dataset is prepared for model training. All preprocessing steps have been successfully completed, with the dataset free of any missing or erroneous values. The features are transformed into suitable formats for machine learning models to ensure all the inputs are of high quality and relevance.

Removing missing data, addressing outliers, and engineering relevant features of the datasets would then be very strong and structured for model development. All these processes would contribute significantly to increased accuracy and reliability for prediction models to perform well when trying to predict the power factor, energy consumption, and any other APFC requirements for the system.

3.2. Machine learning models

3.2.1. Linear regression

A statistical technique called linear regression is used to model the connection between one or more independent variables and a dependent variables $x_1, x_2, x_3, \dots, x_n$. The relationship is represented by the (Eq. 1) (AlShafeey and Csák, 2021).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon. \quad (1)$$

where β_0 is the intercept $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the coefficients for the independent variables, and ϵ is the error term representing unexplained variability. Linear regression aims to find the coefficients that minimize the difference between observed and predicted values of y making it a powerful tool for predicting outcomes based on given input variables.

3.2.2. Random forest

The Random Forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode (classification) or mean prediction. Each individual tree in this forest was trained on random data samples as well as features to ensure differences between these trees. Output last prediction of new input — average predictions from all trees shown (Eq. 2) (Cattani, 2023; Shareef et al., 2017).

$$\hat{y} = \frac{1}{T} \sum_{i=1}^T h_i(X). \quad (2)$$

where T is the number of trees in the forest, and $h_i(x)$ is the prediction of the i -th decision tree for input x . Random Forests are effective in handling complex relationships and noisy data, making them suitable for predicting grid supply power factor and determining additional APFC required in our research.

3.2.3. Decision tree

A decision tree is a flexible supervised learning technique that minimises impurity and predicts the target variable by dividing data into subsets according to feature values. Using metrics like entropy or Gini impurity, the approach chooses the feature and split point that best separates the data at each node in the tree. Using the feature values of a new input x , the tree is traversed from the root to a leaf node in order to determine the prediction. Eq. 3 illustrates how the procedure maximises homogeneity within each subgroup by optimising the splits, despite the absence of a single overarching formula (Mahesh et al., 2022; Li et al., 2024; Wang et al., 2018).

$$Q(t) = \frac{m_{\text{left}}}{m} H(D_{\text{left}}) + \frac{m_{\text{right}}}{m} H(D_{\text{right}}) \quad (3)$$

3.2.4. XG boost classifier

XG Boost is a powerful algorithm for classification tasks, including

predicting grid supply power factor and determining additional APFC required in your research. It sequentially combines predictions from multiple decision trees to improve accuracy (Zou et al., 2022).

$$\hat{y} = \sigma \left(\sum_{t=1}^T f_t(x) \right) \quad (4)$$

Here, $\hat{y}$ represents the predicted class probability for input x aggregated over T trees. XGBoost integrates regularization techniques and parallel computing to enhance model performance and scalability, making it well-suited for handling complex relationships and optimizing predictions based on features like solar irradiance and PV production.

3.2.5. Adaboost classifier

Adaboost (Adaptive Boosting) is an ensemble learning technique that sequentially combines weak learners to improve classification accuracy, making it ideal for predicting grid supply power factors and determining additional APFC required in your research. The final prediction $\hat{y}$ for a new input x is determined by combining the weighted predictions of multiple weak learners (Lodhi et al., 2022).

$$\hat{y}(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right) \quad (5)$$

3.2.6. Gradient boosting

Gradient Boosting is an ensemble learning technique that sequentially combines weak learners, typically decision trees, to improve prediction accuracy. The final prediction $\hat{y}$ for an input x is determined by summing predictions from all trees in the ensemble shown in Eq. 6 (Wang et al., 2018; Qinghe et al., 2022).

$$\hat{y}(x) = F_{T-1}(x) + \gamma \cdot h_T(x) \quad (6)$$

Here, $F_T(x)$ represents the ensemble prediction after T trees, $F_{T-1}(x)$ is the ensemble prediction after $T - 1$ trees, $h_T(x)$ is the prediction of the T -th tree, and γ is the learning rate that controls the contribution of each tree. Gradient Boosting is effective for predicting grid supply power factors and determining APFC requirements by iteratively refining predictions based on features like solar irradiance and PV production, ensuring accurate and robust results.

Lasso regression: Lasso Regression (Least Absolute Shrinkage and Selection Operator) is a linear regression method that includes an L1 penalty to improve model accuracy and interpretability by performing feature selection. It is useful for predicting grid supply power factor and determining additional APFC required by shrinking some coefficients to zero, thus highlighting the most relevant features. The Lasso objective function shown in Eq. 7 (Tang et al., 2018).

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{2N} \sum_{i=1}^N (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (7)$$

Here, y_i is the target variable, x_i is the feature vector, β are the coefficients, and λ is the regularization parameter. This method helps prevent overfitting and improves the interpretability of the model by focusing on the most influential features.

8) Ridge Regression: Ridge Regression is a linear regression method that includes an L2 penalty to improve model robustness by handling multicollinearity and preventing overfitting, making it suitable for predicting grid supply power factor and determining additional APFC required. The Ridge Regression objective function shown in Eq. 8 (Erten, 2022).

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{2N} \sum_{i=1}^N ((y_i - x_i^T \beta))^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (8)$$

Here, y_i is the target variable, x_i is the feature vector, β are the coefficients, and λ is the regularization parameter. This method reduces the variance of the coefficient estimates, ensuring more stable and

accurate predictions by penalizing large coefficients.

The Linear Regression model, Random Forest, Decision Tree, XGBoost, AdaBoost, Gradient Boosting, Lasso Regression, and Ridge Regression models used in this work are good predictors of the grid supply power factor, and thus can be adopted to estimate additional APFC. However, some components of their implementation would benefit more from enhanced technical depth, although these models were chosen because they are capable of representing complex data relationships, the problems associated with applying each model call for further discussion. For instance, in the case of Linear and Ridge Regression models, multicollinearity as an issue was addressed through feature selection, data preprocessing, and hyperparameter tuning while dealing with noisy data in Random Forest models. Additionally, overfitting that is very common in ensemble methods like Random Forest and Gradient Boosting was further controlled by cross-validation with over-regularization techniques. The further deconstruction of these challenges and the strategies adopted to overcome them would give better insight into how robust the model was.

Additionally, the novelty of this approach against literature should be clearly given. An illustration of this approach is the feature selection that Lasso Regression offers to reduce model complexity when compared to previous approaches. Likewise, the use of XGBoost's parallel computing in handling large data as well as its regularization can be highlighted by contrasting it to past methods - the ways through which this research advances energy prediction modeling. A clearer comparison with previous works will not only fortify the findings but also provide a better perspective of the importance of methodology within the framework of energy management systems and predictive modelling (Carneiro et al., 2022; Solar Power Prediction using Regression Models, 2022).

3.3. Model evaluation

- 1) Model evaluation is crucial in assessing the performance of predictive models in our research, predicting grid supply power factor and determining additional APFC required. Common evaluation metrics include MSE, RMSE, MAE, MAPE, and R^2 .
- 2) Mean Squared Error: MSE measures the average of the squares of the errors, which are the differences between predicted and actual values. It gives higher weight to larger errors, making it sensitive to outliers, which is shown in Eq. 9 (Lee et al., 2024; Shuvho et al., 2019).

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (9)$$

where, N is the number of observations y_i is the actual value $\hat{y}_i$ is the predicted value.

- 3) Root Mean Squared Error (RMSE): RMSE is the square root of MSE and provides a measure of error in the same units as the target variable, making it more interpretable, which is shown in Eq. 10 (Lee et al., 2024; Shuvho et al., 2019).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (10)$$

- 4) Mean Absolute Error (MAE): MAE measures the average of the absolute differences between predicted and actual values, providing a linear score that does not square the errors, thus being less sensitive to outliers compared to MSE (Anon, 2023; Kassem et al., 2021a).

$$MAE = \frac{1}{N} |y_i - \hat{y}_i| \quad (11)$$

- 5) Mean Absolute Percentage Error (MAPE): MAPE expresses the prediction accuracy as a percentage, making it easier to interpret relative errors. It is the average absolute percentage error between predicted and actual values (Kassem et al., 2021b; Sarmah et al., 2023).

$$MAPE = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (12)$$

- 6) R-squared (R^2): R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit, which is shown in Eq. 13 (Ansari et al., 2021; Dondariya et al., 2018).

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (13)$$

This extensively describes models performance analysis since model prediction is aiming for grid supply power factors as well APFC requirements represented by different evaluation metrics. MSE and RMSE are good summary statistics for large errors, which provides a more complete picture of how accurate your predictions overall. MAE is easy to understand, it gives the average error where all errors are treated as equally important. Mean Absolute Percentage Error MAPE provides a simple measure of the size of the errors in percent to enable comparison across different scales and datasets Lastly, R-squared or the coefficient of determination R^2 is used to measure how well your model predicts variance in target variable, higher the r-square better it explain. This will allow you to finally evaluate the performance of your models and also help in determine if such an approach for predicting grid supply power factors or APFC requirements was useful.

3.4. Actual v/s predicted value

When we researched for power factor prediction based on grid supply and selection of additional APFC it became mandatory to understand the correlation between actual values with predicted ones. The efficiency of PF is known as a real-time observation in the system, always compare your estimated values by model with observations on site. Comparing these helps measure the accuracy of models using MSE, RMSE, MAE & R^2 as well for robust trend capture and system management insights. Similarly, APFC translates to the measured adjustments in reactive power under maintenance for PF correction and if required updates real-time corrections while predicted APFC values influence a proactive approach. This validation is important for better grid efficiency, planning renewable energy integration and making operational decisions in order to help make electrical power systems both sustainable and reliable. Fig. 4: Structured Method employed in our research for Precise prediction of Grid supply power factor and Extra APFC required to maintain PF = 1 (kVAR) Data preparation is the first step taken to clean up any extraneous or erroneous columns so that it will not impact your dataset. In addition, redundant data points were found and removed to address biasing issue which in turn may increase quality of datasets.

3.5. Hyperparameter tuning

Predicting the power factor and additional APFC required to maintain PF = 1, hyperparameter tuning is crucial in the optimization of the performances of such machine learning models. For instance, for each model, which in this case includes Ridge Regression, Random Forest Regressor, Decision Tree Regressor, XGBoost, AdaBoost, Gradient Boosting, and Lasso Regression, a set of hyperparameters typically de-

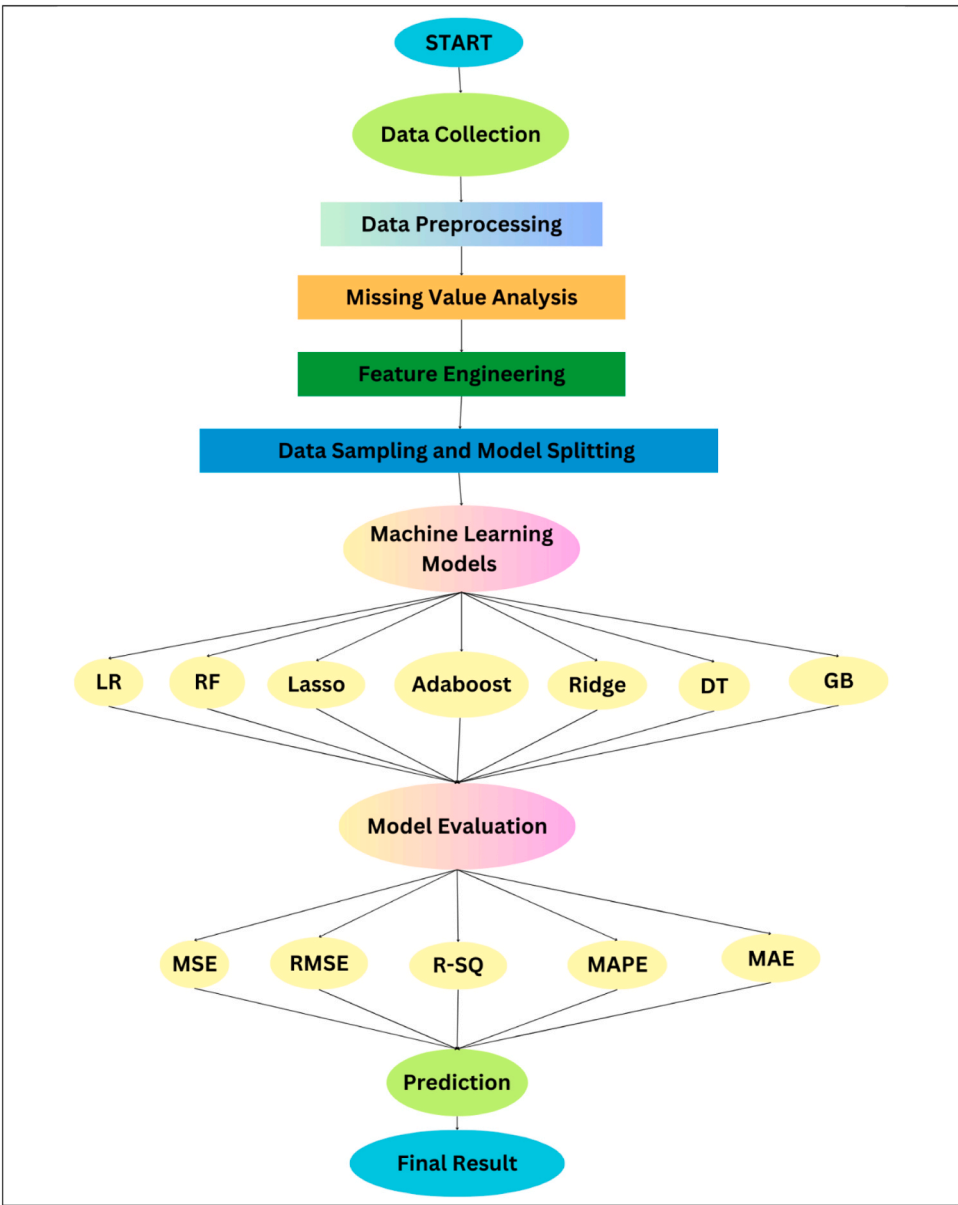


Fig. 4. Overview of the methodology used in the proposed system.

termines the learning process. Ridge Regression has alpha as a main hyperparameter used in order to give control over the strength of regularization and avoid overfitting by penalizing large coefficients. For Random Forest and Decision Tree Regressor, *n_estimators*, *max_depth*, *min_samples_split*, and *min_samples_leaf* are the cardinal hyperparameters that control how complex is the model so that it is not specialized and generalizes well. These depend on parameters such as *learning_rate* and *n_estimators*, *subsample*, and *max_depth* that govern the learning rate, boosting rounds, and depth of individual trees, all critical factors to avoid overfitting and enhance robustness.

We use GridSearchCV to tune these hyperparameters because it carries out an exhaustive search over ranges of values for the parameters. This allows one to identify the best combination of parameters for each model to be able to generalize very well. The goal is to find a balance between bias and variance that allows the model to be able to predict without overfitting or underfitting. Hyperparameter tuning will prove to be a very significant aspect of our efforts in improving the predictive performance of our models in order to correctly predict the power factor and extra APFC necessary to keep PF= 1. Careful selection,

tuning, and fine-tuning of these parameters contribute to the reliability and efficiency of solutions derived from our study.

The Categorical variables were changed to normalized way, then made sure that the entries in each column are same across all the data set. In order to plug the missing data, sophisticated imputations were used (like mean/median and predictive modeling) wherever applicable so that dataset was intact in nature from a statistical perspective. We have used different machine learning algorithms such as LR, RF, DT, XGBoost Classifier, AdaBoost, Gradient Boosting, Ridge Regression and Lasso Regression. We trained the models by splitting our dataset into train and validation datasets to learn from data patterns. We were evaluating the performance using standard metrics like MSE, RMSE, R^2 and MAPE. Tabular & Visual comparisons of actual vs predicted values to evaluate the model performance under different scenarios, enabling an iterative process from a Model development stage to feature engineering and pipeline.

Our methodology systematically guided the prediction of Grid supply power factor and Additional APFC, leveraging advanced machine learning techniques. By integrating meticulous data preparation,

strategic algorithm selection, rigorous training, and comprehensive evaluation, our research aimed to deliver accurate insights essential for informed decision-making in energy management and optimization.

4. Result and analysis

The project's goal is to improve the energy efficiency of grid-connected rooftop solar systems by using several machine learning approaches to anticipate the power factor. The projected values produced by a number of machine learning models, including LR, RF, Lasso, Adaboost, Ridge, DT, GB, The planned and actual power factors are compared in Figs. 5 through 12. The y-axis of these graphs shows the power factor, while the x-axis shows "Month_Year." The orange line represents the expected power factor, and the blue line shows the actual power factor.

The Linear Regression Fig. 5 model shows a consistent and close approximation to the actual power factor values. The model performs well across all months, with minor deviations. For example, in January 2020, the predicted power factor is 0.964, while the actual is 0.98. This slight underestimation persists across the months, indicating the model's stable performance but with a slight tendency to predict lower values. With a few deviations, the Random Forest model in Fig. 6 also forecasts values that are almost identical to the real power factor. In certain months, like December 2020 (0.882 predicted vs. 0.92 real), the model tends to be very conservative; but, in May 2020 (0.855 projected vs. 0.85 actual), it overestimates. Though marginally different, the forecasts generated by the Random Forest model are often correct. Although the Decision Tree Shows in Fig. 7 model achieves an acceptable match for the most of the months, its deviations can occasionally be much larger due to its heightened sensitivity to data changes. For example, in June 2020, the anticipated power factor is 0.887, whereas the actual number is 0.89. The model occasionally overfits, as evidenced by months of noticeable variations in power factor.

High accuracy is demonstrated by XGBoost, Shows in Fig. 8 especially in months with stable power factor values. It roughly tracks the real power factor, with just minor variations. For instance, the actual value of 0.93 in March 2020 roughly matches the anticipated value of 0.929 for power factor. When it comes to power factor prediction throughout several months, XGBoost demonstrates resilience. Fig. 9

shows how AdaBoost and XGBoost function similarly across a range of months, with slight under- and overestimations. The projected values are consistently in close proximity to the actual power factor, with a little trend towards underestimation. For instance, in January 2020, the actual power factor is 0.98, whereas the predicted value is 0.963. With minimal adjustment, the shows in Fig. 10 Gradient Boosting model accurately follows changes in the real power factor. Given that it captures power factor variations well, it is among the most accurate models we looked at. For example, in September 2021, the actual power factor is 0.93, but the projected value is 0.932. As Fig. 11 illustrates, the predictions of the Ridge Regression are stable and reliable, with only a little deviation from the genuine power factor. The model's regularisation helps control variation and yields close predictions. For instance, in October 2021, the power factor is predicted to be 0.910, which is nearly identical to the actual value of 0.91. Fig. 12 displays the Lasso Regression. Lasso Regression predicts well as well, despite a little tendency to overestimate in some months. The model performs well, particularly in the months when the real power factor is lower, such April 2020, when the expected value of 0.871 and the actual value of 0.87 are nearly identical.

The study demonstrates that all models provide an acceptable estimate of the true power factor, but with varying degrees of precision. Since they produce the most accurate predictions, XGBoost, Gradient Boosting, and Ridge Regression are some of the best techniques. Both Random Forest and Linear Regression perform well, despite certain biases. The capacitor banks required to improve the energy efficiency of grid-integrated rooftop PV installations may be predicted using these models. The way energy management systems are developed and implemented, which ensures that the required capacitor banks are exactly proportioned to guarantee perfect power factor values, may help to explain the little differences in the projections. This will eventually increase the energy efficiency and stability of the grid-integrated PV systems.

The assessment criteria for the machine learning models used to estimate the power factor are summarised in Table 1. Power factor is a critical component in identifying the proper capacitor banks required to improve the energy efficiency of rooftop solar systems that are linked to the grid. A number of measures were assessed, including MSE, R^2 , RMSE, MAE and MAPE. Regression also has the lowest average squared and

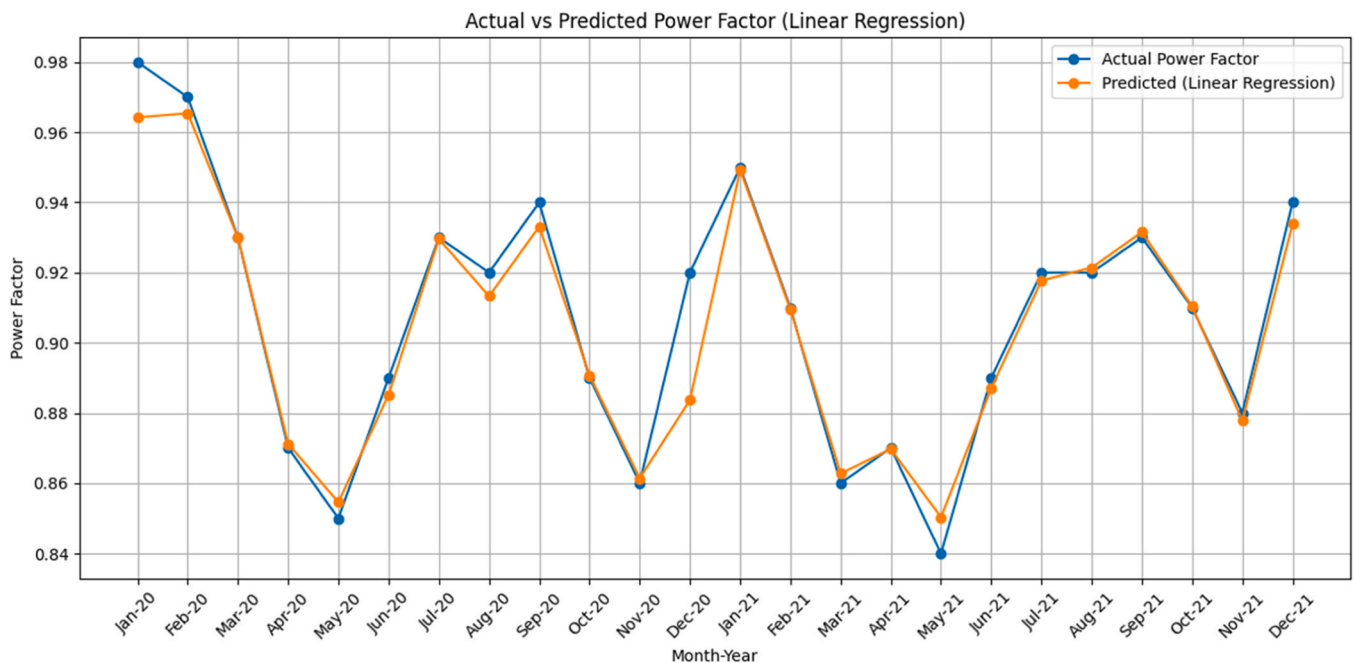


Fig. 5. Actual vs predicted power factor (LR).

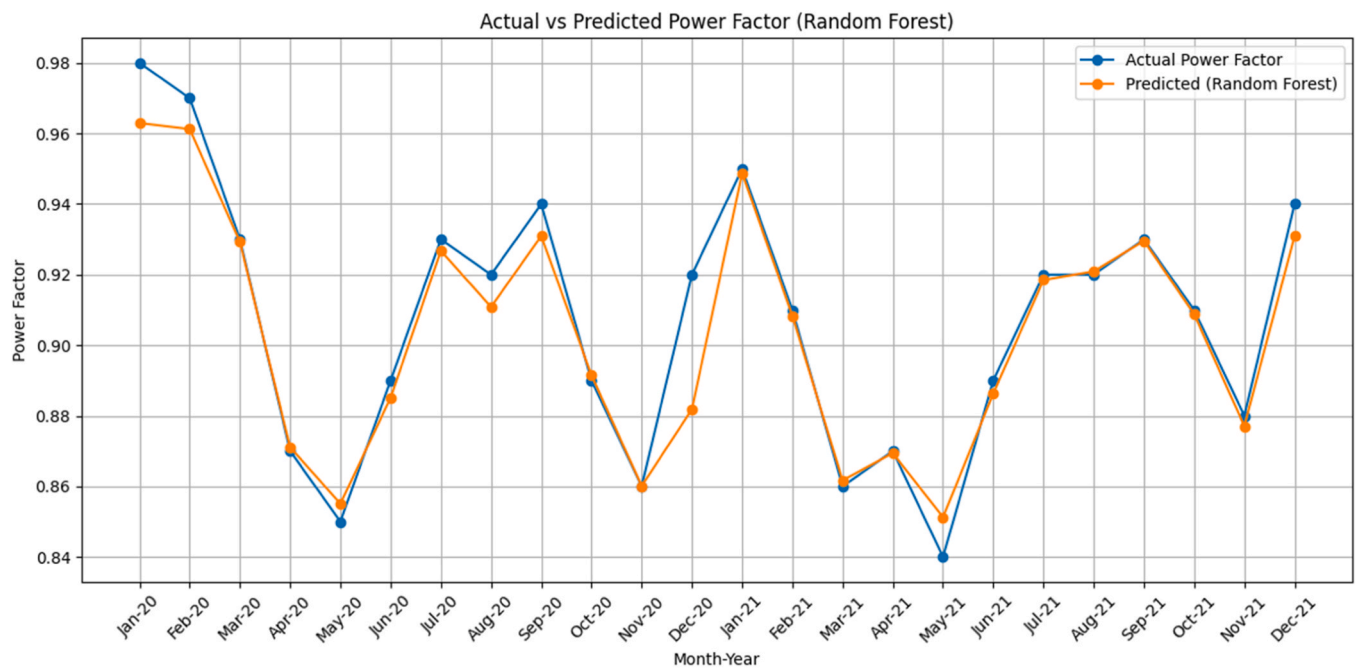


Fig. 6. Actual vs predicted power factor (RF).

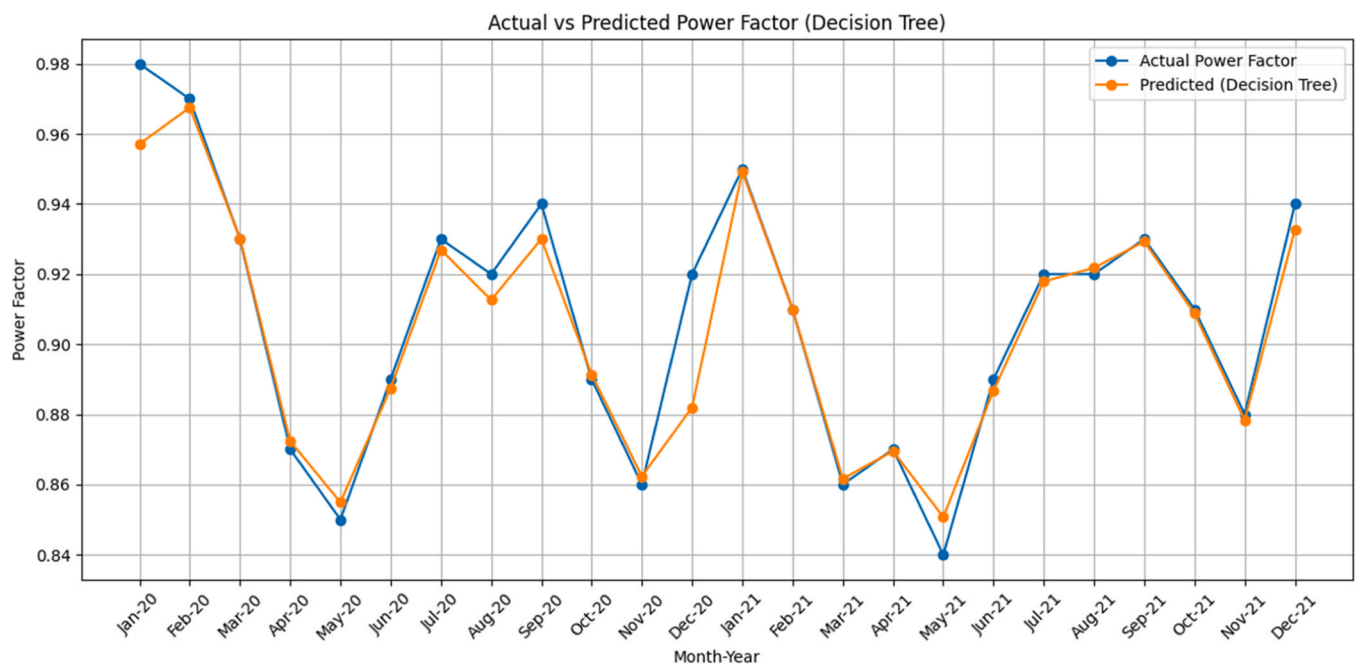


Fig. 7. Actual vs predicted power factor (DT).

absolute disparities between the predicted and real power factors, with MSE (0.000233) and RMSE (0.015) values that are the lowest. This shows that Lasso Regression has a very high accuracy and consistency rate. Conversely, Ridge Regression has the greatest RMSE (0.035) and MSE (0.001234), suggesting a greater average departure from the true values. Compared to other models, this one could be less successful in reducing prediction mistakes. Once more taking the lead, Lasso Regression has the greatest R-Squared value (0.888), which indicates that it accounts for 88.8 % of the variation in the power factor data. The model fits the data rather well, as evidenced by the high R2 score. With its lowest R2 (0.407), Ridge Regression appears to explain just 40.7 % of the variance. This suggests that other models may be better able to

identify the underlying patterns in the data than Ridge Regression.

The Lasso Regression model has the lowest MAE of any model tested, with a value of 0.013. This demonstrates how well it can forecast power factor with accuracy. The MAE (0.032) for Ridge Regression is the highest, indicating bigger average absolute errors and, thus, less accurate predictions. Lasso Regression also excels with the lowest MAPE (1.45 %), indicating that, on average, the predicted values are 1.45 % off from the actual values. This demonstrates its efficiency in minimizing percentage errors. Ridge Regression has the highest MAPE (3.45 %), reflecting a greater average percentage error and implying a less reliable performance. Based on the assessment measures, it can be concluded that Lasso Regression is the best model evaluated for power factor

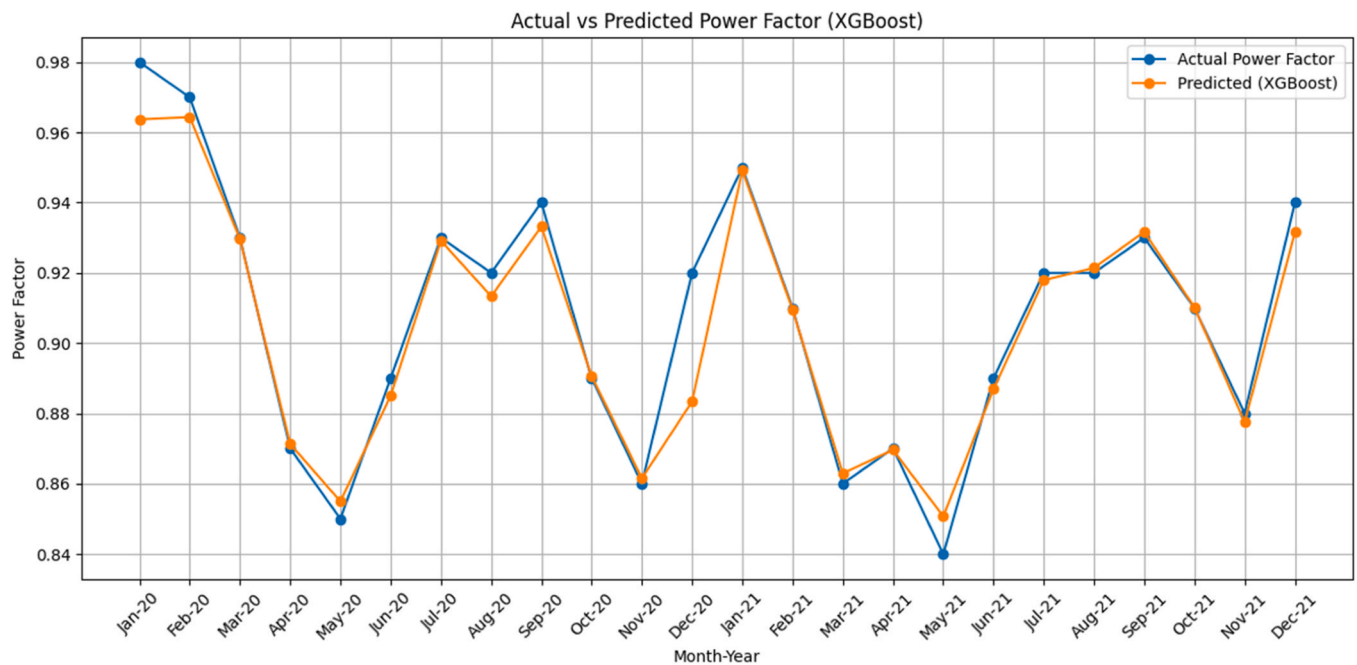


Fig. 8. Actual vs predicted power factor (XG boost).

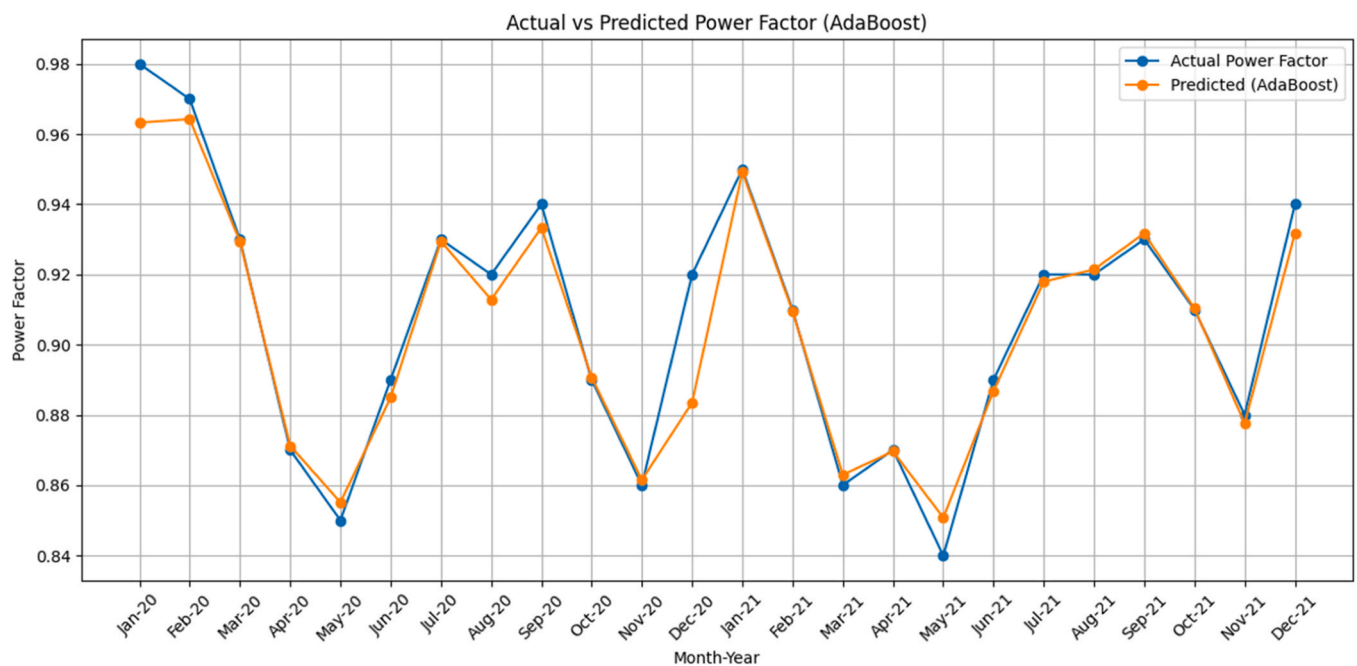


Fig. 9. Actual vs predicted power factor (Adaboost).

forecasting. It frequently has the lowest R^2 and greatest MSE, RMSE, MAE, and MAPE, indicating that it offers the most precise and trustworthy forecasts. Because of its excellent performance, this model is quite appropriate. The least successful regression model is Ridge Regression, which has larger errors and less explanatory power. Although Ridge Regression could still be helpful in some situations, its results indicates that Lasso Regression is a better option for precise power factor predictions. The integration of Lasso Regression into energy management systems can aid in the accurate determination of capacitor bank requirements, resulting in improved grid-integrated rooftop PV system performance and increased energy efficiency.

An analysis of the additional requirements for APFC (Automatic

Power Factor Correction) is necessary in order to improve the energy efficiency of rooftop PV plants connected to the grid. These findings, displayed in Figs. 13 through 20, demonstrate how several machine learning algorithms compare predicted and actual APFC values. The accuracy of these models is measured by their ability to forecast the amount of APFC needed to maintain a power factor of 1 (kVAR).

The accuracy of the Linear Regression model is moderate, as seen in Fig. 13. While it correlates pretty well generally, there are some noticeable deviations from the actual statistics, particularly around August 2020 and May 2021. This model has a tendency to overestimate and underestimate the APFC requirements in different months. Ridge Regression, as seen in Fig. 14, performs better than Linear Regression

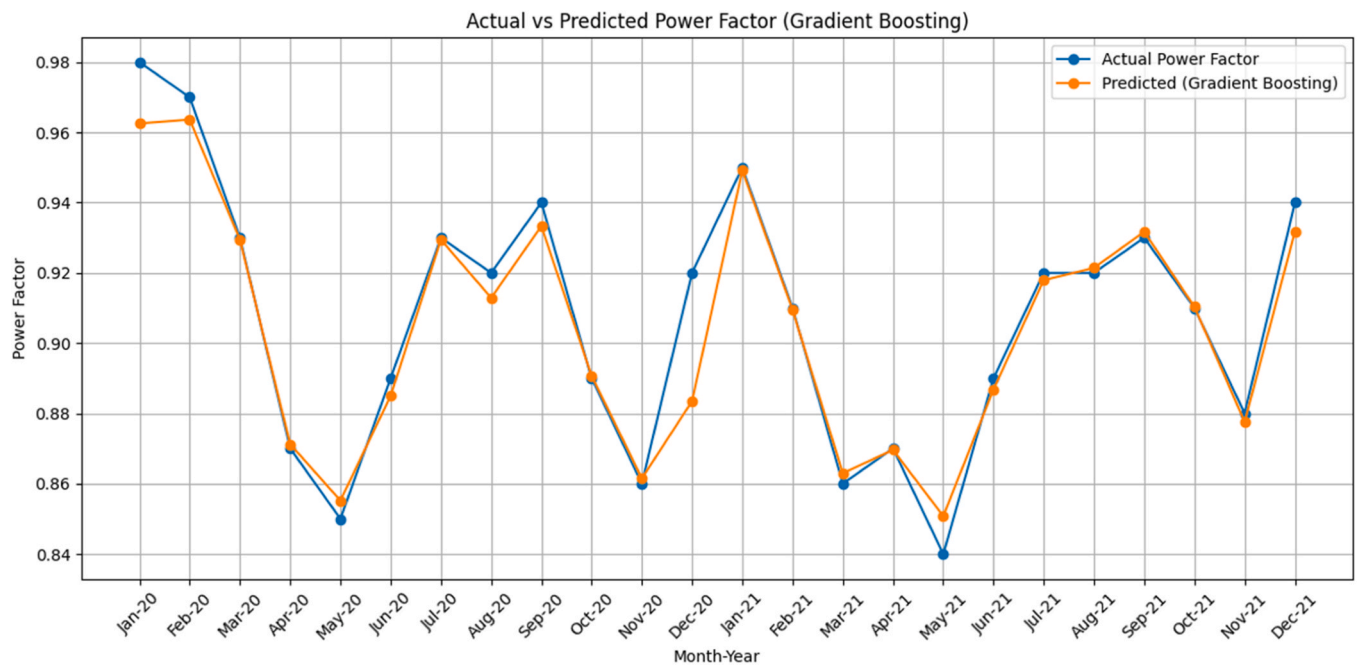


Fig. 10. Actual vs predicted power factor (Gradient boosting).

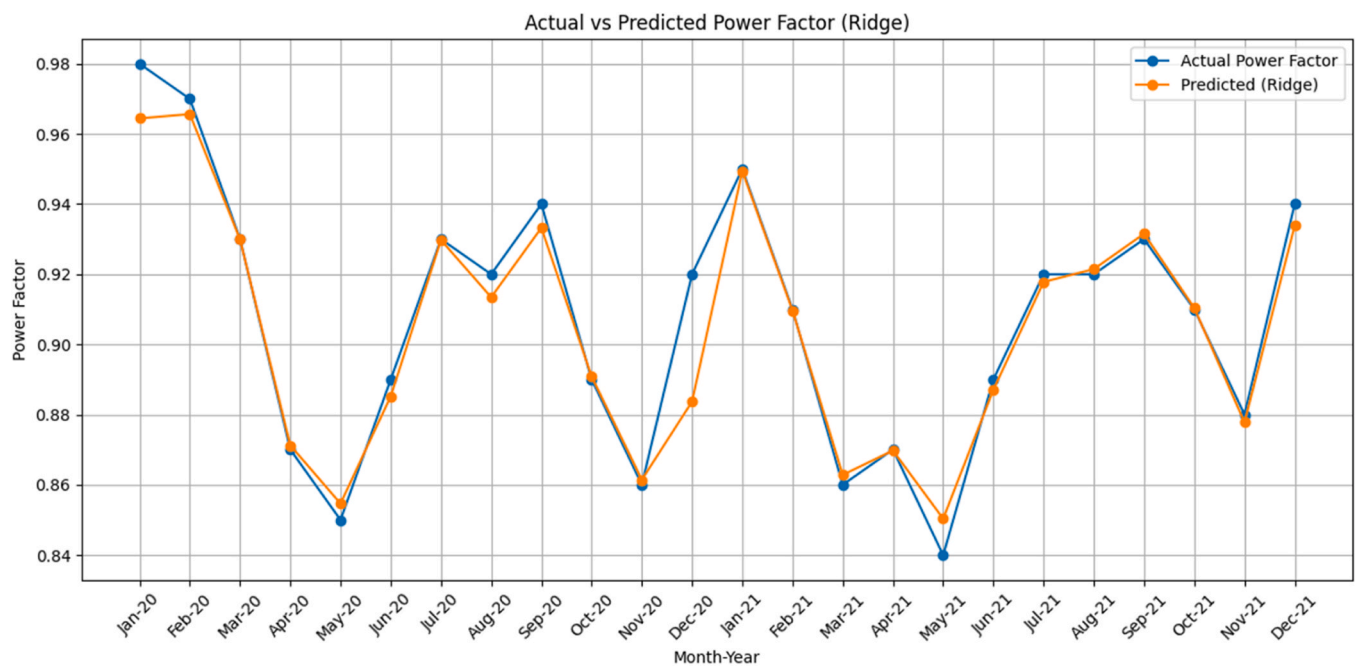


Fig. 11. Actual vs predicted power factor (Ridge).

despite still having some discrepancies since it shows a closer alignment with real values, especially in the latter months. This model does quite well overall, despite its inability to describe oscillations in APFC demand across specific time periods.

The most trustworthy results are produced by the Lasso Regression, as shown in Fig. 15, which effectively manages variances and closely matches the actual APFC values. It is quite accurate and reliable, particularly when it comes to pattern recognition. There are some deviations from actual values in Fig. 16's Random Forest, especially in months where APFC requirements change considerably. Though it varies noticeably over specific months, the Decision Tree model (Fig. 17) generally follows the pattern. While XGBoost (Fig. 18) and Gradient

Boosting (Fig. 19) both offer excellent trend capture and high accuracy, XGBoost shows slight variances. Fig. 20 illustrates how AdaBoost predictions, which are often accurate with a few slight discrepancies, particularly in months where APFC demands fluctuate rapidly.

The most accurate model for calculating increased APFC requirements is Lasso Regression. It accurately depicts trends and variances and often matches actual figures. Our model is particularly helpful in precisely calculating the capacitor bank needs that are needed to maximise the energy efficiency of rooftop photovoltaic systems that are connected into the grid. Both gradient boosting and XGBoost function well, exhibiting less deviance from real values and superior prediction accuracy. These models are strong and able to adapt to changes in APFC

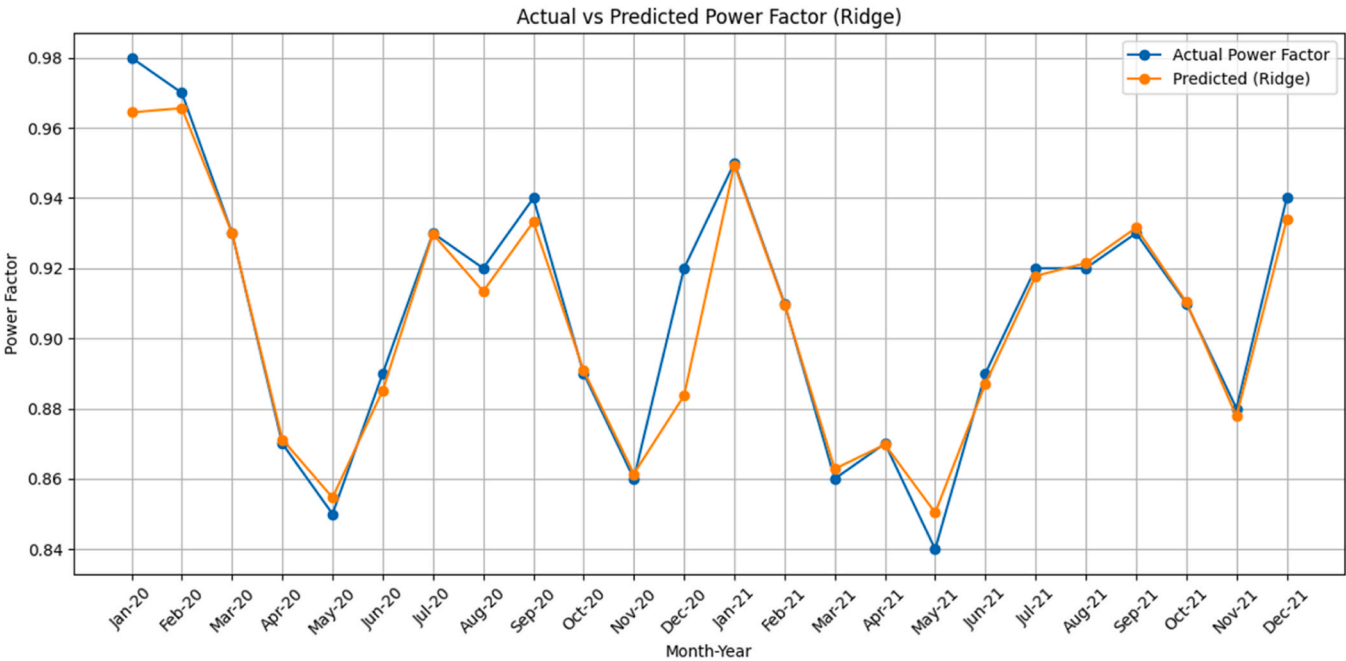


Fig. 12. Actual vs predicted power factor (Lasso).

Table 1
Show's evaluation matrices of power factor.

Model	MSE	R-Squared	RMSE	MAE	MAPE (%)
Linear Regression	0.000324	0.844	0.018	0.016	1.73
Random Forest	0.000610	0.707	0.025	0.020	2.18
Decision Tree	0.000960	0.538	0.031	0.024	2.60
XG Boost	0.000534	0.743	0.023	0.015	1.58
Ada-boost Classifier	0.000824	0.604	0.029	0.022	2.41
Gradient Boosting	0.000688	0.669	0.026	0.017	1.83
Lasso Regression	0.000233	0.888	0.015	0.013	1.45
Ridge Regression	0.001234	0.407	0.035	0.032	3.45

specifications. However, there are more noticeable deviations from Decision Tree and Random Forest's forecasts, especially in cases of abrupt shifts. Ridge Regression and Linear Regression are not able to fully capture all changes, despite their relative accuracy. Overall, by including models like Lasso Regression, XGBoost, and Gradient Boosting into energy management systems, the precision of forecasts for capacitor bank requirements may be increased. This would boost rooftop PV systems that are incorporated into the grid's efficiency.

Based on the performance indicators in Table 2 and the assessment of machine learning models for predicting the extra APFC (Automatic Power Factor Correction) needed for grid-integrated rooftop PV plants yields the following insights. There are notable variations in the predicted accuracy and dependability of different machine learning

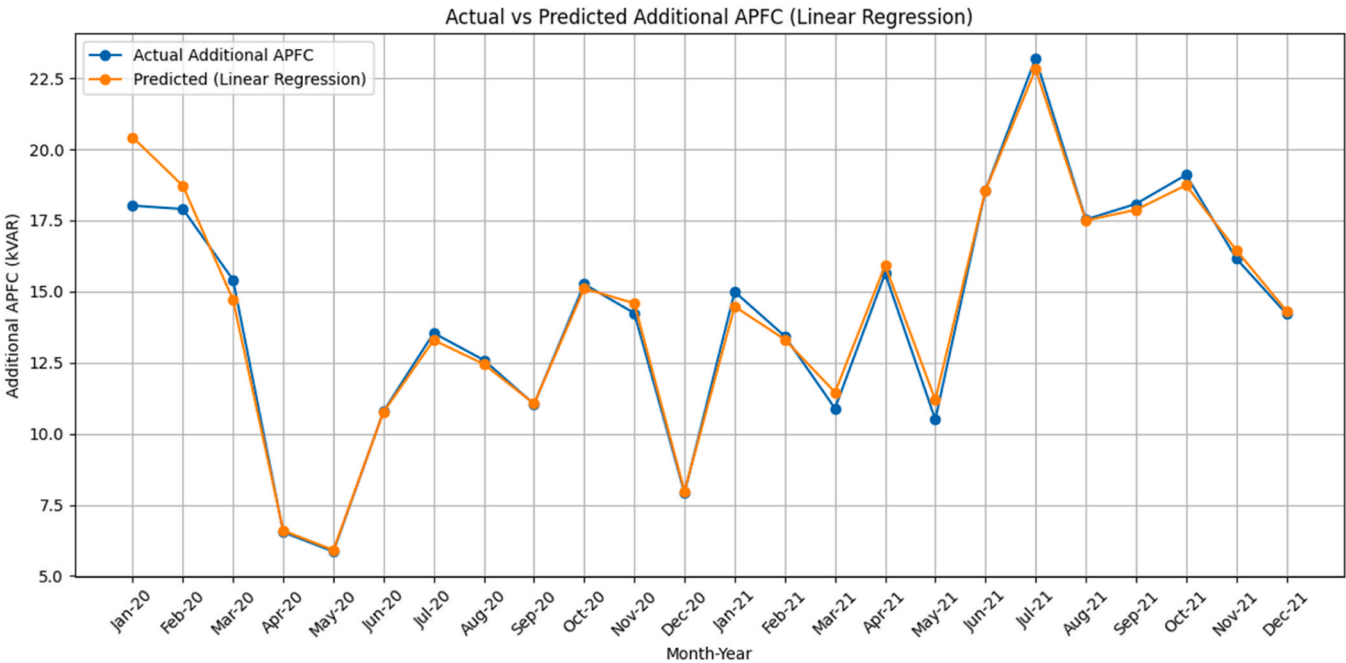


Fig. 13. Actual vs predicted APFC (LR).

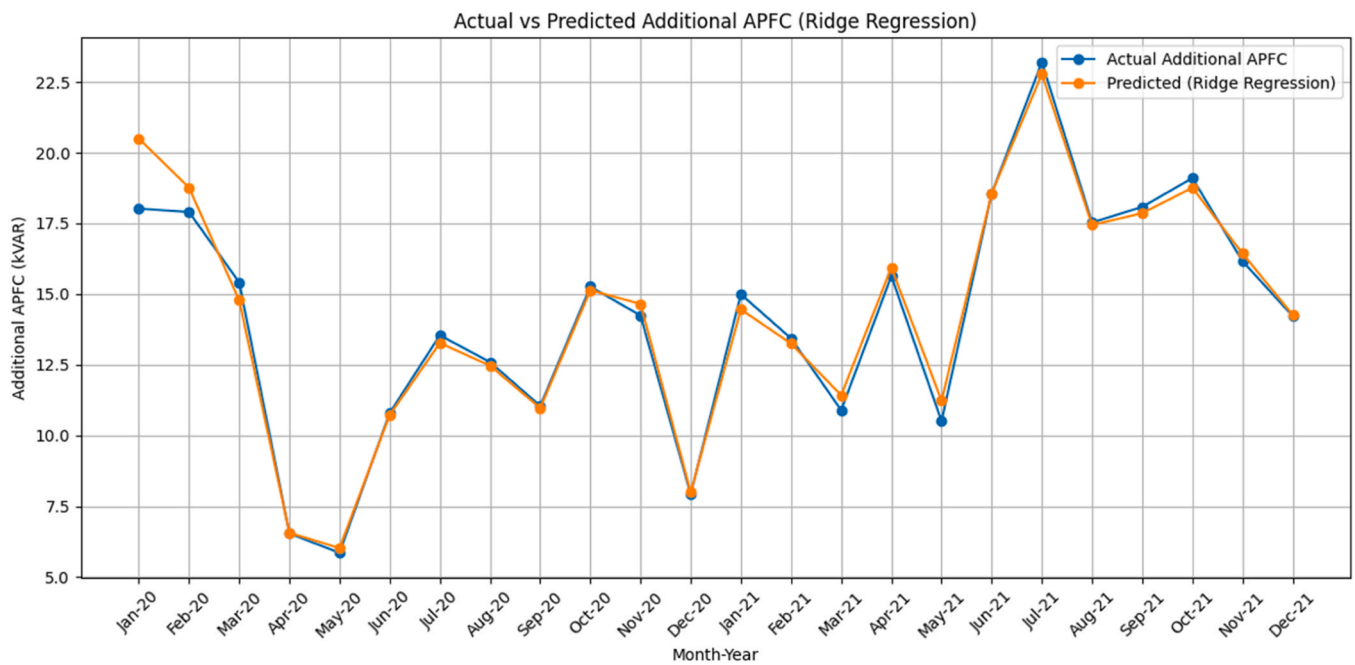


Fig. 14. Actual vs predicted APFC (Ridge).

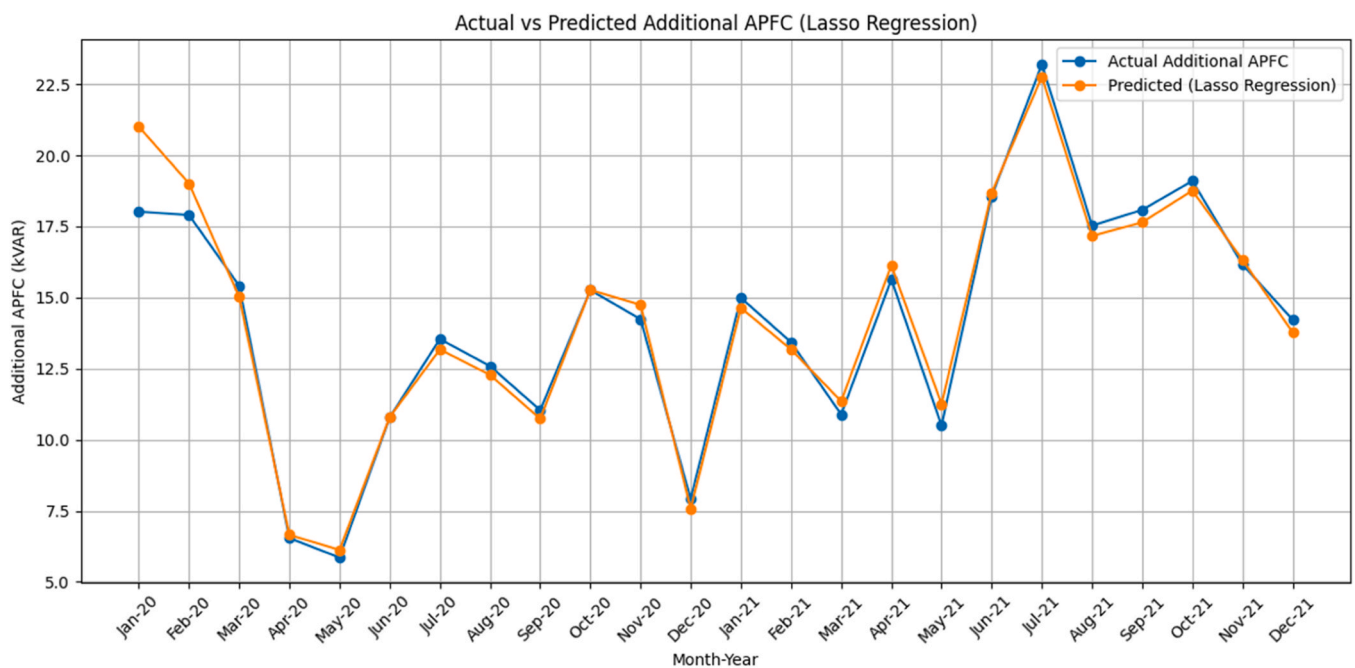


Fig. 15. Actual vs predicted APFC (Lasso).

algorithms when it comes to predicting the extra Automatic Power Factor Correction (APFC) needed for rooftop PV plants that are linked into the grid.

Linear regression shows the highest degree of prediction accuracy, with an MSE of 1.278 and an R-squared value of 0.96. The model that minimises error the best among all the models that were looked at is the one with the lowest RMSE and MAE values. Linear regression's accuracy is further supported by its MAPE of 4.497 %, which shows that it can predict the additional APFC needed with the least amount of deviation from actual data. Ridge regression also performs quite well, with an MSE of 1.368 and an R-squared value of 0.957, placing it just behind Linear regression. Despite having significantly higher error metrics than Linear

Regression, Ridge Regression offers significant prediction accuracy, with an RMSE of 1.17 and an MAE of 0.752. Although it is not as precise as linear regression, with a MAPE of 4.796 %, it nevertheless produces rather accurate predictions. In comparison to the top two models, Lasso Regression is effective but has a lower R-squared value of 0.938 and a higher MSE of 1.98. The bigger RMSE and MAE values suggest that there is more prediction error. With a MAPE of 6.559 %, the Lasso Regression increases the unpredictability of its forecasts.

The performance of Lasso Regression where the predictors are used to predict power factor and more APFC The first key aspect is that Lasso has its L1 regularization in helping to overcome overfitting by setting less important coefficients to zero, leading to feature selection. This

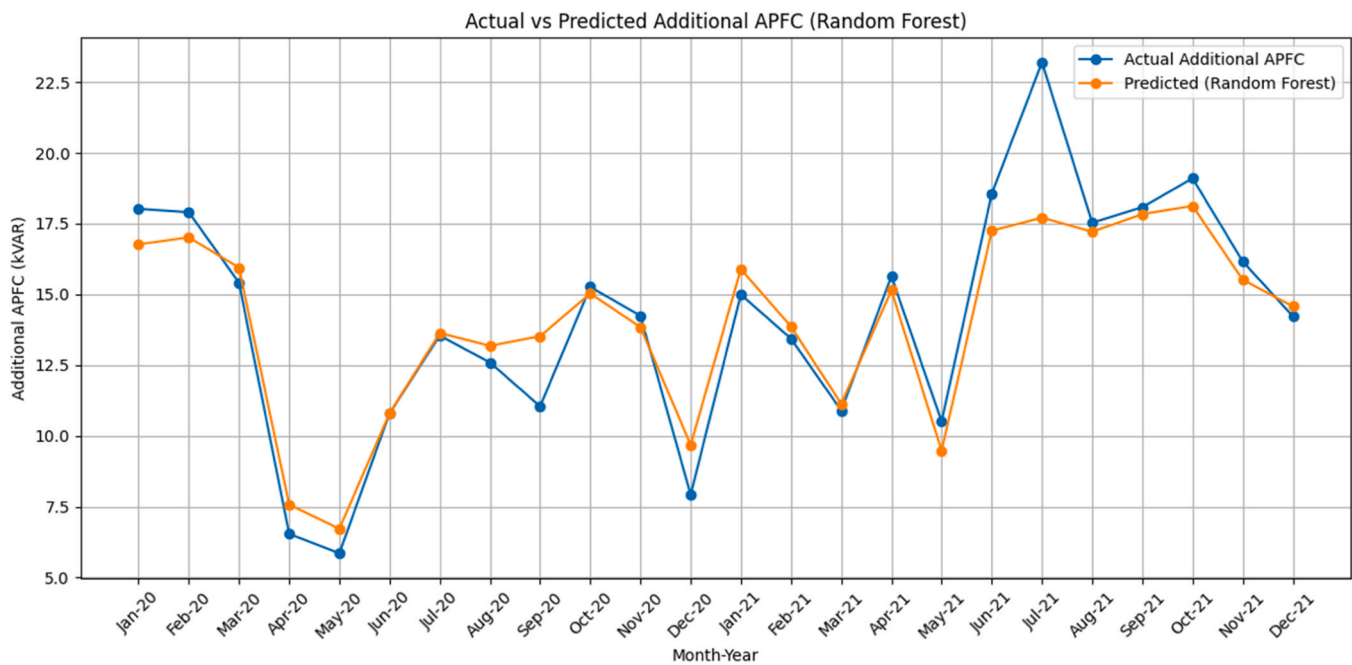


Fig. 16. Actual vs predicted APFC (RF).

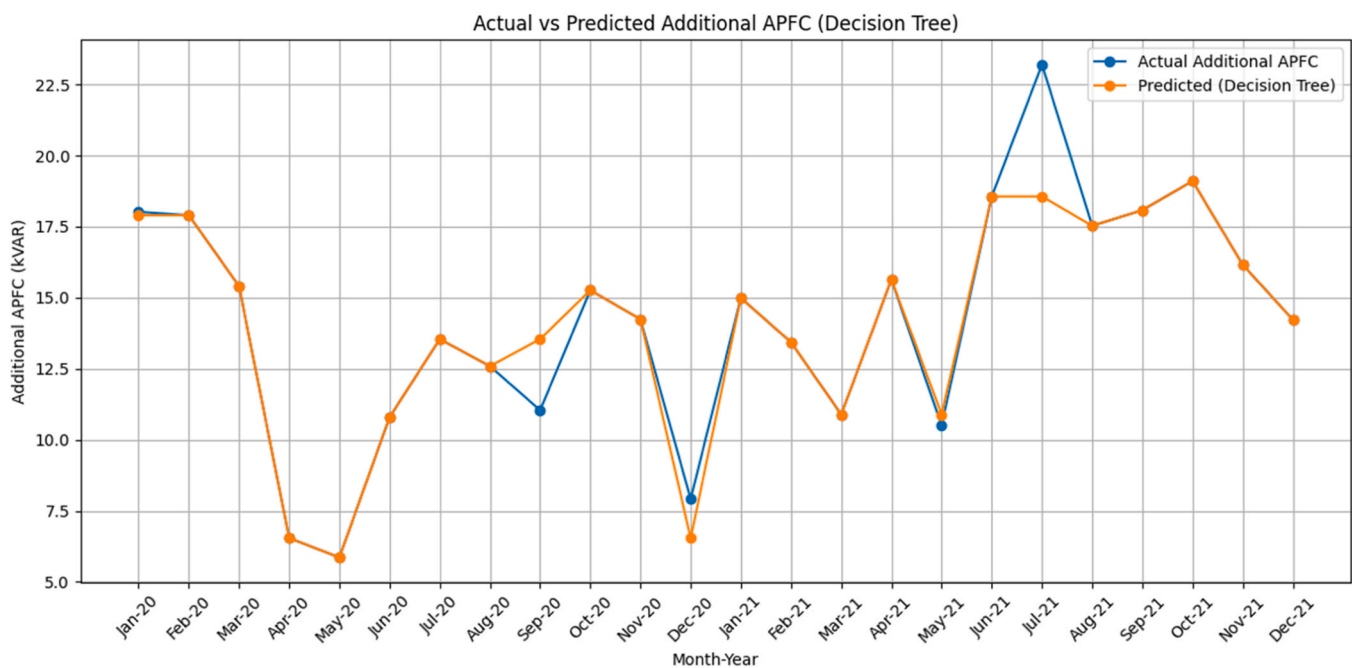


Fig. 17. Actual vs predicted APFC (DT).

means that the model now focuses only on the relevant predictors; therefore, the model would be less complex, hence facilitating better generalization. In the context of power factor and APFC prediction, where the dataset may contain noisy or redundant variables, Lasso helps eliminate unnecessary complexity and ensures more accurate predictions, particularly in real-world applications with fluctuating and noisy data. Lasso is also capable of handling multicollinearity in predictor variables—a problem in which the predictor variables are highly correlated. The automatic elimination of redundant predictors ensures that the multicollinearity does not harm the model, thus allowing for higher predictive accuracy. This is particularly valuable in the prediction of power factor and APFC when they are often interrelated variables

such as energy consumption, grid supply, and solar irradiance. Low values of MSE, RMSE, MAE, and MAPE indicate strong predictive accuracy and stability of Lasso Regression. An R-squared value of 0.888 facilitates the capture of considerable variance in data while making accurate predictions in the energy management application. Even in the presence of small errors, the impact on energy efficiency and stability will be tremendous in the grid. To reflect this idea, minimal prediction errors translate the ability of Lasso Regression to track changes even under fluctuations. It also greatly prevents overfitting, common when models learn noise in the training data and hence yield poor performance on unseen data. Generalization capability is important in this energy system domain as patterns in data may shift due to variations in

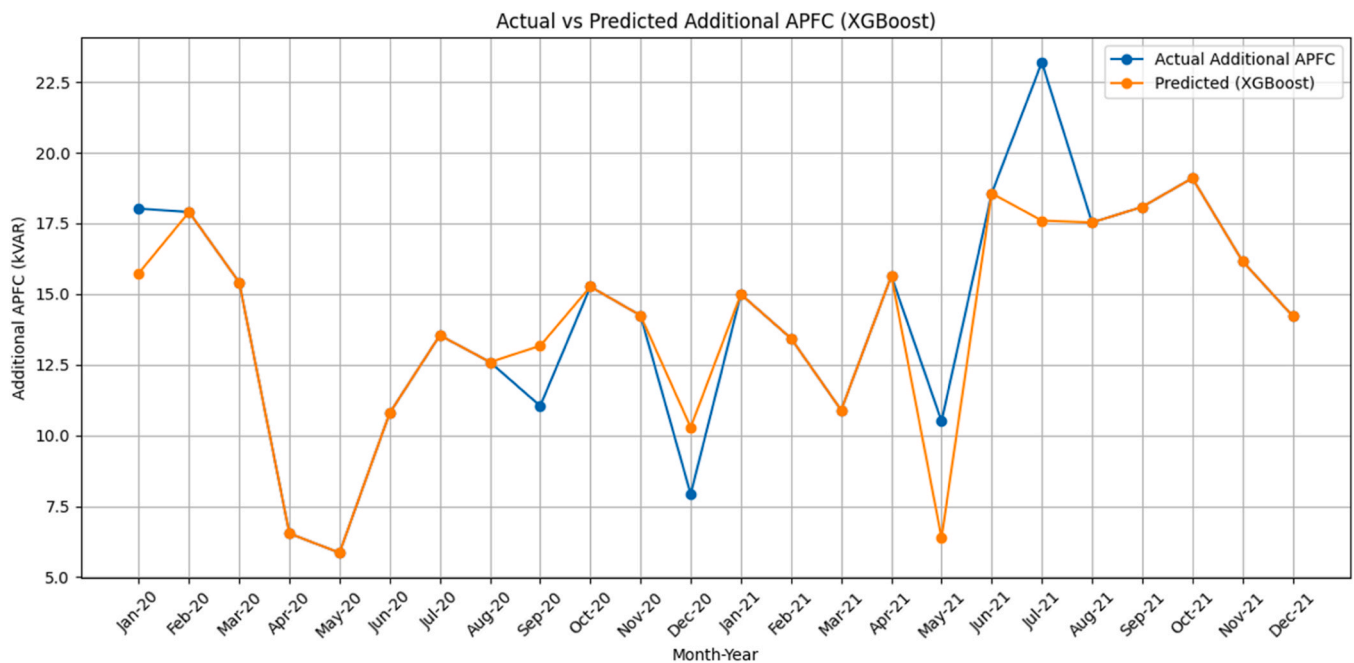


Fig. 18. Actual vs predicted APFC (XGBoost).

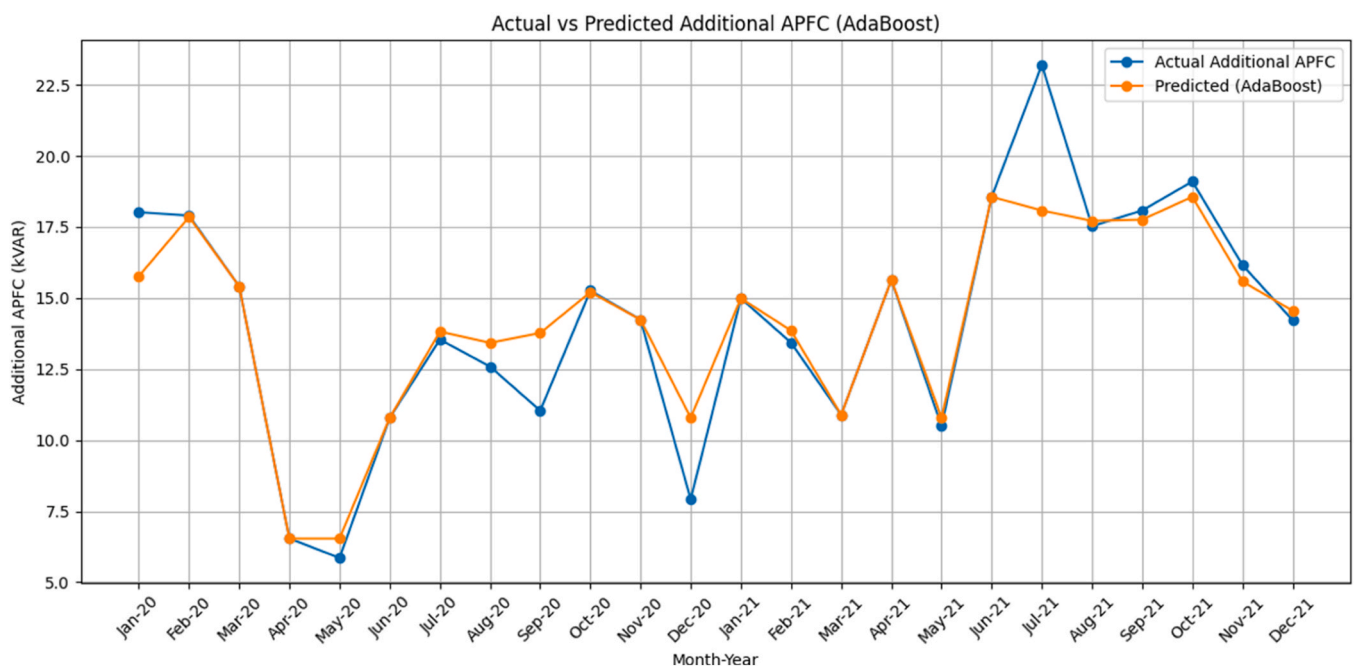


Fig. 19. Actual vs predicted APFC (Gradient boosting).

the operational or seasonal regime. Reduced model complexity by Lasso is of great importance for safe predictors in both power factor and APFC to be really relevant for practical forecasting applications. Simplicity and interpretability are some other reasons behind its success: In the energy system, the question of how energy consumption or solar irradiance impacts power factor and APFC requirements is crucial. Lasso's feature selection makes it easier to interpret how each predictor impacts the model's output and gives actionable information to stakeholders. Lastly, Lasso is quite efficient at tracking fluctuations, thus capturing general trends as well as fluctuations of seasonal patterns of power factors and APFC requirements. It shows consistency in long-run trend tracking even though slight lags during sudden fluctuations-increase of

APFC requirements in this case-occurred. This is why Lasso can be considered reliable while predicting energy efficiency and grid stability, especially in solar PV systems. This is how Lasso Regression excels: with an integrated feature of regularization, feature selection, and resistance to overfitting. The model is thus very ideal for the prediction of power factor and APFC in grid-integrated solar PV systems as it directly optimizes energy management for better stability and efficiency in the grid.

Conversely, the models of Random Forest and Decision Tree perform worse. Random Forest has an MSE of 8.397 and an R-squared value of 0.735, making it difficult to manage higher error rates, such as an RMSE of 2.898 and an MAE of 2.401. Its 17.008 % MAPE further demonstrates how inaccurate its estimates are. The Decision Tree model also shows

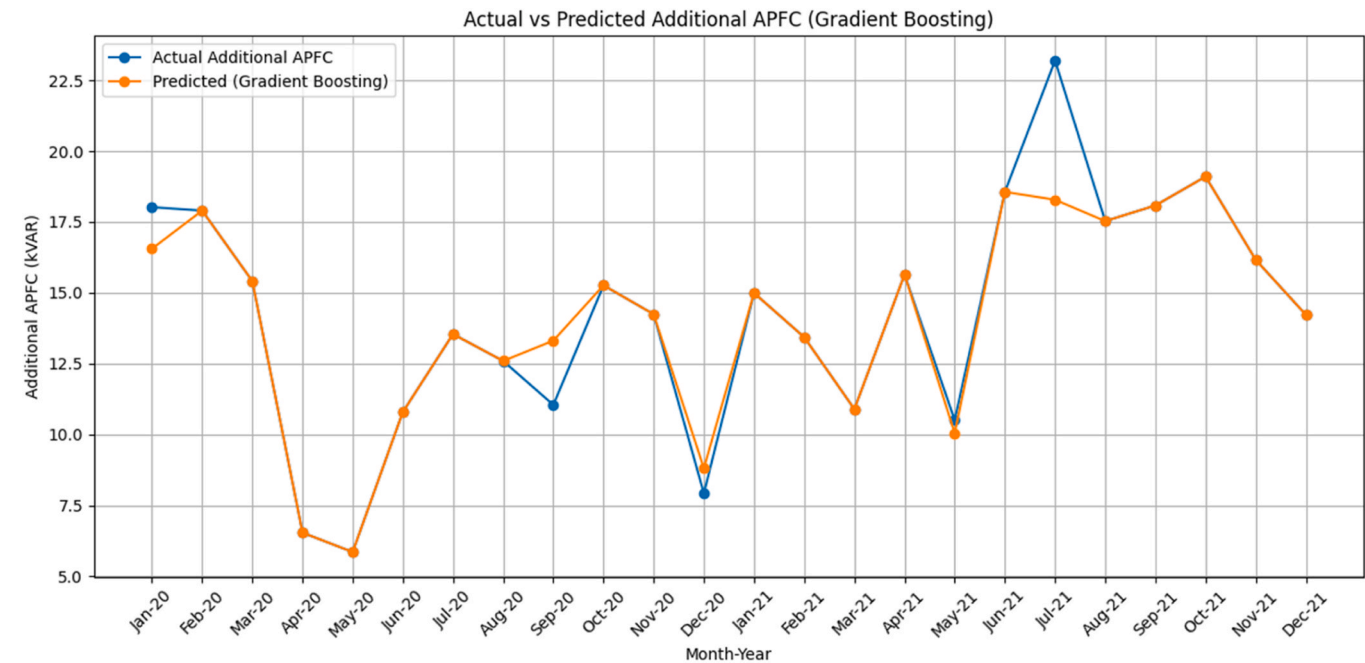


Fig. 20. Actual vs predicted APFC (Adaboost).

Table 2
Show's evaluation matrices of additional AFPC required.

Model	MSE	R-sq	RMSE	MAE	MAPE (%)
Linear Regression	1.278	0.96	1.131	0.711	4.497
Ridge Regression	1.368	0.957	1.17	0.752	4.796
Lasso Regression	1.98	0.938	1.407	0.966	6.559
Random Forest	8.397	0.735	2.898	2.401	17.008
Decision Tree	5.97	0.812	2.443	1.806	12.89
XGBoost	12.734	0.599	3.568	3.298	25.027
AdaBoost	9.41	0.703	3.068	2.652	19.657
Gradient Boosting	6.495	0.795	2.549	2	13.085

significant error, providing higher RMSE and MAE values than the best models, with an MSE of 5.97 and an R-squared value of 0.812. XGBoost, AdaBoost, and Gradient Boosting have the highest error metrics among the models that were evaluated. With an RMSE of 3.568, an MAE of 3.298, an MSE of 12.734, and an R-squared value of 0.599, XGBoost displays notable mistakes. There is a noticeable divergence between the actual figures and the MAPE of 25.027 %. In contrast, large MSE and RMSE values are displayed by AdaBoost and Gradient Boosting, which also have high MAPE percentages of 19.657 % and 13.085 %, respectively. These models have some predictive potential, but they don't perform as well as Ridge and Linear regression. Ridge regression and linear regression are the most reliable models for forecasting increasing APFC demands because they yield estimates that are accurate, consistent, and low in error. Conversely, Random Forest, Decision Tree, XGBoost, AdaBoost, and Gradient Boosting are less effective methods because to their higher mistake rates and less accuracy. Ridge regression or linear regression might be used to optimise the energy efficiency of grid-integrated rooftop PV systems and raise the accuracy of capacitor bank forecasts. The study reveals that all models provide acceptable estimates of the true power factor with varying degrees of precision. Lasso Regression, XGBoost, and Gradient Boosting are the most effective for predicting power factor, with Lasso Regression being the top performer across all evaluation metrics. These models can be integrated into energy management systems to accurately determine capacitor bank requirements, thus enhancing the energy efficiency of grid-connected rooftop solar systems. For predicting additional APFC requirements, Lasso Regression also performed best, closely following

actual values and trends. Integrating models like Lasso Regression into energy management systems can significantly improve the precision of capacitor bank predictions, leading to enhanced efficiency in grid-integrated rooftop PV systems.

5. Conclusion

The objective of this study is to improve the power factor and support the extra APFC required for preserving unitary value in order to maximise the energy efficiency of the 100 kWp Rooftop Solar Power Plant at MITS, Gwalior. Several machine learning models, including Decision Trees, Random Forests, and Linear Regression, are used to achieve these objectives. XGboost, Gradient Boosting Ridge regression, AdaBoost (adaptive threshold model), and Lasso were utilised to forecast the necessary APFC and grid supply power factor. Similar trend was maintained when predicting with the Lasso regression model, for additional APFC requirement to achieve it at unity power factor. The model also predicted the real APFC values for different months to have high energy efficiency. Precise APFC forecasting is necessary because both under-compensation and over-compensation are costly inefficiencies. The model performance was evaluated based on several different metrics, including but not limited to MSE, RMSE, R², MAE and Mape. Lasso Regression was always at the top of all models and accounted for approximately 89 % variance in the power factor data. Ensemble models (XGBoost and Gradient Boosting) have also got good scores but little lower than Lasso weighted stacking. In comparison, models like Decision Tree and Ridge Regression had higher error metrics primarily because they were overfitted and additionally performed worse in generalization as the lower R² indicates.

The applied implications of this research are considerable. Using best performing models mainly Lasso Regression and integrating it into the Energy Management system at MITS, Gwalior can yield in optimal power factor correction for an institute. This is to shelve the energy system running at its highest efficiency, curb losses in electricity and reduce tariffs whilst reducing environmental impacts of carbon emissions. In combination with these predictive analytics the hybrid energy system does not only increase reliability but also becomes a real benchmark for efficient decedents and new ways of managing power in universities. Future effort Ideally, future research would build better

models in these areas — ones which are based on a larger dataset and allow for real-time data to be input into the model so it can explore novel storage solutions that automatically respond as energy demand and supply immediately change. Furthermore, real-time adaptive control systems can be incorporated for superior solutions that may offer precision of power factor correction requirements and more cost effectiveness in greater portion. Finally, this study concludes with a significant contribution to energy management on sustainability. The study has developed and implemented a robust power factor correction based on advanced machine learning techniques that can be adopted by the colleges like MITS, Gwalior growing Solar Rooftop Power plant looking for better energy management strategies. The utilization of these predictive models in energy management systems actually represents a forward looking strategic direction for future sustainable energy solutions.

Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A2C1004743).

CRediT authorship contribution statement

Vivek Saxena: Visualization, Conceptualization. **Saibal Manna:**

Formal analysis, Data curation. **Mohammed H. Alsharif:** Writing – review & editing, Validation, Investigation. **Mun-Kyeom Kim:** Writing – review & editing, Funding acquisition, Conceptualization. **Deepansh Kulshrestha:** Project administration, Data curation. **Nikhil Paliwal:** Supervision, Investigation. **Saurabh Kumar Rajput:** Writing – original draft, Validation, Software, Resources, Methodology, Formal analysis, Conceptualization.

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.

Acknowledgement

The authors would like to express their sincere gratitude to Prakash Enterprises, Gwalior, for their invaluable support in facilitating the research work. Their generous assistance and cooperation provided crucial resources and guidance, greatly contributing to the successful completion of this study.

APPENDIX A. : Specification details of PV inverter

Section	Parameter	Specification
Model	Model	SG50KTL-M–20
Model	S/N	A1812191543
DC-Input	Max. Input Voltage	DC 1000 V
DC-Input	Min. MPP Voltage	DC 300 V
DC-Input	Max. MPP Voltage	DC 950 V
DC-Input	Max. Input Current	DC 4 * 28 A
DC-Input	Isc PV	DC 4 * 35 A
AC-Output	Rated Output Voltage	3/N/PE AC 400/230 V
AC-Output	Rated Output Frequency	50/60 Hz
AC-Output	Max. Output Current	AC 80 A
AC-Output	Rated Output Power	50000 W
AC-Output	Max. Apparent Power	50000 VA

APPENDIX B. : Specification details of PV Module

Parameter		Specification
Rated Peak Power	P_{mpp}	330 W
Open Circuit Voltage	V_{oc}	46.3 V
Short Circuit Current	I_{sc}	9.24 A
Rated Voltage	V_{mpp}	38.0 V
Rated Current	I_{mpp}	8.70 A
Fill Factor	FF	77.28 %
Efficiency	%	17.00 %
Fire Rating	$Type-1$	Class C

Irradiation 1000 W/m², AM 1.5, 25°C Cell Temperature

Data availability

Data will be made available on request.

References

- AlShafeey, M., Csák, C., 2021. Evaluating neural network and linear regression photovoltaic power forecasting models based on different input methods. *Energy Rep.* 7, 7601–7614. <https://doi.org/10.1016/j.egy.2021.10.125>.

- Alsharif, M.H., 2017. Comparative analysis of solar-powered base stations for green mobile networks. *Energies* 10, 1208. <https://doi.org/10.3390/en10081208>.
- Alsharif, M.H., Jahid, Abu, Kannadasan, Raju, Kim, Mun-Kyeom, 2024. Unleashing the potential of sixth generation (6G) wireless networks in smart energy grid management: a comprehensive review. *Energy Rep.* 11, 1376–1398. <https://doi.org/10.1016/j.egy.2024.01.011>.
- Alsharif, M.H., Kannadasan, Raju, Jahid, Abu, Albreem, Mahmoud A., Nebhen, Jamel, Choi, Bong Jun, 2021. Long-term techno-economic analysis of sustainable and zero grid cellular base station. *IEEE Access* 9, 54159–54172. <https://doi.org/10.1109/ACCESS.2021.3071250>.
- Alsharif, M.H., Nordin, Rosdiadee, Ismail, Mahamod, 2016. Green wireless network optimisation strategies within smart grid environments for Long Term Evolution (LTE) cellular networks in Malaysia. *Renew. Energy* 85, 157–170. <https://doi.org/10.1016/j.renene.2015.06.044>.
- Anon, 2023. Predicting PV system parameters using machine learning: a systematic review (Jun.). *Heliyon* 9 (6), e16815. <https://doi.org/10.1016/j.heliyon.2023.e16815>.
- Ansari, S., Ayob, A., Lipu, M.S.H., Saad, M.H.M., Hussain, A., 2021. A review of monitoring technologies for solar PV systems using data processing modules and transmission protocols: progress, challenges and prospects (Jul.). *Sustainability* 13 (15), 8120. <https://doi.org/10.3390/su13158120>.
- Benti, N.E., Chaka, M.D., Semie, A.G., 2023. Forecasting renewable energy generation with machine learning and deep learning: current advances and future prospects. *Sustainability* 15, 7087. <https://doi.org/10.3390/su15097087>.
- Carneiro, T.C., Rocha, P.A.C., Carvalho, P.C.M., Fernández-Ramírez, L.M., 2022. Ridge regression ensemble of machine learning models applied to solar and wind forecasting in Brazil and Spain. *Appl. Energy* 314, 118936. <https://doi.org/10.1016/j.apenergy.2022.118936>.
- Cattani, G., 2023. Combining data envelopment analysis and random forest for selecting optimal locations of solar PV plants. *Energy AI* 11, 100222. <https://doi.org/10.1016/j.egyai.2022.100222>.
- Cortes-Vega, D., Alazki, H., 2021. Robust maximum power point tracking scheme for PV systems based on attractive ellipsoid method ([Online]. Available). *Sustain. Energy, Grids Netw.* 25, 100410. <https://doi.org/10.1016/j.segan.2020.100410>.
- Dondariya, C., Porwal, D., Awasthi, A., Shukla, A.K., Sudhakar, K., M. M. S. R., Bhimte, A., 2018. Performance simulation of grid-connected rooftop solar PV system for small households: a case study of Ujjain, India (Aug.). *Energy Rep.* 4, 555–564. <https://doi.org/10.1016/j.egy.2018.08.002>.
- Erten, M.Y., 2022. Solar power prediction using regression models (pp. ?, Dec.). Ulus. Muhendis. Arastirma ve Gelistirme Derg. 10 (?). <https://doi.org/10.29137/umagd.1100957>.
- Gómez Medina, J.M., de la Torre y Ramos, J., López Monteagudo, F.E., Ríos Rodríguez, Ld.C., Esparza, D., Rivas, J.M., Ruvalcaba Arredondo, L., Romero Moyano, A.A., 2022. Power factor prediction in three phase electrical power systems using machine learning. *Sustainability* 14, 9113. <https://doi.org/10.3390/su14159113>.
- Gusman, L.S., Pereira, H.A., Callegari, J.M.S., Cupertino, A.F., 2020b. Design for reliability of multifunctional PV inverters used in industrial power factor regulation. *Int. J. Electr. Power Energy Syst.* 119, 105932. <https://doi.org/10.1016/j.ijepes.2020.105932>.
- Gusman, L.S., Pereira, H.A., Callegari, J.M.S., Cupertino, A.F., 2020a. Design for reliability of multifunctional PV inverters used in industrial power factor regulation ([Online]. Available). *Int. J. Electr. Power Energy Syst.* 119, 105932. <https://doi.org/10.1016/j.ijepes.2020.105932>.
- Kalyuzhny, A., Reshef, B., Yehuda, G., Day, T., [additional authors], 2013. Design of capacitor bank in parallel to photovoltaic power plant. In *Eurocon 2013. IEEE Conference*. <https://doi.org/10.1109/EUROCON.2013.6625156>.
- Kassem, Y., Gökçekuş, H., Güvensoy, A., 2021a. Techno-economic feasibility of grid-connected solar pv system at near east university hospital, Northern Cyprus (Nov.). *Energies* 14 (22), 7627. <https://doi.org/10.3390/en14227627>.
- Kassem, Y., Gökçekuş, H., Güvensoy, A., 2021b. Techno-economic feasibility of grid-connected solar PV System at near east university hospital, Northern Cyprus (article). *Energies* 14 (22), 7627. <https://doi.org/10.3390/en14227627>.
- Lee, Da-Sheng, Lai, Chih-Wei, Fu, Shih-Kai, 2024. A short- and medium-term forecasting model for roof PV systems with data pre-processing (Mar.). *Heliyon* 10 (6), e27752. <https://doi.org/10.1016/j.heliyon.2024.e27752>.
- Leghari, Z.H., Kumar, M., Shaikh, P.H., Kumar, L., Tran, Q.T., 2022. A critical review of optimization strategies for simultaneous integration of distributed generation and capacitor banks in power distribution networks. *Energies* 15, 8258. <https://doi.org/10.3390/en15218258>.
- Li, D., Zhu, D., Tao, T., Qu, J., 2024. Power generation prediction for photovoltaic system of hose-drawn traveler based on machine learning models. *Processes* 12 (1), 39. <https://doi.org/10.3390/pr12010039>.
- Lodhi, E., Wang, F.-Y., Xiong, G., Dilawar, A., 2022. An adaboost ensemble model for fault detection and classification in photovoltaic arrays (Jan.). *IEEE J. Radio Freq. Identif.* 1-1. <https://doi.org/10.1109/JRFID.2022.3212310>.
- Mahesh, P.V., Meyyappan, S., Alla, R.K., 2022. Maximum power point tracking using decision-tree machine-learning algorithm for photovoltaic systems (Oct.). *Clean. Energy* 6 (5), 762–775. <https://doi.org/10.1093/ce/zkac057>.
- Mitra, I., Heinemann, D., Ramanan, A., et al., 2022. Short-term PV power forecasting in India: recent developments and policy analysis. *Int. J. Energy Environ. Eng.* 13, 515–540. <https://doi.org/10.1007/s40095-021-00468-z>.
- Palahalli, H., Maffezzoni, P., Daniel, L., Grusso, G., 2022. Statistical analysis of PV penetration impact on residential distribution grids ([Online]. Available). *Sustain. Energy, Grids Netw.* 32, 100949. <https://doi.org/10.1016/j.segan.2022.100949>.
- Preda, S., Oprea, S.-V., Băra, A., Belciu, A., 2018. PV forecasting using support vector machine learning in a big data analytics context. *Symmetry* 10, 748. <https://doi.org/10.3390/sym10120748>.
- Qinghe, Z., Xiang, W., Huang, B., Wang, J., Fang, J., 2022. Optimised extreme gradient boosting model for short term electric load demand forecasting of regional grid system (Nov.). *Sci. Rep.* 12 (19282), 1. <https://doi.org/10.1038/s41598-022-22024-3>.
- Roy, A., Verma, V., Gampa, S.R., Bansal, R.C., 2023. Planning of distribution system considering residential roof top photovoltaic systems, distributed generations and shunt capacitors using gravitational search algorithm. *Comput. Electr. Eng.* 111 (B), 108960. <https://doi.org/10.1016/j.compeleceng.2023.108960>.
- Sarmah, P., Das, D., Saikia, M., Kumar, V., Yadav, S.K., Paramasivam, P., Dhanasekaran, S., 2023. Comprehensive analysis of solar panel performance and correlations with meteorological parameters (Dec.). *ACS Omega* 8 (50), 47897–47904. <https://doi.org/10.1021/acsomega.3c06442>.
- Shabbir, N., Kütt, L., Astapov, V., Daniel, K., Jawad, M., Husev, O., Rosin, A., Martins, J., 2024. Enhancing PV hosting capacity and mitigating congestion in distribution networks with deep learning based PV forecasting and battery management ([Online]. Available). *Appl. Energy* 372, 123770. <https://doi.org/10.1016/j.apenergy.2024.123770>.
- Shafiqullah, M., Ahmed, S.D., Al-Sulaiman, F.A., 2022. Grid integration challenges and solution strategies for solar PV systems: a review. *IEEE Access* 10, 3174555. <https://doi.org/10.1109/ACCESS.2022.3174555>.
- Shareef, H., Mutlag, A.H., Mohamed, A., 2017. Random forest-based approach for maximum power point tracking of photovoltaic systems operating under actual environmental conditions (Article ID). *Hindawi J. Renew. Energy* 2017, 1673864. <https://doi.org/10.1155/2017/1673864>.
- Shuvo, Md. Bengir, Ahmed, Mohammad Asaduzzaman, Chowdhury, Shameem, Ahmed, Mohammad Abul, Kashem, 2019. Prediction of solar irradiation and performance evaluation of grid connected solar 80KWp PV plant in Bangladesh (Jul.). *Energy Rep.* 5, 1354–1363. <https://doi.org/10.1016/j.egy.2019.06.011>.
- Smaili, I.H., Almalawi, D.R., Shaheen, A.M., Mansour, H.S.E., 2024. Optimizing PV sources and shunt capacitors for energy efficiency improvement in distribution systems using subtraction-average algorithm. *Mathematics* 12, 625. <https://doi.org/10.3390/math12050625>.
- Solar Power Prediction using Regression Models, 2022. *Uluslararası Mühend.-. Araştırma ve Geliştirme Derg.* <https://doi.org/10.29137/umagd.1100957>.
- Tang, N., Mao, S., Wang, Y., Nelms, R.M., 2018. Solar power generation forecasting with a LASSO-based approach (Apr.). *IEEE Internet Things J.* 5 (2), 1090–1099. <https://doi.org/10.1109/JIOT.2018.2812155>.
- Vernica, I., Wang, H., Blaabjerg, F., 2018. Uncertainty analysis of capacitor reliability prediction due to uneven thermal loading in photovoltaic applications. *Microelectron. Reliab.* 88–90, 1036–1041. <https://doi.org/10.1016/j.microrel.2018.07.059>.
- Wang, J., Li, P., Ran, R., Che, Y., Zhou, Y., 2018. A short-term photovoltaic power prediction model based on the gradient boost decision tree (May). *Appl. Sci.* 8 (5), 689. <https://doi.org/10.3390/app8050689>.
- Zou, Miao, Jiang, Wu-Gui, Qin, Qing-Hua, Liu, Yu-Cheng, Li, Mao-Lin, 2022. Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting (Aug.). *Materials* 15 (15), 5298. <https://doi.org/10.3390/ma15155298>.