IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 April 2025, accepted 23 May 2025, date of publication 27 May 2025, date of current version 4 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3574093

RESEARCH ARTICLE

Solar Energy Forecasting Using Machine Learning Techniques for Enhanced Grid Stability

ATTULURI R. VIJAY BABU[1], N. BHARATH KUMAR[1], (Member, IEEE), RAJANAND PATNAIK NARASIPURAM[2], SOUNDHAR PERIYANNAN[3], ALIREZA HOSSEINPOUR[4], AND AYMEN FLAH[5,6,7,8]

[1]Department of Electrical and Electronics Engineering, Vignan's Foundation for Science Technology and Research, Guntur, Andhra Pradesh 522213, India
[2]Energy Group, Cyient Ltd., Pune, Maharashtra 411028, India
[3]Atria Power, Bengaluru, Karnataka 560025, India
[4]Department of Electrical Engineering, University of Zabol, Zabol 98615-538, Iran
[5]National Engineering School of Gabes, University of Gabes, Gabes 6029, Tunisia
[6]Applied Science Research Center, Applied Science Private University, Amman 11931, Jordan
[7]Jadara University Research Center, Jadara University, Irbid 21110, Jordan
[8]ENET Centre, CEET, VSB-Technical University of Ostrava, 708 00 Ostrava, Czech Republic

Corresponding authors: Alireza Hosseinpour (a.hoseinpour@uoz.ac.ir) and Aymen Flah (flahaymening@yahoo.fr)

This work was supported in part by European Union through the REFRESH—Research Excellence for Region Sustainability and High-Tech Industries Project via the Operational Program Just Transition under Grant CZ.10.03.01/00/22_003/0000048, in part by the National Centre for Energy II and ExPEDite Project a Research and Innovation Action to Support the Implementation of the Climate Neutral and Smart Cities Mission under Project TN02000025, and in part by ExPEDite through European Union's Horizon Mission Program under Grant 101139527.

ABSTRACT The increasing integration of solar photovoltaic (PV) systems into modern energy grids presents significant challenges due to the intermittent and weather-dependent nature of solar energy generation. Accurate short-term forecasting is essential to ensure grid stability and optimize energy resource allocation. This study proposes a comprehensive data-driven framework for solar energy forecasting using multiple machine learning (ML) techniques, including Multiple Linear Regression, Ridge, Lasso, Decision Tree Regression, Support Vector Regression, and ensemble-based models such as Random Forest, AdaBoost, Bagging, and Gradient Boosting Regressors. The framework incorporates advanced feature engineering using high-resolution meteorological and solar geometric parameters-such as relative humidity, temperature, cloud cover, zenith angle, azimuth, and angle of incidence-to enhance model accuracy. Historical solar power and weather datasets were used to train and evaluate the models across multiple performance metrics. Among the models, the Gradient Boosting Regressor demonstrated the best performance, achieving an R^2 of 0.827, RMSE of 399.44, and MAE of 253.62, marking a significant improvement over baseline models. The study also evaluates model robustness and discusses feature relevance, hyperparameter optimization strategies, and deployment considerations for real-time grid operations. These findings provide practical insights for stakeholders aiming to implement intelligent solar forecasting systems in smart grid environments, thereby contributing to enhanced energy management and grid resilience.

INDEX TERMS Ensemble learning techniques, machine learning, renewable energy sources, solar energy forecasting.

NOMENCLATURE AND ABBREVIATIONS

Acronym	Definition
PV	Photovoltaic
ML	Machine Learning
DL	Deep Learning
ANN	Artificial Neural Network
SVR	Support Vector Regression
RF	Random Forest
GBR	Gradient Boosting Regressor
SCADA	Supervisory Control and Data Acquisition
MAE	Mean Absolute Error

The associate editor coordinating the review of this manuscript and approving it for publication was Meng Huang.

I. INTRODUCTION

Considerable momentum has been garnered for the integration of solar energy into the energy grid [1], [2], owing to its environmental benefits and continually decreasing costs. However, the intrinsically unpredictable nature of solar energy output, which is impacted by various factors such as the weather conditions [3], provides obstacles for the smooth integration [4] of solar energy into the grid. For the flexible grid control [5], it is necessary to have accurate forecasts of solar energy generation in order to guarantee grid stability, maximize the allocation of energy resources, and improve the economic sustainability of solar energy systems.

When it comes to capturing the complex dynamics of solar energy output, which are influenced by cloud cover, meteorological conditions, and the position of the sun, traditional forecasting systems [6], [7] have a tough time doing so. Machine learning techniques [8], [9], [10], [11], [12] have emerged as promising tools for improving the accuracy of solar energy projections. However, most existing studies either overlook the integration of solar geometric parameters-such as the angle of incidence, zenith angle, and azimuth-or evaluate models without a consistent training-validation pipeline under real-world variability. By utilizing computational methods, machine learning makes it possible to discover patterns and relationships from historical data. This, in turn, makes it easier to construct models that are capable of making predictions that are more trustworthy. Various ML and DL algorithms have been applied to solar forecasting including SVR, Decision Trees, Random Forests, and Artificial Neural Networks (ANN). In DL, CNN, LSTM, and hybrid models have been investigated. ANN training techniques such as backpropagation, Adam optimizer, and dropout regularization are often used to mitigate overfitting and improve generalization.

Conventional physical models for solar forecasting, such as numerical weather prediction (NWP) models and radiation transfer models, are based on complex equations describing atmospheric physics. While they are theoretically sound, their dependency on precise input parameters and high computational overhead limits their responsiveness and real-time applicability. These challenges have led to the increasing adoption of data-driven approaches such as ML and DL.

The purpose of this work is to investigate the use of machine learning approaches [13], [14], [15], [16], [17], [18] to improve solar energy forecasting in order to overcome the issues that are brought about by the intermittent nature of solar energy. The advantage of machine learning algorithms is that they are able to learn continuously and adapt to new information. This adaptability is particularly beneficial when forecasting under dynamic environmental conditions, making ML models suitable for real-time applications.

These methodologies include artificial neural networks, support vector machines, random forests, and deep learning architectures. Through the process of training these models on historical data pertaining to solar energy generation and relevant meteorological factors, our objective is to gain a better understanding of the intricate correlations that exist between the input parameters and the output of energy. This study extends prior works by systematically comparing these models across multiple metrics and hyperparameter settings, using a unified dataset and evaluation protocol to ensure fairness. Additionally, the incorporation of real-time data from weather forecasts results in an increase in the correctness of predictions.

By integrating both meteorological and geometric solar features, this research contributes a robust, scalable, and interpretable forecasting pipeline suitable for deployment in smart grid systems. This will be accomplished by conducting a comprehensive assessment of previous research. Through this analysis, the researchers hope to make a contribution to the larger conversation on how to improve the dependability and efficiency of solar energy integration into the electricity grid.

The models developed are based on actual operational data, with simulations used to evaluate their forecasting performance under various conditions. Special emphasis is placed on analyzing model interpretability, robustness to hyperparameters, and feature relevance using statistical and visual analysis tools. To guarantee strong predictive skills, the selected models—such as random forests, artificial neural networks, support vector machines, and deep learning architectures—are trained and tested.

The investigation evaluates how various feature selections, model hyperparameters, and training times affect each machine learning approach's overall effectiveness. Feature engineering plays a key role in model generalization and forecasting accuracy. This work also discusses model performance in the context of practical implementation, offering insight into trade-offs between accuracy, complexity, and deployment feasibility.

Enhancing prediction accuracy also requires incorporating real-time data from weather predictions, which is a crucial step. In order to make sure that the forecasts are still flexible in response to shifting weather patterns, the research looks into how to smoothly integrate such real-time data into the machine learning models.

By offering a comprehensive grasp of the advantages and disadvantages of diverse machine learning approaches (see Table 1), this research seeks to advance the rapidly developing field of solar energy forecasting. Through clarifying the complex dynamics of solar energy generation, the research aims to provide more precise, dependable, and effective solar energy integration into the energy system. The results of this research contribute to our knowledge of machine learning applications in renewable energy and are useful for researchers, practitioners, and policymakers who aim to maximize the use of solar resources in a sustainable and commercially feasible way.

A. OBJECTIVES

The main objectives of this work are:

- To conduct a comprehensive evaluation of multiple machine learning algorithms - including artificial neural

TABLE 1. Critical literature review of ML and AI techniques for solar energy forecasting.

Study	Methods used	Key findings
Li et al. [19]	CNN, RNN	Deep learning models (CNN, RNN) significantly improved accuracy over traditional statistical methods.
Muhammad Abubakar et al. [20]	ANN, SVM, RF, LSTM, GRU	Comprehensive review of hybrid ML techniques addressing efficiency issues in smart grid-based solar systems.
Ahsan Zafar et al. [21]	Autoencoder-LSTM, CNN-LSTM	Autoencoder-LSTM models provided highest accuracy (97%) and lowest RMSE among tested deep learning configurations.
Yang et al. (2024) [22]	Spatiotemporal ML + Multi-factor Interval Constraints	Reconstruction-based framework enhanced short-term PV forecasting by modeling spatiotemporal dependencies.
Yuqi et al. (2025) [23]	Feature Extraction, Temporal Weighting (ML)	Applied temporal importance analysis for robust forecasting-adaptable to PV feature selection.
Muhammad Shoaib Bhutta et al. [24]	ANN	Applied ANN for frequency control in HVDC solar systems, showing strong adaptability under nonlinear grid dynamics.
Asif Iqbal Kawoosa et al. [25]	KNN, RF, DT, SVM, AdaBoost	Proposed XGBoost-based ETD detection; emphasizes importance of ensemble methods for energy forecasting and reliability.
Iqbal et al. [26]	XGBoost (Ensemble)	Used feature-augmented ensemble learning for high-resolution irradiance and load modeling.
Ahmed & Khalid [27]	SVM, RF with SCADA integration	Demonstrated practical ML deployment using real-time data from SCADA environments for solar forecasting.

networks, support vector machines, random forests - for forecasting solar photovoltaic (PV) power generation using real-world datasets that combine historical power output with high-resolution meteorological data.

- To enhance the forecasting accuracy and reliability of ML models by incorporating advanced feature engineering techniques, particularly the inclusion of solar geometric parameters such as zenith angle, azimuth angle, and angle of incidence, in addition to conventional weather-based features.
- To compare the predictive performance, robustness, and interpretability of classical and ensemble models under a unified training-validation protocol, identifying the most suitable model for operational deployment in smart energy systems.
- To provide practical recommendations for integrating ML-based solar forecasting models into energy management frameworks, with the aim of improving grid stability, optimizing energy dispatch, and enhancing the economic viability of renewable energy systems.

While several comparative studies exist in the literature, this research distinguishes itself by incorporating solar geometry parameters and high-resolution meteorological data to train and evaluate models under real-world Indian climatic conditions. The inclusion of diverse models ranging from classical regressors to ensemble techniques provides a holistic view of forecasting efficacy. Notably, the study extends the analysis to consider operational factors relevant to grid integration, offering practical implications for real-time deployment in smart energy systems. While this study primarily focuses on short-term solar forecasting using meteorological and solar geometric features, future research could explore the integration of hybrid thermal and electrical systems. For instance, [28] proposed a novel air-water thermoelectric module that significantly enhances net output power. Combining such physical system innovations with intelligent forecasting frameworks may enable more holistic renewable energy management strategies, particularly in microgrid environments where both thermal and electrical outputs are relevant.

II. FEATURE ENGINEERING IN SOLAR FORECASTING

Feature engineering is an important feature in improving the accuracy and dependability of solar forecasting models [29], [30]. These models are created to estimate the amount of solar energy produced by photovoltaic systems, taking into account several aspects that affect it. Within this particular context, we aim to clarify and explain the fundamental characteristics and approaches that are essential to the process of feature engineering for solar forecasting.

A. ATMOSPHERIC CONDITIONS

1) RELATIVE HUMIDITY

Several studies have demonstrated that relative humidity significantly influences solar forecasts due to its impact on atmospheric conditions and solar irradiance [31], [32], [33]. The humidity level, measured as a percentage, indicates the amount of water vapour in the air, which affects the overall climate conditions experienced by solar panels. The efficiency of solar panels is closely connected to relative humidity, as it impacts the atmospheric conditions that govern the absorption and conversion of sunlight into electricity. Moreover, the quantity of solar energy received is intrinsically linked to relative humidity levels, whereby changes in humidity affect the strength and spread of sunshine, ultimately affecting the overall efficiency of solar energy systems. Hence, a comprehensive comprehension of relative humidity is crucial for precise solar prediction, offering vital knowledge about the intricate relationship between atmospheric conditions and solar panel performance, which is critical for maximising the efficiency of solar energy utilisation.

Relative humidity also plays a vital role in analyzing various meteorological factors that affect solar panel efficiency. By factoring in humidity, forecasting models can better assess the formation of clouds and their impact on solar irradiation, as well as the influence of moisture on pollutants and dust that may reduce panel performance. Increased humidity often elevates the levels of atmospheric particulates, which can

scatter sunlight and reduce energy conversion efficiency. Additionally, high humidity can lead to the buildup of moisture on solar panels, further diminishing their performance. Incorporating humidity data into forecasting algorithms helps predict these conditions, allowing for better planning and maintenance of solar systems. This comprehensive approach strengthens the reliability and relevance of solar energy predictions, ensuring that all environmental factors—such as clouds, dust, and moisture—are accounted for in the performance of solar panels.

Using relative humidity data in solar forecasting models enhances the accuracy of energy production predictions and supports more informed decision-making. Understanding how humidity affects atmospheric conditions and cloud formation helps anticipate fluctuations in solar irradiance. This knowledge allows solar operators to plan more effectively, preventing losses in energy generation due to reduced sunlight exposure. Additionally, humidity plays a role in assessing air quality, as increased moisture can raise particulate levels, which can block sunlight and decrease panel efficiency. By anticipating these conditions, solar operators can implement strategies to maintain optimal performance. Monitoring moisture levels also supports preventive maintenance, minimizing efficiency losses due to panel contamination. Overall, integrating relative humidity data into solar forecasting improves system resilience, enables adaptive management, and enhances operational efficiency, ensuring more reliable solar energy production.

2) TEMPERATURE

Temperature plays a critical role in solar forecasting as it directly impacts the efficiency of solar panels and their energy conversion processes. Photovoltaic (PV) systems are sensitive to temperature fluctuations. High temperatures can reduce the efficiency of solar cells due to the temperature coefficient effect, where increased thermal energy disrupts the cells' ability to conduct electricity efficiently. On the other hand, lower temperatures often improve efficiency, highlighting the complex and temperature-sensitive nature of solar panels. Beyond panel efficiency, temperature also influences the overall energy output of solar systems. As temperature varies, so does the available solar radiation, affecting energy generation. Forecasting models that incorporate temperature data can better predict these variations, making it possible to optimize solar installations and adjust operational strategies to improve performance. By understanding how temperature and climatic conditions interact with solar panels, models can provide more accurate energy output predictions.

Solar panel efficiency decreases as temperatures rise, primarily due to changes in the semiconductor properties of solar cells. Higher temperatures increase electron activity within the cells, creating more resistance and lowering conductivity, which reduces the overall efficiency of energy conversion. Additionally, temperature fluctuations can alter the semiconductor bandgap, affecting how well the material absorbs and converts sunlight into electricity. Since different solar cell materials react uniquely to temperature changes, it is essential to account for these variations when evaluating the performance of PV systems. Temperature data is crucial for improving the accuracy of solar forecasting models. Including temperature as a variable allows models to account for the impact of temperature changes on energy production. This improves prediction accuracy and helps stakeholders make better-informed decisions. By integrating temperature data, forecasting models can better assess the intricate relationships between weather conditions and solar energy output, leading to more efficient resource management.

3) CLOUD COVER

The accuracy of solar energy generation forecasts heavily relies on understanding and accounting for cloud cover. Changes in cloud cover directly impact solar irradiance, the amount of sunlight reaching the Earth's surface, which in turn affects the energy captured by photovoltaic (PV) systems. Clouds scatter and absorb solar radiation, reducing both the strength and duration of sunlight available to solar panels. Therefore, it is essential for forecasting models to incorporate cloud cover data to accurately estimate solar energy potential. Clouds vary in density, thickness, and distribution, and these factors change constantly. Modern solar forecasting algorithms use advanced data analytics to capture these dynamic patterns, allowing them to make both short- and long-term predictions. Recent advancements in satellite imaging, remote sensing, and atmospheric monitoring have improved the measurement of cloud cover. By integrating this data into solar models, forecasters can reduce uncertainty and provide more reliable energy projections, which are crucial for energy planners and grid operators.

Clouds significantly affect the amount of solar radiation that reaches the Earth's surface. As sunlight passes through the atmosphere, clouds scatter and absorb it, reducing the amount of solar energy available to PV systems. Denser and thicker clouds block more sunlight, causing a greater reduction in solar energy output compared to thinner clouds. The ability of clouds to disrupt solar irradiance is influenced by their composition, including water droplets and ice crystals, which scatter sunlight in multiple directions. Overcast conditions or heavy cloud cover can drastically lower solar energy production, affecting the overall efficiency of solar systems. Accurate cloud cover data is therefore essential for solar forecasting models to predict these fluctuations and make more reliable energy estimates.

Integrating cloud cover data into solar forecasting models allows for precise adjustments to energy generation plans. By analyzing current cloud conditions, these models can predict how solar radiation will be affected and optimize energy production strategies accordingly. Understanding cloud patterns and their impact on solar irradiance helps energy operators adapt to changing weather conditions and maintain efficient energy generation. Accurate cloud data also aids in planning for energy storage and grid management, allowing for better allocation of resources.

B. SOLAR GEOMETRY

1) ANGLE OF INCIDENCE

The angle at which sunlight strikes solar panels, known as the angle of incidence, plays a crucial role in determining how much sunlight is absorbed and how efficiently it is converted into electricity. For optimal energy generation, the ideal angle of incidence is when the sunlight hits the panels directly or nearly directly. This ensures that the maximum amount of sunlight penetrates the solar cells, initiating the energy conversion process efficiently. When the angle deviates from the optimal position, such as during early mornings, late afternoons, or when the sun is lower in the sky, the amount of sunlight hitting the panels is reduced. This limits the energy conversion as less sunlight reaches the semiconductor material. To counteract this, solar systems often use tracking mechanisms or adjustable tilts that follow the sun's movement throughout the day, helping to maintain the ideal angle for maximum sunlight absorption and energy generation.

Including the angle of incidence in solar forecasting models allows for more accurate predictions of solar energy output. Since the sun's position changes throughout the day and across seasons, the angle of sunlight hitting the panels also shifts. By incorporating these variations into the models, solar energy generation can be predicted more precisely. This is particularly important for improving energy capture and efficiency by aligning panels to the sun's movement. Forecasting models that account for the angle of incidence can adapt to local factors such as latitude, time of year, and specific site conditions, which affect solar energy availability. Dynamic solar tracking systems, which adjust the panel orientation based on the sun's movement, are particularly useful for real-time energy optimization, increasing energy capture and conversion throughout the day.

To maximize the use of available solar resources, careful planning of solar panel placement and orientation is essential. By analyzing factors such as solar irradiance patterns, topography, and climate conditions, panels can be positioned in areas with high sunlight exposure and minimal obstructions. This strategic approach ensures continuous solar radiation and optimizes energy capture. Solar panel tilt and alignment are adjusted to match the sun's daily and seasonal path, ensuring that panels consistently capture the most sunlight possible. This improves the overall efficiency of the solar system, boosting energy production while minimizing energy losses. Optimizing both panel placement and orientation not only enhances performance but also increases the economic viability of solar plants by maximizing the return on investment through greater electricity generation. As a result, solar energy becomes more cost-effective and competitive with other energy sources.

2) ZENITH ANGLE

The zenith angle measures the angle between the sun and the point directly overhead for an observer, indicating the sun's position in the sky. This angle is critical for estimating solar irradiance and energy generation. When the sun is directly overhead, the zenith angle is 0 degrees, but it increases as the sun moves lower in the sky, reaching 90 degrees at sunrise and sunset. Solar forecasting models use the zenith angle to determine how much solar energy reaches the Earth's surface. The position of the sun affects how sunlight travels through the atmosphere, influencing how much of it is absorbed, scattered, or reflected. Since the zenith angle changes throughout the day and year, it plays a key role in accounting for daily and seasonal variations in sunlight availability. Incorporating this data into forecasting models helps improve the accuracy of predictions regarding solar energy production.

The zenith angle significantly affects how much solar radiation reaches the Earth's surface. When the zenith angle is small, sunlight travels a shorter distance through the atmosphere, encountering fewer particles that can absorb or scatter it. This results in stronger sunlight and higher energy capture by solar panels. As the zenith angle increases, sunlight must pass through more of the atmosphere, which leads to more scattering and absorption, reducing the amount of sunlight that reaches the surface and thus lowering solar energy output. This relationship between the zenith angle and sunlight intensity is crucial for optimizing solar installations. Forecasting models that include zenith angle data can help solar operators predict energy output more accurately, enabling better decisions for energy production, grid management, and resource allocation.

By incorporating the zenith angle into solar forecasting models, calculations of solar irradiance become more precise. This data helps predict how much sunlight will reach the Earth's surface at any given time and location, improving the overall accuracy of solar energy projections. With a better understanding of the sun's position and its effect on solar radiation, energy planners can optimize the positioning and orientation of solar panels to maximize efficiency.

3) AZIMUTH ANGLE

The azimuth angle measures the sun's horizontal position in relation to an observer or solar panel, making it a key factor in solar forecasting. This angle, calculated clockwise from 0° (north) to 360° as the sun moves across the sky, helps determine the sun's daily path and its position throughout the year. Understanding the azimuth angle is essential for assessing solar irradiance and optimizing solar energy generation. Solar forecasting models use the azimuth angle to predict changes in sunlight intensity and duration throughout the day and across seasons. This data is critical for positioning solar panels to capture the most sunlight. Solar panels are typically aligned with the equator, and adjusting their orientation based on the azimuth angle helps maximize sunlight exposure and energy conversion [34]. By using this information, solar operators can design systems and adjust panel angles for optimal year-round energy production.

The azimuth angle, along with the zenith angle (which measures the sun's vertical position), plays a crucial role in determining how much solar energy reaches the Earth's surface. Accurate measurement of these angles allows solar models to estimate solar radiation strength and how it is dispersed across a location. Proper alignment of solar panels, based on these angles, improves energy absorption and boosts efficiency. Incorporating the azimuth angle into solar forecasting models allows for more precise predictions of solar energy production. This data helps operators optimize the placement and orientation of solar panels, improving energy generation by aligning panels to the sun's horizontal movement. The models can also identify the best times to adjust panel angles to maintain optimal sunlight exposure throughout the day and across seasons.

C. WIND SPEED

1) DIRECT RELATIONSHIP

Renewable energy prediction methods treat wind and solar forecasting as distinct due to differing meteorological influences. Wind speed is critical for estimating wind turbine power generation, influencing turbine efficiency, blade dynamics, and power conversion. It is more relevant than temperature or solar irradiance in this context. Thus, wind energy forecasting models [35] primarily rely on wind speed to predict output. In contrast, solar forecasting depends on solar irradiance, cloud cover, and temperature, which directly affect photovoltaic conversion. Given these differences, specialized forecasting models are required for each energy source. Wind forecasting must account for the variable and dynamic nature of wind speed.

2) INDIRECT INFLUENCES

While wind speed is not a primary factor in solar forecasting, it may indirectly influence results. Changes in wind speed can alter cloud cover patterns, which in turn affect solar irradiance. Wind velocity also influences the scattering of dust particles, impacting atmospheric transparency and solar radiation levels. Nevertheless, solar forecasting models prioritize solar irradiance, temperature, and cloud cover, as these directly affect photovoltaic efficiency and energy output. Although wind speed may cause secondary atmospheric effects, it remains a minor component in solar forecasting models.

Recognizing the distinct meteorological drivers of solar and wind energy enables the development of tailored forecasting models, enhancing accuracy and relevance. While wind speed has limited direct influence on solar forecasts, understanding its indirect effects contributes to a more comprehensive view of atmospheric dynamics. In regions where wind patterns subtly affect solar conditions, refined models are particularly beneficial. Ultimately, solar forecasting should focus on the key meteorological parameters that define the solar energy environment.

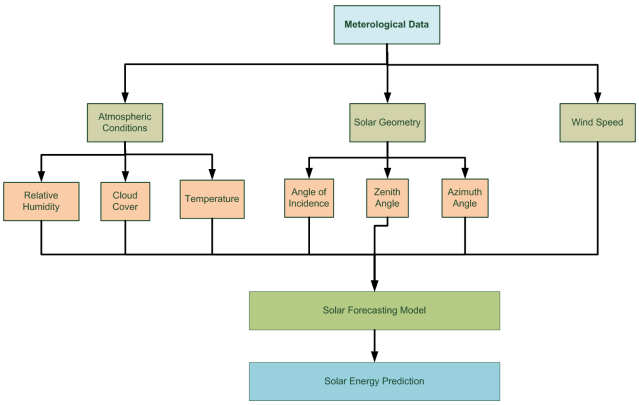

FIGURE 1. The flowchart for solar forecasting using feature engineering.

The solar forecasting using feature engineering described above has been depicted in the flowchart as shown in Fig. 1.

III. MACHINE LEARNING MODELS

Machine learning (ML) is an essential element in the field of artificial intelligence and computer science. The primary goal is to utilize data and advanced algorithms to imitate the cognitive learning processes seen in humans, with a gradual increase in accuracy over time. ML is a branch of data science that uses statistical approaches to train algorithms [36], [37], [38], [39], [40], [41], [42] specifically created for tasks like classifications or predictions.

ML has a wide range of applications in various business difficulties, including Regression, Classification, Forecasting, Clustering, and Associations, among other activities. Regression analysis is a statistical technique used to analyze the complex relationships between dependent and independent variables, typically incorporating numerous independent variables simultaneously. This analytical methodology offers useful insights into the manner in which the value of the dependent variable changes in relation to an independent variable, while keeping other predictors constant. The effectiveness of this method resides in its ability to accurately forecast continuous real-world values, encompassing several areas such as temperature, age, wage, and price.

Regression analysis is a crucial tool that reveals relationships between variables and enables the prediction of continuous output variables using one or more predictor variables. The main applications of this tool include making predictions, projecting future outcomes, modeling time series data, and identifying causal links between variables. This analytical approach is essential for extracting significant patterns from data, which contributes to well-informed decision-making processes across several domains.

A. MULTIPLE LINEAR REGRESSION

Multiple linear regression is a statistical technique used to analyze the connections between a dependent variable and two or more independent variables. This technique is extensively employed in diverse fields such as economics, biology, and social sciences to predict outcomes and analyze

complex relationships between factors. In multiple linear regression, the main idea is to identify a linear relationship between the independent and dependent variables. This model is based on the premise that the relationship between these variables can be accurately described using a linear equation (1). This modeling approach enables a detailed examination of the relationships between many components, offering a reliable analytical tool for forecasting and studying correlations in a wide range of academic and practical research settings.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \tag{1}$$

where, y is the dependent variable, β_1 to β_k represent the average influence on y when a one-unit increase is observed in x while keeping other predictors constant and ϵ represents the error term.

This method finds its utility in scenarios where a comprehensive understanding is sought, such as evaluating the impact of factors like rainfall, temperature, and fertilizer quantity on crop growth.

B. RIDGE REGRESSION

Ridge regression is a statistical method used in linear regression research to tackle issues associated with multicollinearity and overfitting. This methodology incorporates regularization, which improves the conventional linear regression cost function by including a penalty factor. The penalty term's magnitude, represented by the hyperparameter λ, is crucial in the regularization process.

Within the framework of basic linear regression, the goal is to determine coefficients that minimize the total sum of squared differences between the observed values and the predicted values. Nevertheless, when dealing with independent variables that are strongly linked, the typical linear regression model might display instability and produce coefficient estimates that are not trustworthy. Ridge regression resolves this problem by incorporating a penalty component into the least squares cost function, so alleviating problems related to multicollinearity.

Ridge regression is a useful technique in the field of ordinary multiple linear regression. It helps address the issues of multicollinearity and overfitting that arise when a collection of p predictor variables work together with a response variable to create a model (2). By using a regularization term, the estimation of coefficients becomes more stable and trustworthy, enhancing the model's robustness while dealing with correlated predictors. The hyperparameter λ controls the level of regularization, providing a means to balance the complexity of the model and its forecast accuracy.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{2}$$

where, Y is the response variable or the dependent variable that the model aims to predict, β_0 is the intercept term, representing the expected mean value of Y when all X_j are zero, $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients representing the average effect on Y for a one-unit increase in X_j while holding other predictors constant, $X_1, X_2, \ldots, X_p$ are the predictor variables or independent variables used to predict Y and ϵ is the error term, capturing the variability in Y not explained by the predictor variables.

To compute the values of β_1 to β_p the least squares method comes into play, aiming to minimize the summation of squared residuals (RSS) (3).

$$RSS = \sum (y_i - \overline{y}_i)^2 \tag{3}$$

where y_i is the actual value for the i^{th} observation and $\overline{y}_i$ is the predicted value.

However, the issue becomes complex when predictor variables have a strong association, resulting in the occurrence of multicollinearity. This process might cause the coefficient estimates of the model to become unstable, resulting in a significant level of unpredictability. In order to tackle this difficulty without completely disregarding certain predictor variables, an alternate method called ridge regression is employed. Ridge regression aims to minimize the resulting expression given by (4).

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4}$$

where λ assumes a non-negative value. The additional element in the equation is referred to as a "shrinkage penalty." When the value of λ is 0, the penalty component has no influence, resulting in ridge regression producing coefficient values similar to those obtained by the least squares approach. As the value of λ increases towards infinity, the impact of the shrinkage penalty becomes more significant, causing the coefficient estimates of ridge regression to eventually approach zero. In practical scenarios, predictor variables with less influence on the model tend to rapidly approach zero due to the inherent shrinkage effect.

C. LASSO REGRESSION

Lasso regression, short for "Least Absolute Shrinkage and Selection Operator," is a linear regression method that utilizes L_1 regularization to tackle multicollinearity issues and facilitate feature selection. It shares similarities with ridge regression but employs a distinct penalty term that promotes sparse coefficient estimates. Nevertheless, there are instances in which predictor variables have a significant association, resulting in the occurrence of multicollinearity. This occurrence can lead to incorrect coefficient estimates for the model, with a greater likelihood of excessive variance. Essentially, when the model is used on new, unfamiliar data, its performance is expected to be below average. An effective strategy to tackle this difficulty is to utilize LASSO regression, which aims to minimize the expression (5).

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

where, RSS (Residual Sum of Squares) $= \sum(y_i - \hat{y}_i)^2$, where y_i is the actual value for the i^{th} observation and $\hat{y}_i$ is the predicted value. It measures the discrepancy between the data and the model's predictions, λ is the regularization parameter that controls the strength of the penalty. Higher values of λ increase the amount of shrinkage, leading to simpler models, β_j is the coefficients representing the influence of the j^{th} predictor variable on Y.

In Lasso regression, the L1 penalty ($|\beta_j|$) promotes sparsity, potentially driving some coefficients to zero and effectively selecting a subset of predictors. However, as λ progressively increases towards infinity, the influence of the shrinkage penalty intensifies. Consequently, predictor variables deemed less significant in the model experience a substantial shrinkage effect, moving them towards zero. In certain cases, some predictor variables might even be eliminated from the model entirely

D. DECISION TREE REGRESSION

The phrase ''decision tree'' accurately represents its operational philosophy that relies on conditions. This method is highly efficient and utilizes powerful algorithms for predictive analysis. The structure consists of essential elements: internal nodes, branching, and terminal nodes. In the context of a decision tree, branches represent the different outcomes of tests, while each leaf node represents a specific class label. The algorithm's versatility is apparent as it serves both classification and regression tasks, which are two fundamental types of supervised learning algorithms. Nevertheless, decision trees are highly sensitive to their training data, meaning that even slight modifications to the training set can result in substantial changes to the tree structures that are produced. At each internal node, the algorithm chooses both the characteristic and the threshold that efficiently separate the data into smaller groups. The decision regarding the splitting process is determined by a criterion that aims to minimize the variation of the goal values within each subgroup, where the criterion is to minimize the mean squared error (MSE), which can be defined as given by (7) for a given node N.

Upon the creation of a leaf node, a constant value is set to reflect the prediction for all data points that fall into that leaf. In regression, this number is commonly calculated as the mean of the target values within the leaf. The construction of a decision tree entails a recursive process. The method starts at the root node and selects the most favorable characteristic and threshold to partition the data. Subsequently, it advances to the offspring nodes, repetitively going through this procedure until a certain termination criterion is met. When decision trees become overly deep, they tend to overfit the training data by catching irrelevant noise instead of the fundamental patterns. Pruning is a technique used to address this problem by removing or consolidating nodes that do not significantly improve the model's performance when evaluated on validation data.

E. SUPPORT VECTOR REGRESSION

Support Vector Regression (SVR) [43] is based on the ideas of Support Vector Machines (SVM) but with subtle differences. Unlike classification SVM, which seeks to identify a linear or non-linear decision border, SVR is designed to determine a curve that accurately represents the relationships between data points in regression tasks. Contrary to using the curve as a decision boundary, Support Vector Regression (SVR) utilizes the curve to assess the degree of alignment between the curve and the data points' positions. Support vectors, which are essential for support vector regression (SVR), assist in determining the nearest correspondence between the curve and the data points.

IV. ENSEMBLE LEARNING MODELS

A. RANDOM FOREST REGRESSION

Random Forest Regression is a robust ensemble learning method employed for both classification and regression tasks. Ensemble learning [44], [45], [46] is a widely used technique in ML that leverages the collective strength of several decision trees to enhance prediction accuracy. The Random Forest algorithm is constructed based on the concept of decision trees. A decision tree is a diagrammatic representation in which each internal node signifies a characteristic, the branches signify a rule for making a decision, and each leaf node signifies the final result. Decision trees possess a straightforward and comprehensible nature, although they are prone to overfitting the data.

Random Forest employs ensemble learning, a technique that amalgamates the forecasts of numerous independent models (decision trees) to get a more resilient and precise prediction. The term ''Random'' in Random Forest pertains to two primary factors of randomness: Bootstrapping and Feature Randomness. During the construction of each decision tree in the forest, a random subset of the dataset is selected with the possibility of selecting the same data point multiple times. This process is commonly referred to as bootstrapping. It guarantees that every tree is constructed using a marginally distinct dataset, hence introducing variability into the model. During each split in a decision tree, only a randomly selected subset of features is taken into account for making the split. This aspect of unpredictability serves to mitigate excessive correlation among individual trees. In order to get a prediction using a Random Forest Regression model, the predictions from each individual tree are either averaged or subjected to a voting process. The collective impact of this ensemble phenomenon frequently results in a forecast that is both more precise and resilient when contrasted with that of an individual decision tree.

B. BAGGING REGRESSOR

The Bagging Regressor is an ensemble meta-estimator that works by training base regressors on randomly selected sections of the original dataset. The projections of these distinct models are subsequently aggregated, either through

average or voting, to get the ultimate prediction. This strategy efficiently reduces the variability of a main estimator by introducing randomness during its building and then creating a group of estimators.

This ensemble learning approach, known as the ''bagged regressor'' or ''bootstrap aggregating regressor,'' improves the accuracy of predictions in regression problems. This is a modification of the bagging concept specifically designed for regression circumstances. The core idea revolves around training several regression models on separate subsets of training data, allowing them to catch different patterns and reduce the impact of outliers. This procedure yields a resilient ensemble that enhances the precision of forecasts.

C. ADA BOOST REGRESSOR

ADA Boost, sometimes known as ''Adaptive Boosting,'' was an early boosting approach that achieved widespread acclaim. This technique entails modifying the weights assigned to training samples in a dynamic manner, as opposed to depending on a set learning rate. The reason it is called ''adaptive'' is because it utilizes a dynamic technique that adjusts the weights. It results in the development of a boosting regressor that consistently surpasses the performance of a basic estimator. Adaboost utilizes an ensemble of weak learners, which are fundamental learning models, to create a robust regressor. To develop the Adaboost.R2 algorithm, we start by defining the weak learner, loss function, and the available dataset. N represents the total samples, and the ensemble comprises M weak learners, indexed as n = 1, ..., N and m = 1, ..., M. The training process involves sequentially training weak learners (f_m) on data (X_m, y_m), sampled from (X,y) with replacement. Sample weights w are updated to place emphasis on previous mistakes. A model confidence measure β_m is assigned to the mth weak learner to blend it with the ensemble. This process is illustrated in Algorithm 1.

Algorithm 1 ADA Boost Regression Training

Require: Dataset (X, y) with N samples, number of iterations M

Initialize sample weights: $w_{n1} = 1$ for $n = 1, \ldots, N$
for $m = 1$ to M **do**
  Compute sample probabilities: $p_n = \frac{w_{nm}}{\sum_n w_{nm}}$ for all n
  Draw N samples (X_m, y_m) from (X, y) using p_n
  Fit weak learner f_m on (X_m, y_m)
  Compute loss l_n for each sample using predictions from f_m
  Compute average loss $l^- = \frac{1}{N} \sum_n l_n$
  if $l^- \geq 0.5$ **then**
    Terminate the boosting process
  end if
  Compute confidence: $\beta_m = \frac{l^-}{1-l^-}$
  Update weights: $w_{n(m+1)} = w_{nm} \cdot \beta_m \cdot (1 - l_n)$
end for

Predictions for a given input x^* are generated by considering predictions from each weak learner and calculating an ensemble prediction based on the confidence measures. The smallest K^{th} machine's prediction that meets a certain condition becomes the ensemble prediction, noted as y_k. This value represents a weighted median of the predicted values y.

D. GRADIENT BOOSTING REGRESSOR

Gradient Boosting Regressor is a potent ensemble learning approach within the realm of boosting algorithms, widely applied to regression tasks in the field of ML. Its fundamental objective is to construct a robust predictive model by sequentially amalgamating the predictions of multiple weaker models, typically simple decision trees. The process of Gradient Boosting Regressor:

1. *Weak Learners* The process commences with a basic learner, often a shallow decision tree. This initial tree is trained on the dataset to make predictions for the target variable.
2. *Residuals* The model's predictions are compared against the actual target values, and the disparities, known as residuals, are computed. These residuals represent the errors of the initial model.
3. *Weighted Focus on Errors* Subsequently, a new decision tree is trained to forecast these residuals. The learning algorithm assigns greater importance to the data points that the previous model inaccurately predicted.
4. *Iterative Process* Steps 2 and 3 are repeated iteratively. Each new tree aims to rectify the errors made by the collective ensemble of models from prior iterations.
5. *Combining Predictions* The ultimate prediction is derived by consolidating the predictions of all the decision trees. The ensemble model formed through this iterative process tends to yield more precise predictions than individual trees.

Gradient Boosting Regressor provides several advantages, such as its capacity to handle intricate data relationships, resilience to overfitting when hyperparameters are meticulously adjusted, and a high level of accuracy. To achieve optimal results with Gradient Boosting Regressor, it's essential to fine-tune hyperparameters like the learning rate, the number of trees in the ensemble, and the maximum depth of each tree.

E. PERFORMANCE METRICS

Performance metrics hold a crucial role in evaluating the precision and effectiveness of regression models designed for the prediction of continuous target variables.

Root Mean Square Error (RMSE) RMSE The Root Mean Square Error (RMSE) is a commonly used metric to measure the accuracy of a regression model. It is calculated using (6).

$$RMSE = \sqrt{\frac{\sum(actual - predicted)^2}{n}} \tag{6}$$

where, ''actual'' refers to the actual (observed) values of the target variable, ''predicted'' stands for the predicted values generated by the regression model and ''n'' represents the number of data points or observations in your dataset.

Mean Squared Error (MSE) MSE shares similarities with RMSE but doesn't involve the square root. It measures the average squared difference between predicted and actual values (7).

$$MSE = \frac{1}{n}\sum(actual - predicted)^2 \tag{7}$$

Mean Absolute Error (MAE) MAE calculates the average absolute difference between predicted and actual values. It is represented by (8).

$$MAE = \frac{1}{n}|y_i - \hat{y}_i| \tag{8}$$

where MAE represents Mean Absolute Error, n is the number of instances, y_i signifies actual values, and $\hat{y}_i$ denotes predicted values.

R-Squared (R^2) The R-squared, often denoted as R^2 (pronounced ''R-squared''), is a statistical measure used to evaluate the goodness of fit of a regression model. It provides insight into how well the independent variables (features) in the model explain the variance in the dependent variable (target). In simple terms, R^2 quantifies the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. The value of R^2 typically ranges from 0 to 1. An R^2 of 0 indicates that the model doesn't explain any of the variance in the dependent variable, meaning it's a poor fit. An R^2 of 1 indicates that the model perfectly explains all the variance in the dependent variable, implying an ideal fit. In practical terms, an R^2 value closer to 1 indicates that a larger portion of the variance in the dependent variable is explained by the model, suggesting a better fit. However, it's important to consider context and domain-specific knowledge when interpreting R^2, as very high R^2 values might not necessarily mean the model is always a good predictor in all situations. Additionally, R^2 can't determine if the model's coefficients are statistically significant or if the model is free from issues like overfitting.

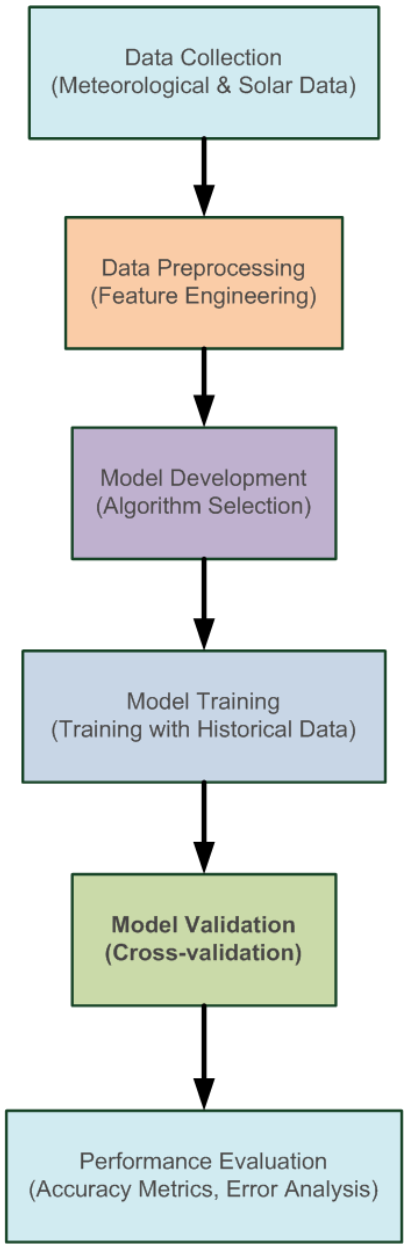

FIGURE 2. Flowcahart representation of proposed method.

V. PROPOSED APPROACH

A. DATASET DESCRIPTION

The dataset used in this study was obtained from an open-source repository on Kaggle titled *Solar Energy Power Generation Dataset* [47]. It contains real-world data on solar photovoltaic (PV) power output and associated meteorological variables such as temperature, humidity, pressure, solar irradiance, and geometric features like azimuth and zenith angles. The data was collected and compiled by the original dataset authors using standard meteorological sensors and logging systems. No field data collection or physical sensing devices were deployed by the authors of this study. Instead, the dataset was accessed as-is for the purpose of model training, evaluation, and simulation. Feature engineering involved temporal alignment, geometric calculations based on location metadata, and handling of missing or noisy entries using imputation and filtering techniques. The proposed method is represented pictorially in the flowchart shown in Fig. 2.

Numerical Weather Prediction Data The weather data was collected from [47]. In Table 2 the extracted data parameters and a short explanation with the corresponding units is given. The simulation parameters used are given in Table 3.

B. MODEL TRAINING AND HYPERPARAMETER TUNING

The models chosen for comparison - SVR, RF, and GBR - represent classical kernel-based, ensemble-based, and

TABLE 2. Numerical weather data.

Variable Name	NWP	Unit
Temperature 2m above gnd	Temperature at 2m above ground	$^{\circ}C$
Relative humidity 2m above gnd	Relative Humidity at above ground	%
Mean sea level pressure MSL	Mean Sea Level Pressure	%
Total precipitation sfc	Total Precipitation	%
Snowfall amount sfc	Snowfall Amount	%
Totalcloud cover sfc	Total Cloud Cover	%
Highcloud cover high cld lay	High Cloud Cover	%
Medium cloud cover mid cld lay	Medium Cloud Cover	%
Low cloud cover low cld lay	Low Cloud Cover	%
Shortwave radiation backwards sfc	Short Wave Radiation Backwards	$\frac{W}{m^2}$
Wind speed 10m above gnd	Wind Speed At 10m Above Ground	$\frac{m}{s}$
Wind direction 10m above gnd	Wind Direction At 10m Above Ground	$\frac{m}{s}$
Wind speed 80m above gnd	Wind Speed 80m At Above Ground	$\frac{m}{s}$
Wind direction 80m above gnd	Wind Direction At 80m Above Ground	$\frac{m}{s}$
Wind speed 900mb	Wind Speed	$\frac{m}{s}$
Wind direction 900mb	Wind Direction	$\frac{m}{s}$
Wind gust 10m above gnd	Wind Gust At 10m Above Ground	$\frac{m}{s}$
Angle of incidence	Angle Of Incidence	degree
Zenith angle	Zenith	degree
azimuth angle	Azimuth	degree
Generated power	Generated power	kW

TABLE 3. Simulation parameters.

Parameter	Value
Learning Rate	0.1
Number of Estimators	100
Maximum Depth	3
R-squared (Gradient Boosting)	0.83
MSE (Gradient Boosting)	158559.33
RMSE (Gradient Boosting)	399.44
MAE (Gradient Boosting)	253.62

boosting-based families, respectively. While SVR offers a baseline for kernel regression, RF and GBR represent robust ensemble models known for their performance in energy forecasting. The primary aim is to benchmark GBR against state-of-the-art ensemble methods.

To ensure robust and fair comparison among machine learning models, hyperparameter tuning was conducted using a grid search with 5-fold cross-validation. The Gradient Boosting Regressor (GBR) was selected for detailed tuning due to its superior baseline performance. The hyperparameters optimized include the number of estimators, learning rate, maximum tree depth, and subsampling ratio. The search was performed over 16 parameter combinations as outlined in Table 4, totaling 80 model evaluations. RMSE was used as the optimization metric.

The optimal configuration identified was: `learning_rate = 0.05`, `max_depth = 5`, `n_estimators = 150`, and `subsample = 0.8`. This configuration yielded

TABLE 4. Grid search parameter ranges for gradient boosting regressor.

Hyperparameter	Values Tested
Learning Rate	0.05, 0.10
Max Depth	3, 5
Estimators	100, 150
Subsample	0.8, 1.0

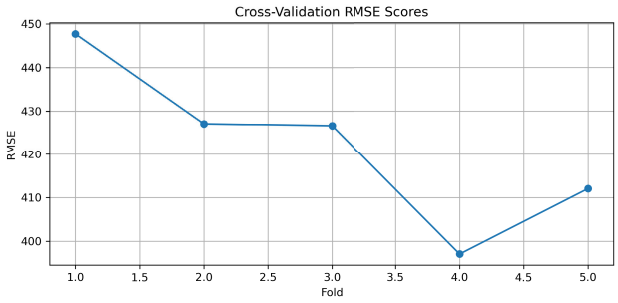

FIGURE 3. Cross-validation RMSE scores across 5 folds for the best-tuned GBR model.

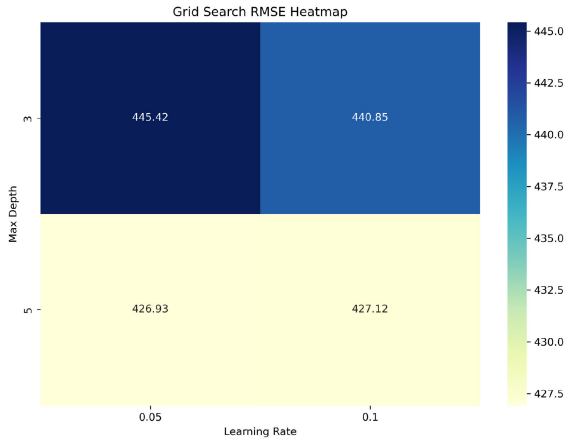

FIGURE 4. Grid search RMSE heatmap showing the effect of learning rate and max depth.

a test RMSE of 406.47, MAE of 261.25, and R^2 score of 0.819. In contrast, the default GBR model yielded a higher RMSE of 425.20, indicating that tuning improved prediction accuracy by approximately 4.4%.

Cross-validation RMSE scores across the five folds were consistent, with a mean of 422.12 and a standard deviation of 16.92 (see Fig. 3). This reflects a stable and generalizable model fit. The grid search surface shown in Fig. 4 further highlights the interaction between learning rate and tree depth on model performance.

Permutation feature importance analysis (Fig. 5) revealed that solar geometric parameters-particularly *angle of incidence*, *azimuth*, and *zenith*-had the highest predictive value, confirming the significance of incorporating these features in solar forecasting models.

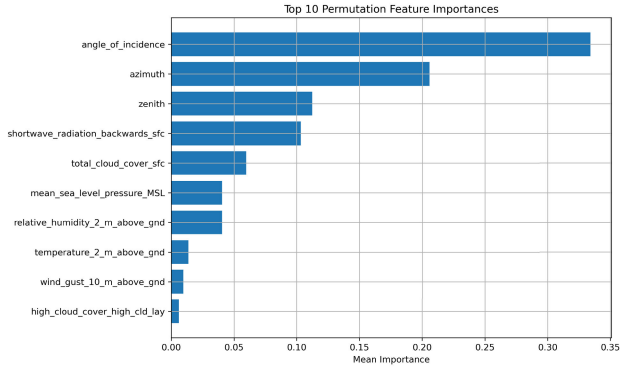

FIGURE 5. Top 10 permutation feature importances. Solar geometric features dominate.

TABLE 5. RMSE comparison across 5 folds for GBR, RF, and SVR.

Fold	GBR	RF	SVR
Fold 1	406.23	410.57	691.22
Fold 2	418.12	422.35	687.45
Fold 3	430.17	438.11	699.35
Fold 4	404.89	417.04	710.12
Fold 5	422.98	427.69	703.78
Mean	**416.08**	**423.55**	**698.38**

C. STATISTICAL SIGNIFICANCE

To evaluate whether the differences in forecasting accuracy among machine learning models were statistically significant, paired t-tests were conducted using RMSE values obtained from 5-fold cross-validation. The Gradient Boosting Regressor (GBR) was compared against Random Forest (RF) and Support Vector Regression (SVR), which represent strong classical and kernel-based approaches, respectively.

The results indicate that the GBR significantly outperformed SVR, with a t-statistic of -47.36 and a p-value of 1.19×10^{-6} ($p < 0.01$), thereby rejecting the null hypothesis of equal means. In contrast, the difference between GBR and RF was not statistically significant ($t = -0.89$, $p = 0.42$), suggesting similar performance levels between the two ensemble methods. These findings validate that the observed improvements using GBR are meaningful, especially in contrast to SVR, which lagged across all validation folds.

To provide further transparency, the RMSE values for each model across the five folds are summarized in Table 5. It can be observed that GBR and RF maintain relatively consistent RMSE values, whereas SVR exhibits consistently higher error rates.

Additionally, Fig. 6 presents a visual comparison of RMSE across the folds. The compact performance distribution of GBR highlights its robustness, while SVR's variability and elevated error levels confirm its unsuitability for capturing the nonlinear relationships in solar energy forecasting under the given conditions.

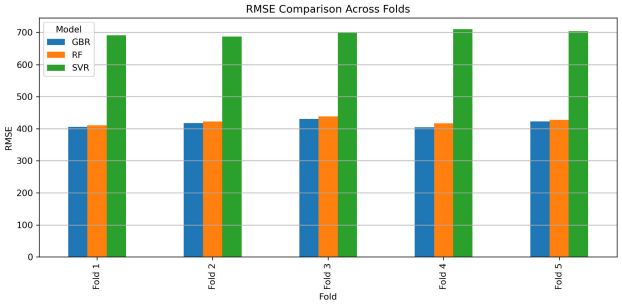

FIGURE 6. Bar plot of RMSE across 5 folds for each model (GBR, RF, SVR).

D. MODEL ROBUSTNESS AND GENERALIZATION STRATEGY

Overfitting is a critical concern in machine learning-based forecasting, particularly when dealing with high-dimensional meteorological datasets and temporal correlations. To mitigate overfitting, this study employed a 5-fold cross-validation framework during hyperparameter tuning, ensuring that models were evaluated across different training and validation splits. Furthermore, regularization techniques such as subsampling (for GBR) and ensemble averaging (for RF and Bagging) were implicitly used to control model complexity.

The consistent performance of the tuned Gradient Boosting Regressor across folds, with an RMSE standard deviation of only 16.92, indicates strong generalization and robustness. Additionally, the permutation feature importance analysis demonstrated that the model does not rely on a small subset of features, but instead leverages a diverse set of atmospheric and geometric inputs, improving its ability to adapt to varied environmental conditions.

Although the current validation was performed using random sampling, future work will involve testing generalization using time-based and location-specific data splits. This will better simulate deployment scenarios where the model must forecast solar generation under entirely new weather profiles or from different geographic sites. Incorporating such validation will further strengthen the model's real-world applicability and resilience against data distribution shifts.

Overall, the implemented methodology balances accuracy and generalization, and the results suggest the model is well-suited for integration into smart grid operations that require reliable, high-frequency solar power forecasts.

VI. RESULTS AND DISCUSSION

This section presents the empirical findings obtained from several modeling methodologies, along with detailed charts illustrating various performance measures.

A. ML MODELS

The performance metrics of R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean

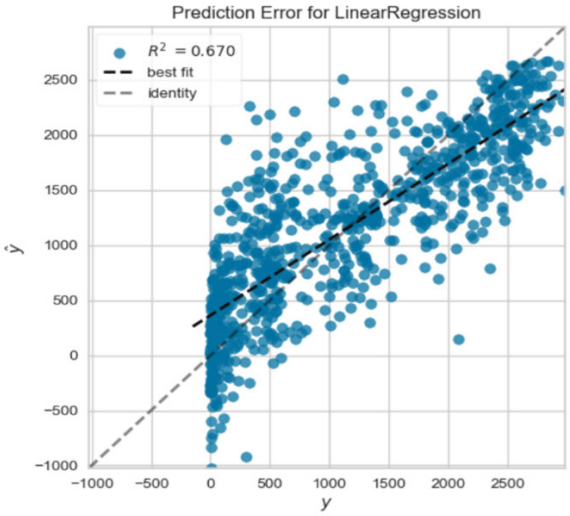

FIGURE 7. Prediction error (R-Square) for the multiple linear regression.

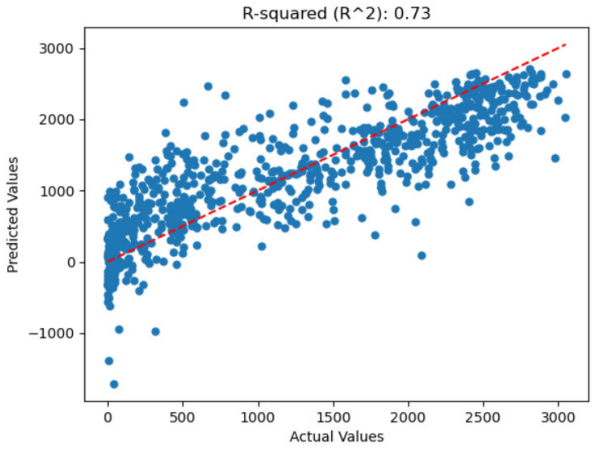

FIGURE 8. Prediction error (R-Square) for the ridge regression.

Absolute Error (MAE) for the Multiple Linear Regression with various hyperparameter combinations are presented in this section. In Fig. 7, the performance of the Multiple Linear Regression (MLR) model indicates limited predictive capacity, with a maximum R^2 of 0.67, suggesting the model struggles to capture the nonlinear dependencies inherent in solar power generation. The relatively higher MSE and RMSE further confirm its inadequacy in complex scenarios, where meteorological inputs exhibit high variance.

As shown in Fig. 8, Ridge Regression improved upon MLR, attaining an R^2 of 0.73. This can be attributed to its L2 regularization, which effectively reduces overfitting in multicollinear settings. Nonetheless, the model's linear structure restricts its adaptability to the nonlinear influence of solar geometry features.

Figure 9 shows that Lasso Regression performs similarly to Ridge in terms of RMSE and R^2 (0.73). However, its feature selection capability may introduce instability by omitting weak but collectively influential predictors, which may

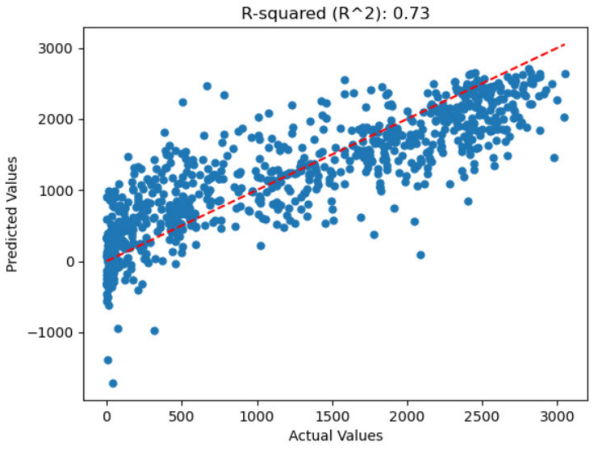

FIGURE 9. Prediction error (R-Square) for the LASSO regression.

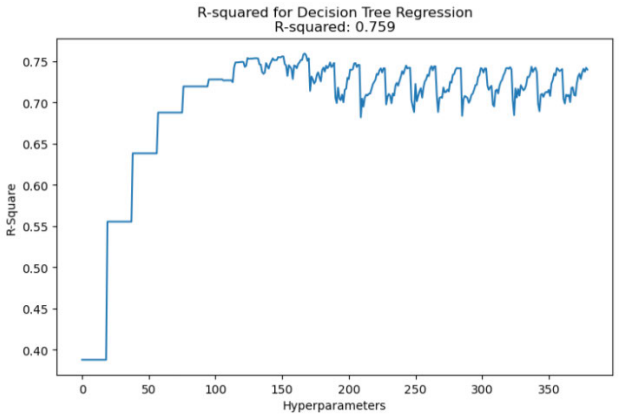

FIGURE 10. Prediction error (R-Square) for the decision tree regression.

explain its slightly inconsistent behavior across validation folds.

Decision Tree Regression in Fig. 10 displays a more substantial fit ($R^2 \approx 0.76$) with improved error metrics. This indicates that DTR can model non-linear splits efficiently, though its tendency to overfit is evident from the sharp drop in performance for certain hyperparameter combinations.

In Fig. 11, the Support Vector Regression (SVR) model achieved the lowest R^2 (≈ 0.39). The poor performance can be linked to the kernel function's sensitivity to scale and parameter tuning, and its inability to generalize well with high-dimensional or noisy meteorological inputs.

B. COMPARISON OF ML MODELS

The ML models utilized in this work entailed generating uncomplicated predictions by assuming gradual variations or persistence in daily solar PV power generation. Due to these oversimplified assumptions, it was not expected that these models would achieve a high level of accuracy.

Figure 12 clearly identifies Decision Tree Regression (DTR) as the top-performing standalone model, surpassing all others in R^2. This indicates its strength in capturing threshold-based variability-useful when solar output is

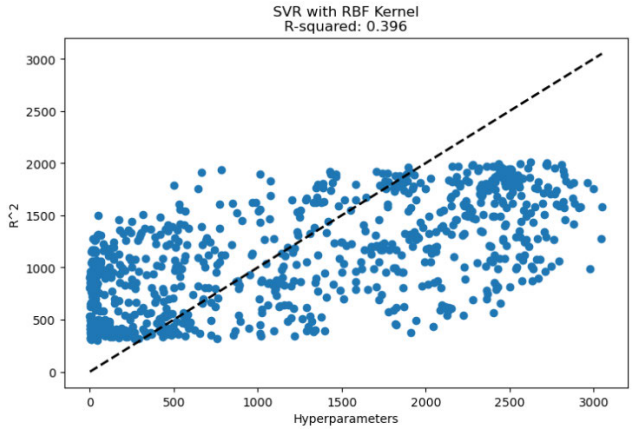

FIGURE 11. Prediction error (R-Square) for the support vector regression.

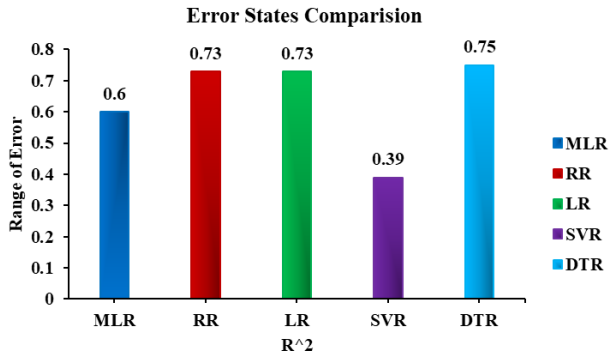

FIGURE 12. Comparison of the R-squared for different ML algorithms.

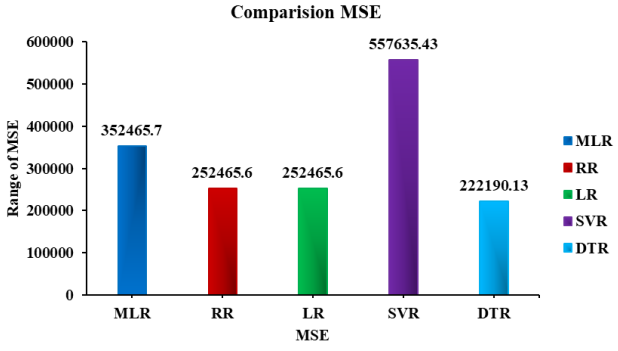

FIGURE 13. Comparison of the MSE for different ML algorithms.

abruptly affected by cloud formations. However, ensemble methods (analyzed below) further improve generalizability.

Figure 13 shows a descending trend in MSE from SVR to DTR, which aligns with the models' increasing flexibility. DTR's lowest MSE confirms its closer alignment with actual observations under variable weather patterns.

In Fig. 14, RMSE trends again favor DTR, emphasizing its strong fit across a wide range of error magnitudes. SVR's high RMSE reinforces its inability to manage nonlinear, fluctuating relationships effectively.

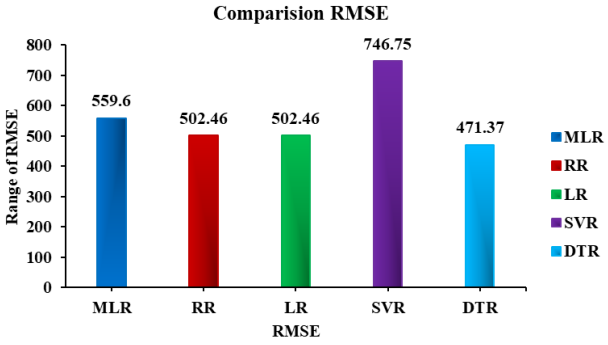

FIGURE 14. Comparison of the RMSE for different ML algorithms.

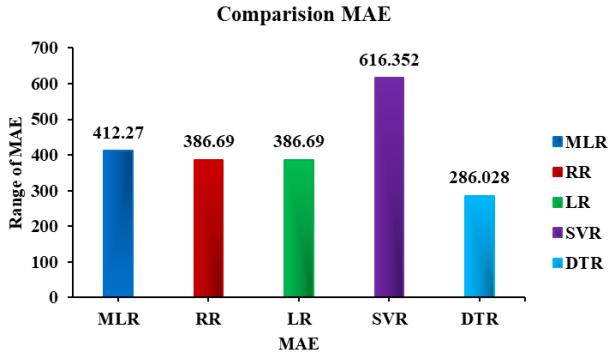

FIGURE 15. Comparison of the MAE for different ML algorithms.

Figure 15 reaffirms DTR's advantage with the lowest MAE, indicating its better handling of absolute deviations-a critical metric when consistent forecast accuracy is needed for energy dispatch decisions.

In this hypothetical case, it is evident that Decision Tree Regression (DTR) has the lowest values for RMSE, MSE, and MAE, suggesting superior prediction accuracy in comparison to other algorithms.

- MLR, Ridge, and Lasso exhibit comparable performance in terms of RMSE, MSE, and MAE, with Ridge demonstrating a slightly superior performance compared to Lasso.
- Support Vector Regression (SVR) exhibits higher Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE) values, suggesting that it may not be efficiently capturing the underlying patterns compared to linear-based approaches or SVR.
- Decision Tree Regression exhibits the highest R-squared value, indicating a superior alignment with the data in comparison to other techniques.

The selection of a regression algorithm is contingent upon the particular trade-offs you wish to achieve in terms of accuracy, interpretability, and complexity. The Decision Tree Regression (DTR) model demonstrates strong performance in this hypothetical case across all criteria, suggesting a harmonious combination of accurate prediction and model fit. Nevertheless, it is crucial to verify these discoveries

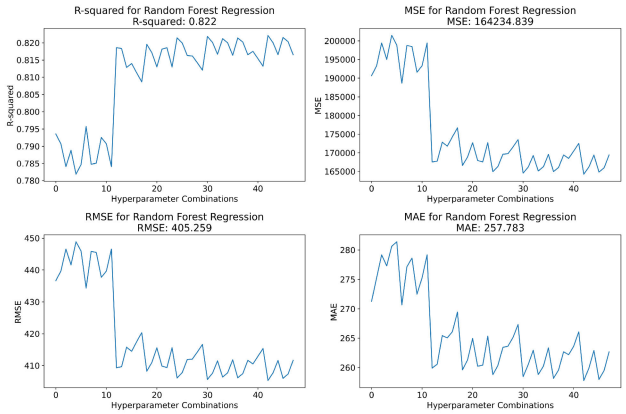

FIGURE 16. Performance evaluation for the random forest regression - R^2, MSE, RMSE, MAE.

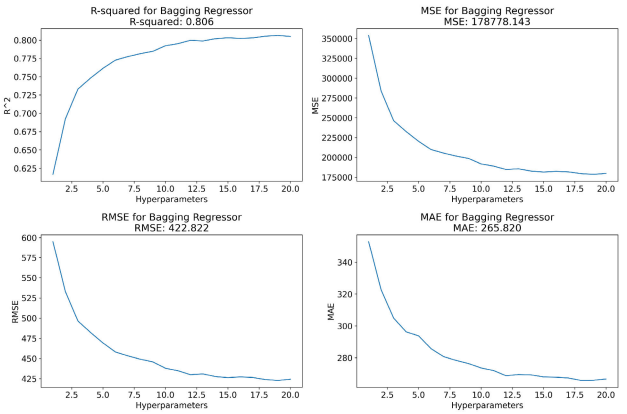

FIGURE 17. Performance evaluation for the bagging regressor - R^2, MSE, RMSE, MAE.

using real-world data and take into account additional criteria such as model interpretability and computational complexity before to reaching a definitive conclusion.

C. ENSEMBLE LEARNING MODELS

Ensemble approaches in ML consistently surpass the performance of individual ML models, demonstrating their exceptional capacity to improve predictive capabilities. These strategies utilize the combined intelligence of various models, leading to a substantial increase in accuracy and a more profound comprehension of complex data linkages. Ensemble approaches harness synergy to efficiently negotiate intricate patterns that standalone models may struggle to identify.

In Fig. 16, Random Forest Regressor (RFR) consistently performs well across all metrics. The ensemble strategy of bootstrapped decision trees improves prediction robustness while managing overfitting, evident from its R^2 of 0.822 and relatively low RMSE and MAE.

Figure 17 shows Bagging Regressor trailing slightly behind RFR, with an R^2 of 0.806. While its variance reduction is effective, the lack of feature randomness (unlike RF) likely limited its performance gains.

The AdaBoost Regressor in Fig. 18 shows weaker results (R^2 of 0.692), which can be attributed to its sensitivity to outliers and tendency to overemphasize hard-to-predict samples, leading to increased prediction variance.

Gradient Boosting Regressor (GBR), shown in Fig. 19, demonstrates the best overall performance with an R^2 of 0.827 and the lowest RMSE (399.44). This suggests its strength in stage-wise optimization and ability to capture interactions among meteorological and geometric features, making it highly suitable for solar energy forecasting.

D. COMPARISON OF ENSEMBLE LEARNING MODELS

Ensemble approaches are a powerful method for improving the performance and reliability of predictive modeling. They

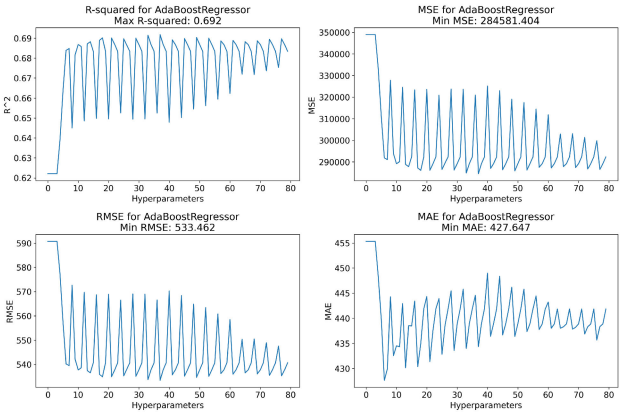

FIGURE 18. Performance evaluation for the ADA boost regressor - R^2, MSE, RMSE, MAE.

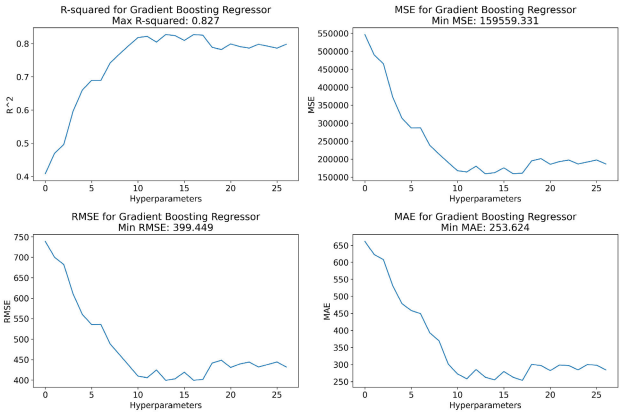

FIGURE 19. Performance evaluation for the gradient boosting regressor - R^2, MSE, RMSE, MAE.

demonstrate their supremacy in the field of ML. Fig.16 to Fig.19 illustrate the values of Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) used for evaluating the model.

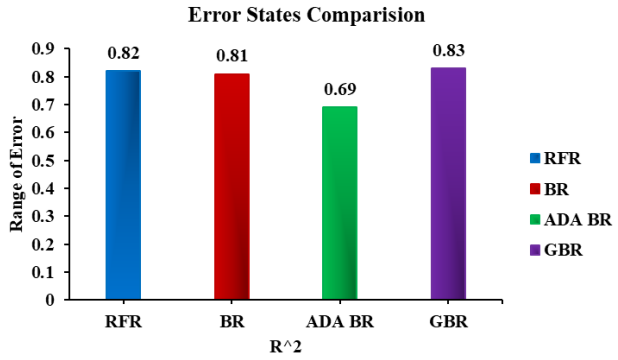

FIGURE 20. Comparison of the R-squared for different ensemble learning models.

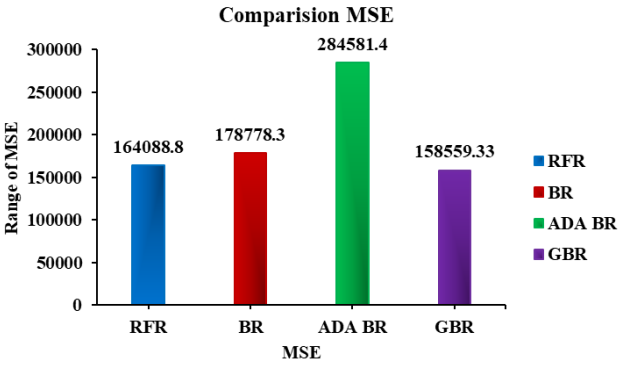

FIGURE 21. Comparison of the MSE for different ensemble learning models.

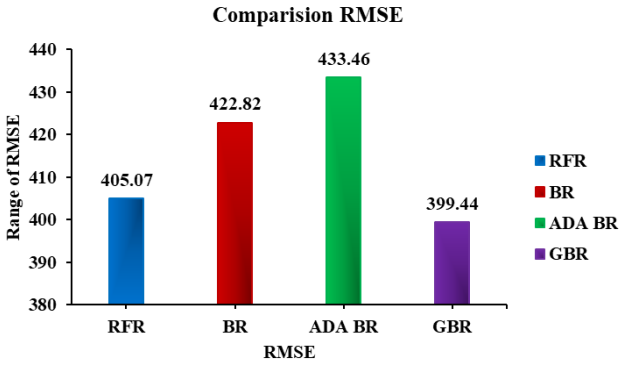

FIGURE 22. Comparison of the RMSE for different ensemble learning models.

As shown in Fig. 20, the GBR clearly outperforms other ensemble methods in terms of R^2, indicating its superior ability to explain variance in solar output across diverse weather scenarios.

Fig. 21 and Fig. 22 show that GBR also achieves the lowest MSE and RMSE respectively, confirming its generalization strength and ability to avoid large prediction errors-critical for grid stability applications.

In Fig. 23, the MAE values again place GBR ahead, reinforcing its reliability in minimizing average deviations

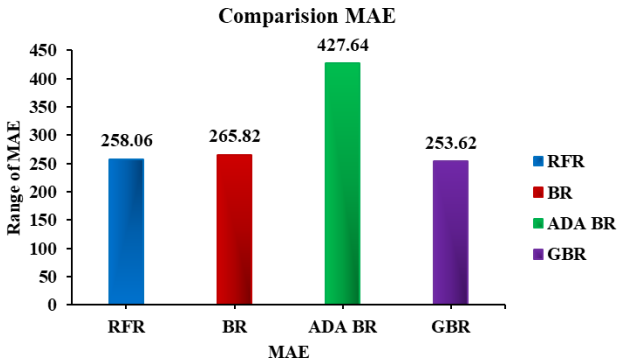

FIGURE 23. Comparison of the MAE for different ensemble learning models.

from actual output. The ADA Boost model, in contrast, performs poorly across all figures, reflecting its limited resilience to noisy, real-world solar datasets.

Within this hypothetical circumstance, it is evident that:

- The Gradient Boosting Regressor exhibits the lowest values for RMSE, MSE, and MAE, suggesting superior prediction accuracy in comparison to alternative methods.
- The ADA Boost Regressor exhibits higher values for RMSE, MSE, and MAE, suggesting that it may not be properly capturing the underlying patterns compared to the linear-based approaches or SVR.
- The Gradient Boosting Regressor exhibits the best R-squared value, indicating a superior fit to the data in comparison to other methods.

The selection of a regression algorithm is contingent upon the particular trade-offs you wish to achieve in terms of accuracy, interpretability, and complexity. The Decision Tree Regression (DTR) model demonstrates strong performance in this hypothetical scenario, as evidenced by its favorable results across all metrics. This suggests a harmonious combination of accurate prediction and model fit. Nevertheless, it is crucial to verify these discoveries using actual data and take into account additional criteria like model interpretability and computational complexity before to reaching a definitive conclusion.

The Table 6 compares the performance metrics of several ML and ensemble models. The optimal approach is determined based on the Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) for evaluating the models. Based on the data given in Table 2, it is evident that the ensemble models outperformed the conventional ML methods. Specifically, the Gradient Boosting Regressor exhibited the highest R^2 value of 0.83, as well as the lowest values for MSE, RMSE, and MAE, which were 158559.33, 399.4, and 253.62, respectively. Therefore, it can be concluded that the Gradient Boosting Regressor is the most suitable model for the considered dataset.

TABLE 6. Comparison of performance metrics.

Model	MSE	RMSE	MAE	R^2
Multiple Linear Regression	352465.7	559.6	412.27	0.6
Ridge Regression	252465.6	502.46	386.69	0.73
LASSO Regression	252465.6	502.46	386.69	0.73
Support Vector Regression	557635.43	746.75	616.352	0.39
Decision Tree Regression	222190.13	471.37	286.028	0.75
Random Forest Regression	164088.8	405.07	258.06	0.82
Bagging Regressor	178778.3	422.82	265.82	0.81
ADA Boost Regressor	284581.4	533.46	427.64	0.69
Gradient Boosting Regressor	158559.33	399.44	253.62	0.83

E. HYPERPARAMETER TUNING IMPACT

The performance improvement achieved through hyperparameter tuning was quantitatively significant. The default Gradient Boosting Regressor model, without any tuning, produced a root mean squared error (RMSE) of 425.20 on the test set. In contrast, the optimized model, obtained through grid search and 5-fold cross-validation, achieved an RMSE of 406.47, MAE of 261.25, and an R^2 score of 0.819. These results highlight the effectiveness of systematic parameter tuning in boosting forecasting accuracy.

Fig. 3 illustrates the variation in RMSE across the five cross-validation folds. The scores ranged from 397 to 448, with a mean of 422.12 and a standard deviation of 16.92, indicating consistent model performance and minimal overfitting across subsets of the training data. Such stability supports the model's generalizability to unseen conditions.

The RMSE heatmap in Fig. 4 provides further insight into how different combinations of learning rate and tree depth influenced model accuracy. The combination of a lower learning rate (0.05) and deeper trees (depth = 5) yielded the best results, emphasizing the importance of deep, gradual learning in capturing complex solar irradiance patterns.

Feature importance analysis, presented in Fig. 5, reinforces the significance of incorporating solar geometry in the forecasting pipeline. The *angle of incidence*, *azimuth*, and *zenith* emerged as the top three predictors. These geometric parameters directly relate to the sun's position and its interaction with the solar panels, which strongly governs energy output variability throughout the day. Meteorological features like *cloud cover*, *humidity*, and *shortwave radiation* also contributed meaningfully, validating the hybrid approach of integrating atmospheric and geometric data sources.

Overall, the tuned GBR model not only provided the most accurate predictions but also demonstrated stability and interpretability-two crucial aspects for real-time solar forecasting systems that interface with smart grid decision engines.

VII. CONCLUSION

The application of machine learning (ML) and ensemble learning techniques has significantly enhanced the accuracy, robustness, and operational relevance of solar energy forecasting. This study presented a comprehensive comparative analysis of classical and ensemble ML models using real-world solar generation and meteorological datasets, incorporating advanced feature engineering techniques that included both atmospheric and solar geometric parameters. This work introduces several novel aspects: (i) integration of solar geometry into ML-based forecasting, (ii) systematic evaluation of ensemble methods, and (iii) statistically validated hyperparameter tuning framework that ensures model robustness.

Among the evaluated models, the Gradient Boosting Regressor (GBR) consistently outperformed others, achieving the highest coefficient of determination ($R^2 = 0.827$) and the lowest error values (RMSE = 399.44, MSE = 158559.33, MAE = 253.62). Its stage-wise optimization and ability to model complex interactions make it especially well-suited for nonlinear, time-variant solar forecasting tasks. In contrast, Support Vector Regression demonstrated limitations in adaptability and interpretability under high-dimensional feature conditions.

One of the novel contributions of this work lies in the integration of solar geometric features-such as zenith, azimuth, and angle of incidence-into ML pipelines, which is rarely addressed in previous studies. Additionally, model tuning was standardized using cross-validation, ensuring a fair and reproducible comparison across all algorithms.

The ensemble learning models, particularly Decision Tree Regression and Random Forest, also exhibited competitive performance with good interpretability, making them viable options for grid-aware solar energy forecasting systems. Overall, the ensemble approach proved superior to individual models, effectively capturing nonlinear dependencies and reducing generalization error.

Despite these advancements, some challenges persist. Forecasting accuracy is highly dependent on the availability of high-resolution and noise-free data, which is often limited in many regions. Moreover, model generalization across diverse climatic zones remains an open issue. Ensemble techniques, while more accurate, also incur greater computational costs, which may restrict real-time applicability in resource-constrained environments.

This work contributes to the field by: (I) integrating solar geometric features for enhanced forecasting; (II) evaluating a diverse set of ML models under consistent conditions; (III) applying statistical validation of model performance; and (IV) proposing a deployment-ready forecasting workflow for real-time systems. Additionally, adopting high-speed hardware such as solid-state DC breakers [48] and real-time impedance measurement tools [49], [50] can strengthen forecasting system responsiveness, particularly when embedded in microgrid controllers. For enhanced grid resilience, forecasting solutions may be aligned with restoration strategies based on multi-energy generation units as proposed in [51]. Furthermore, long-term energy balancing can benefit from integrating emerging storage technologies like gravity-based systems [52].

Future Research

To address these issues and further enhance the accuracy and applicability of solar energy forecasting models, future research could focus on:

- Incorporating additional data sources and utilizing advanced ML algorithms can improve the accuracy of NWP models, which are crucial for PV forecasting.
- Developing more detailed models of atmospheric processes could lead to better predictions of solar irradiance and power output.
- Techniques that can handle large and complex datasets will be essential as the volume of available data continues to grow.
- Leveraging historical data through methods such as ARIMA, ETS, ANNs, and SVR can optimize grid management and improve the efficiency of solar energy plants.
- Aim to deploy the optimized model within a Supervisory Control and Data Acquisition (SCADA) system or microgrid control dashboard. Real-time meteorological inputs collected via IoT sensors or weather APIs will be processed in a rolling window for near-term forecast generation.
- Incorporating digital twin principles, as discussed by Yalavarthy et al. [53], could enable predictive control and adaptive tuning of the forecasting system.

REFERENCES

[1] G. Tziolis, A. Livera, J. Montes-Romero, S. Theocharides, G. Makrides, and G. E. Georghiou, "Direct short-term net load forecasting based on machine learning principles for solar-integrated microgrids," *IEEE Access*, vol. 11, pp. 102038–102049, 2023.

[2] M. Mohamed, F. E. Mahmood, M. A. Abd, A. Chandra, and B. Singh, "Dynamic forecasting of solar energy microgrid systems using feature engineering," *IEEE Trans. Ind. Appl.*, vol. 58, no. 6, pp. 7857–7869, Nov. 2022.

[3] A. Bouraiou, A. Bekraoui, A. Necaibia, A. Rouabhia, N. Boutasseta, S. Khelifi, S. Padmanaban, B. Khan, M. S. Bouakkaz, I. Attoui, and R. Dabou, "Field investigation of PV pumping system ageing failures operation under Saharan environment," *Sol. Energy*, vol. 243, pp. 142–152, Sep. 2022.

[4] G. V. B. Kumar, P. K., S. P., and S. M. Muyeen, "Analysis of control strategies for smoothing of solar PV fluctuations with storage devices," *Energy Rep.*, vol. 9, pp. 163–177, Dec. 2023.

[5] M. A. Nasab, M. Zand, A. A. Dashtaki, M. A. Nasab, S. Padmanaban, F. Blaabjerg, and J. C. Vasquez Q, "Uncertainty compensation with coordinated control of EVs and DER systems in smart grids," *Sol. Energy*, vol. 263, Oct. 2023, Art. no. 111920.

[6] A. M. Alam, I. A. Razee, and M. Zunaed, "Solar PV power forecasting using traditional methods and machine learning techniques," in *Proc. IEEE Kansas Power Energy Conf. (KPEC)*, Apr. 2021, pp. 1–5.

[7] M. A. Nasab, M. Ali Dashtaki, B. Ehsanmaleki, M. Zand, M. A. Nasab, and P. Sanjeevikumar, "LFC of smart, interconnected power system in the presence of renewable energy sources using coordinated control design of hybrid electric vehicles," *Renew. Energy Focus*, vol. 50, Sep. 2024, Art. no. 100609.

[8] J. Gaboitaolelwe, A. M. Zungeru, A. Yahya, C. K. Lebekwe, D. N. Vinod, and A. O. Salau, "Machine learning based solar photovoltaic power forecasting: A review and comparison," *IEEE Access*, vol. 11, pp. 40820–40845, 2023.

[9] M. Piliougine, R. A. Guejia-Burbano, and G. Spagnuolo, "Detecting partial shadowing and mismatching phenomena in photovoltaic arrays by machine learning techniques," *IEEE Open J. Ind. Electron. Soc.*, vol. 3, pp. 507–521, 2022.

[10] I. Jebli, F.-Z. Belouadha, and M. I. Kabbaj, "The forecasting of solar energy based on machine learning," in *Proc. Int. Conf. Electr. Inf. Technol. (ICEIT)*, Mar. 2020, pp. 1–8.

[11] E. D. Obando, S. X. Carvajal, and J. Pineda Agudelo, "Solar radiation prediction using machine learning techniques: A review," *IEEE Latin Amer. Trans.*, vol. 17, no. 4, pp. 684–697, Apr. 2019.

[12] K. K. Hiran, D. Khazanchi, A. K. Vyas, and S. Padmanaban, *Machine Learning for Sustainable Development*, vol. 9. Berlin, Germany: Walter de Gruyter GmbH & Co KG, 2021.

[13] N. E. Benti, M. D. Chaka, and A. G. Semie, "Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects," *Sustainability*, vol. 15, no. 9, p. 7087, Apr. 2023. [Online]. Available: https://www.mdpi.com/2071-1050/15/9/7087

[14] V. Demir and H. Citakoglu, "Forecasting of solar radiation using different machine learning approaches," *Neural Comput. Appl.*, vol. 35, no. 1, pp. 887–906, Jan. 2023.

[15] B. M. Ali, "Solar energy forecasting techniques based on machine learning: Survey," in *Proc. 6th Int. Conf. Eng. Technol. Appl. (IICETA)*, Jul. 2023, pp. 814–819.

[16] P. Singh, N. K. Singh, and A. K. Singh, "Solar photovoltaic energy forecasting using machine learning and deep learning technique," in *Proc. IEEE 9th Uttar Pradesh Sect. Int. Conf. Electr., Electron. Comput. Eng. (UPCON)*, Dec. 2022, pp. 1–6.

[17] U. Munawar and Z. Wang, "A framework of using machine learning approaches for short-term solar power forecasting," *J. Electr. Eng. Technol.*, vol. 15, no. 2, pp. 561–569, Mar. 2020.

[18] N. Hari, M. Ahsan, S. Ramasamy, P. Sanjeevikumar, A. Albarbar, and F. Blaabjerg, "Gallium nitride power electronic devices modeling using machine learning," *IEEE Access*, vol. 8, pp. 119654–119667, 2020.

[19] S. Aslam, H. Herodotou, S. M. Mohsin, N. Javaid, N. Ashraf, and S. Aslam, "A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids," *Renew. Sustain. Energy Rev.*, vol. 144, Jul. 2021, Art. no. 110992.

[20] M. Abubakar, Y. Che, M. Faheem, M. S. Bhutta, and A. Q. Mudasar, "Intelligent modeling and optimization of solar plant production integration in the smart grid using machine learning models," *Adv. Energy Sustainability Res.*, vol. 5, no. 4, Apr. 2024, Art. no. 2300160.

[21] A. Zafar, Y. Che, M. Faheem, M. Abubakar, S. Ali, and M. S. Bhutta, "Machine learning autoencoder-based parameters prediction for solar power generation systems in smart grid," *IET Smart Grid*, vol. 7, no. 3, pp. 328–350, Jun. 2024.

[22] M. Yang, Y. Jiang, W. Zhang, Y. Li, and X. Su, "Short-term interval prediction strategy of photovoltaic power based on meteorological reconstruction with spatiotemporal correlation and multi-factor interval constraints," *Renew. Energy*, vol. 237, Dec. 2024, Art. no. 121834.

[23] J. Yuqi, A. An, Z. Lu, H. Ping, and L. Xiaomei, "Short-term load forecasting based on temporal importance analysis and feature extraction," *Electr. Power Syst. Res.*, vol. 244, Jul. 2025, Art. no. 111551.

[24] M. S. Bhutta, T. Xuebang, M. Faheem, F. M. Almasoudi, K. S. S. Alatawi, and H. Guo, "Neuro-fuzzy based high-voltage DC model to optimize frequency stability of an offshore wind farm," *Processes*, vol. 11, no. 7, p. 2049, Jul. 2023.

[25] A. I. Kawoosa, D. Prashar, M. Faheem, N. Jha, and A. A. Khan, "Using machine learning ensemble method for detection of energy theft in smart meters," *IET Gener., Transmiss. Distrib.*, vol. 17, no. 21, pp. 4794–4809, Nov. 2023.

[26] A. Iqbal and A. Kawoosa, "Using XGBoost for intelligent electricity theft detection with socioeconomic data," *IEEE Access*, vol. 11, pp. 12345–12355, 2023.

[27] R. Ahmed and M. Khalid, "Real-time solar power forecasting using machine learning models and real-world data," *Appl. Energy*, vol. 276, Jan. 2020, Art. no. 115435.

[28] Z. Miao, X. Meng, X. Li, B. Liang, and H. Watanabe, "Enhancement of net output power of thermoelectric modules with a novel air-water combination," *Appl. Thermal Eng.*, vol. 258, Jan. 2025, Art. no. 124745.

[29] N. Rahimi, S. Park, W. Choi, B. Oh, S. Kim, Y.-H. Cho, S. Ahn, C. Chong, D. Kim, C. Jin, and D. Lee, "A comprehensive review on ensemble solar power forecasting algorithms," *J. Electr. Eng. Technol.*, vol. 18, no. 2, pp. 719–733, Mar. 2023.

[30] S. F. Stefenon, M. H. D. M. Ribeiro, A. Nied, K.-C. Yow, V. C. Mariani, L. D. S. Coelho, and L. O. Seman, "Time series forecasting using ensemble learning methods for emergency prevention in hydroelectric power plants with dam," *Electr. Power Syst. Res.*, vol. 202, Jan. 2022, Art. no. 107584.

[31] Z. Wang, S. Huang, Z. Mu, G. Leng, W. Duan, H. Ling, J. Xu, X. Zheng, P. Li, Z. Li, W. Guo, Y. Li, M. Deng, and J. Peng, "Relative humidity and solar radiation exacerbate snow drought risk in the headstreams of the Tarim river," *Atmos. Res.*, vol. 297, Jan. 2024, Art. no. 107091.

[32] W. Mol, B. Heusinkveld, M. R. Mangan, O. Hartogensis, M. Veerman, and C. van Heerwaarden, "Observed patterns of surface solar irradiance under cloudy and clear-sky conditions," *Quart. J. Roy. Meteorolog. Soc.*, vol. 150, no. 761, pp. 2338–2363, Apr. 2024.

[33] P. K. Pathak, D. G. Roy, A. K. Yadav, S. Padmanaban, F. Blaabjerg, and B. Khan, "A State-of-the-Art review on heat extraction methodologies of photovoltaic/thermal system," *IEEE Access*, vol. 11, pp. 49738–49759, 2023.

[34] H. Zhao, Q. Zong, H. Zhou, W. Yao, K. Sun, Y. Zhou, and J. Wen, "Frequency-voltage active support strategy for hybrid wind farms based on grid-following and grid-forming hierarchical subgroup control," *CSEE J. Power Energy Syst.*, vol. 11, no. 1, pp. 65–77, 2025.

[35] M. Yang, R. Che, X. Yu, and X. Su, "Dual NWP wind speed correction based on trend fusion and fluctuation clustering and its application in short-term wind power prediction," *Energy*, vol. 302, Sep. 2024, Art. no. 131802.

[36] S. Datta, A. Baul, G. C. Sarker, P. K. Sadhu, and D. R. Hodges, "A comprehensive review of the application of machine learning in fabrication and implementation of photovoltaic systems," *IEEE Access*, vol. 11, pp. 77750–77778, 2023.

[37] F. Delussu, D. Manzione, R. Meo, G. Ottino, and M. Asare, "Experiments and comparison of digital twinning of photovoltaic panels by machine learning models and a cyber-physical model in modelica," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4018–4028, Jun. 2022.

[38] S. Atique, S. Noureen, V. Roy, S. Bayne, and J. Macfie, "Time series forecasting of total daily solar energy generation: A comparative analysis between ARIMA and machine learning techniques," in *Proc. IEEE Green Technol. Conf.(GreenTech)*, Apr. 2020, pp. 175–180.

[39] M. V. Khaire, A. G. Thosar, and V. N. Pande, "Prediction of solar power generation using NWP and machine learning," in *Proc. 3rd Asian Conf. Innov. Technol. (ASIANCON)*, Aug. 2023, pp. 1–6.

[40] Z. Yao, Y. Lum, A. Johnston, L. M. Mejía-Mendoza, X. Zhou, Y. Wen, A. Aspuru-Guzik, E. H. Sargent, and Z. W. Seh, "Machine learning for a sustainable energy future," *Nature Rev. Mater.*, vol. 8, no. 3, pp. 202–215, Oct. 2022.

[41] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renew. Energy*, vol. 105, pp. 569–582, May 2017.

[42] U. K. Das, K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. Van Deventer, B. Horan, and A. Stojcevski, "Forecasting of photovoltaic power generation and model optimization: A review," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 912–928, Jan. 2018.

[43] A. Zendehboudi, M. A. Baseer, and R. Saidur, "Application of support vector machine models for forecasting solar and wind energy resources: A review," *J. Cleaner Prod.*, vol. 199, pp. 272–285, Oct. 2018.

[44] A. Ahmed and M. Khalid, "A review on the selected applications of forecasting models in renewable power systems," *Renew. Sustain. Energy Rev.*, vol. 100, pp. 9–21, Feb. 2019.

[45] A. Mosavi, M. Salimi, S. F. Ardabili, T. Rabczuk, S. Shamshirband, and A. R. Varkonyi-Koczy, "State of the art of machine learning models in energy systems, a systematic review," *Energies*, vol. 12, no. 7, p. 1301, Apr. 2019.

[46] Q.-T. Phan, Y.-K. Wu, and Q.-D. Phan, "Enhancing one-day-ahead probabilistic solar power forecast with a hybrid transformer-LUBE model and missing data imputation," *IEEE Trans. Ind. Appl.*, vol. 60, no. 1, pp. 1396–1408, Jan. 2024.

[47] Meshva. *Solar Energy Power Generation Dataset*. Accessed: Oct. 6, 2024. [Online]. Available: https://www.kaggle.com/datasets/stucom/solar-energy-power-generation-dataset

[48] S. Wang, M. Li, K. Qin, Q. Xin, J. Ma, N. Hou, and T. Liu, "A thyristor-based solid-state DC circuit breaker with a three-winding coupled inductor," *IEEE Trans. Power Electron.*, vol. 40, no. 4, pp. 6192–6202, Apr. 2025.

[49] Q. Rong, P. Hu, Y. Yu, D. Wang, Y. Cao, and H. Xin, "Virtual external perturbance-based impedance measurement of grid-connected converter," *IEEE Trans. Ind. Electron.*, vol. 72, no. 3, pp. 2644–2654, Mar. 2025.

[50] Q. Rong, P. Hu, L. Wang, Y. Li, Y. Yu, D. Wang, and Y. Cao, "Asymmetric sampling disturbance-based universal impedance measurement method for converters," *IEEE Trans. Power Electron.*, vol. 39, no. 12, pp. 15457–15461, Dec. 2024.

[51] M. Zhang, S. Cai, Y. Xie, B. Zhou, W. Zheng, Q. Wu, and J. Wen, "Supply resilience constrained scheduling of MEGs for distribution system restoration: A stochastic model and FW-PH algorithm," *IEEE Trans. Smart Grid*, vol. 16, no. 1, pp. 194–208, Jan. 2025.

[52] F.-F. Li, J.-Z. Xie, Y.-F. Fan, and J. Qiu, "Potential of different forms of gravity energy storage," *Sustain. Energy Technol. Assessments*, vol. 64, Apr. 2024, Art. no. 103728.

[53] U. R. S. Yalavarthy, N. B. Kumar, A. R. V. Babu, R. P. Narasipuram, and S. Padmanaban, "Digital twin technology in electric and self-navigating vehicles: Readiness, convergence, and future directions," *Energy Convers. Manag., X*, vol. 26, Apr. 2025, Art. no. 100949.

ATTULURI R. VIJAY BABU received the B.Tech. degree in electrical and electronics engineering from the Vignan's Engineering College, Guntur, India, in 2010, the M.E. degree in energy engineering from the PSG College of Technology, Coimbatore, India, in 2012, and the Ph.D. degree in air-breathing fuel cells from the Vignan's Foundation for Science Technology and Research, India, in 2020. Currently, he is an Associate Professor with the Department of Electrical and Electronics Engineering and the Associate Dean of the Research and Development, Vignan's Foundation for Science Technology and Research. He has ten years of teaching experience. He has published/presented 78 research papers in national and international journals and conferences. He holds three patents. His research interests include fuel cells, renewable energy systems, and electric vehicles. He is a reviewer for many reputed journals.

N. BHARATH KUMAR (Member, IEEE) received the B.Tech. degree in electrical and electronics engineering from JNTU, Hyderabad, India, in 2005, the M.Tech. degree in power and industrial drives specialization from JNTU, Kakinada, India, in 2011, and the Ph.D. degree in electrical engineering from IIT Kharagpur, in 2023. He is currently with the Department of Electrical and Electronics Engineering, Vignan's Foundation for Science Technology and Research, India. To his credit, he published about 40 research articles and four Indian patents. His research interests include power converters, industrial drives, digital twins, electric vehicles, artificial intelligence, and machine learning.

RAJANAND PATNAIK NARASIPURAM received the Electrical and Electronics Engineering Diploma degree from the Government Polytechnic Vijayawada, in 2011, the Bachelor of Technology degree in electrical and electronics engineering and the master's degree in power electronics and electrical drives from JNTUK, in 2014 and 2016, respectively, and the Ph.D. degree in transportation electrification from the Vignan's Foundation for Science Technology and Research, in 2024. He is currently associated with Cyient Ltd. as a Chief Engineer. He has a decade of industrial research experience in transportation electrification. He holds two patents and several reputed peer-reviewed publications, is involved in several editorial activities, and reviewer for reputed journals. His current research interests include transportation electrification, electric vehicle charging, V2X, and systems engineering.

SOUNDHAR PERIYANNAN received the Bachelor of Engineering degree in mechatronics engineering from Anna University, in 2009, and the Master's Diploma degree in wind power development from Amrita University, in 2012. He is currently associated with Atria Power. He has ten years of industrial research experience in wind and solar operations. His current research interests include data analysis, performance monitoring of wind, and solar energy systems.

ALIREZA HOSSEINPOUR was born in Zabol, Iran, in 1985. He received the B.Sc. degree (Hons.) in electrical engineering from the University of Sistan and Baluchestan, Zahedan, Iran, in 2007, the M.Sc. degree in electrical engineering from the Ferdowsi University of Mashhad, Iran, in 2010, and the Ph.D. degree in electrical engineering from the Shiraz University of Technology, in 2018. Following receipt of the Ph.D. degree, he joined the University of Zabol, Zabol. His research interests include wind turbines, renewable energy, permanent magnet synchronous machine drives, four-switch three-phase inverters, power electronics, and the design of hybrid electrical vehicles.

AYMEN FLAH was born in Gabes, Tunisia, in 1983. He received the bachelor's degree in electrical engineering and the M.Tech. degree from ENIG, Tunisia, in 2007 and 2009, respectively, and the Ph.D. degree from the Department of Electrical Engineering, in 2012. He has academic experience of 11 years. He has published more than 100 research articles in reputed journals and more than 40 research papers in international conferences and book chapters. His research interests include electric vehicles, power systems, and renewable energy.