IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 May 2024, accepted 13 July 2024, date of publication 18 July 2024, date of current version 13 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430385

RESEARCH ARTICLE

Off-Grid Small-Scale Power Forecasting Using Optimized Machine Learning Algorithms

AADYASHA PATEL[1] **AND O. V. GNANA SWATHIKA**[2]**, (Senior Member, IEEE)**
[1]School of Electrical Engineering, Vellore Institute of Technology, Chennai 600127, India
[2]Centre for Smart Grid Technologies, School of Electrical Engineering, Vellore Institute of Technology, Chennai 600127, India

Corresponding author: O. V. Gnana Swathika (gnanaswathika.ov@vit.ac.in)

This work was supported by the Vellore Institute of Technology, India.

ABSTRACT Solar energy is highly unstable. Total photovoltaic energy generated varies based on changes in weather, climate and seasons. Owing to the arbitrariness of solar energy, photovoltaic output is prone to frequent power oscillations. Due to this reason, photovoltaic power prediction becomes obligatory. Through comprehensive analysis and critical examination, this research endeavours to shed light on the pursuit with regard to embracing solar energy as a sustainable channel of power generation. Additionally, the study aligns with the United Nations' commitment on achieving the Sustainable Development Goals. The objective of this study is to prognosticate the electricity output from photovoltaic panel through compilation of meteorological parameters and computation of hourly global solar radiation upon an inclined surface, alongside the resultant energy yield from an inclined photovoltaic panel. This prediction is executed through the deployment of four optimized machine learning methods, specifically Support Vector Machines, Ensemble of Trees, Gaussian Process Regression and Neural Networks. Bayesian Optimization is utilized to optimize the machine learning models by tuning its hyperparameters. The analysis is conducted across three distinct dataset classifications: annual, monthly and seasonal. The empirical findings underscore that optimized Ensemble of Trees exhibits superior performance across all dataset classifications and also necessitates shortest training duration compared to its counterparts.

INDEX TERMS Bayesian optimization, Ensemble of Trees, hyperparameter tuning, surface inclination angle, LSboost, machine learning, power forecasting, sustainable energy.

I. INTRODUCTION

India, a nation which is one of the world's largest and fastest growing economies, grapples with the task of ensuring that every facet of the society has access to electricity. The nation has grown 7.2% economically during the fiscal year 2022-23, and the growth is projected to accelerate even further in the coming years, propelling the country to become the 3rd largest economy in the next 5 years [1]. As such, with economic progress, India has made commendable progress towards the complete electrification of the country. Although, there still exists some significant disparities with the level of infrastructure between the urban and the rural areas. While urban centres enjoy relatively robust electricity infrastructure, with constant improvements, the rural areas have only in the recent years been electrified. India had launched government initiative called ''Saubhagya'', with the goal of electrifying the entire nation completely by 2018 [2]. By the October of 2018, through the Saubhagya initiative [3], the country has a total electricity coverage of 95%.

Historically, India has heavily relied on non-renewable energy sources, such as natural gas, oil, and coal, to meet its growing demands of energy. However, this dependence comes at a considerable cost. The combustion of fossil fuels contributes to air pollution, environmental degradation, and climate change posing significant health risks and ecological consequences. Furthermore, the finite nature of non-renewable resources underscores the urgency of transitioning towards sustainable alternatives. In recent years, there has been a paradigm shift towards renewable energy sources driven by environmental concerns, energy security, and technological advancements. Solar, wind, hydroelectric,

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

and biomass energy offer clean, sustainable alternatives to fossil fuels with numerous advantages. These renewable sources foster economic growth, boost energy security, and decrease greenhouse gas emissions through job creation and technological innovation. Furthermore, investments in renewable energy infrastructure stimulate innovation and technological advancement, positioning nations at the forefront of the global energy transition.

Among renewable energy sources, solar energy holds immense promise for India, given its abundant sunlight and vast solar potential. As per the energy profile [4] published by the International Renewable Energy Agency, the installed capacity of solar energy in India was logged at 39.2 GW in 2020, which was up from a mere 0.1 GW in 2010. By 2021, the total installed capacity of solar energy has been further increased to 65.9 GW and it contributes to about 4% of the total electricity generating capacity of the country, out of which 19% is only renewable energy. India has implemented several policies and schemes to promote solar energy deployment, including the National Solar Mission, Solar Parks, and subsidies for rooftop solar installations. The newest policy is named ''Pradhan Mantri Suryodaya Yojana'' [5], and the goal of this policy is to install solar panels on 1 crore Indian households. These policies aim to incentivize investment in solar energy infrastructure, enhance grid integration, and accelerate the transition towards clean energy. Solar energy offers numerous advantages, including abundant availability, low operating costs, and minimal environmental impact. However, challenges such as intermittency, storage limitations, and upfront costs remain significant barriers to widespread adoption. The United Nations' Sustainable Development Goals (SDG) emphasize the prominence of renewable energy in achieving global sustainability targets [4]. The aim of the SDG is to ensure access to affordable, reliable, sustainable, and modern energy for all, with a specific focus on increasing the share of renewable energy in the global energy mix.

Photovoltaic (PV) panels also known as solar panels are a cornerstone of renewable energy technology, converting sunlight into electricity through the photovoltaic effect. As the world increasingly embraces clean energy solutions, the demand for PV panels has surged, driving innovations in efficiency, affordability, and scalability. However, to maximize the potential of solar energy and ensuring optimal utilization of PV panels, accurate prediction of energy generation is paramount. Several factors influence the energy generated by PV panels including:

- Solar Irradiance: The intensity of sunlight received by the PV panel influenced by factors such as seasons, weather conditions, latitude, and time.
- Temperature: The temperature of the PV panel affects its efficiency with higher temperatures typically leading to reduced performance.
- Panel Characteristics: The type, size, orientation, tilt angle, and condition of the PV panel impact its energy generation capacity.
- Environmental Factors: Shading, dust, dirt, and other obstructions can diminish the performance of PV panels.

Climatological parameters like temperature, wind speed, solar irradiance, and humidity play a crucial role in predicting the energy generated by PV panels. Real-time data on these parameters obtained from weather stations, satellites, or on-site sensors enables precise modelling of solar energy generation. By incorporating meteorological data into predictive models, analysts can account for variations in sunlight intensity, temperature fluctuations, and other environmental factors thereby enhancing the accuracy of energy predictions for PV panels.

Machine Learning (ML) techniques offer a powerful framework for energy prediction, leveraging algorithms to analyse historical data, identify patterns, and make accurate forecasts. In the context of PV panels, ML models can learn from past energy generation data, meteorological parameters, and other relevant factors to predict future energy output. Common ML approaches for energy prediction include Support Vector Machines (SVM), Ensemble of Trees (ET), Gaussian Process Regression (GPR), and Neural Networks (NN). Certain benefits of using ML for energy prediction are:

- Adaptability: ML models possess the ability to acclimatize with varying environmental conditions and integrate new data to refine predictions over time.
- Accuracy: By learning from historical data and meteorological parameters, ML models can generate highly accurate predictions of energy generation for PV panels.
- Efficiency: ML algorithms can process large volumes of data rapidly enabling real-time or near-real-time energy prediction for operational decision-making.
- Optimization: ML models optimizes PV panel performance by identifying factors that influence energy generation and recommending adjustments to maximize efficiency.

From the existing literature, the authors in [6] offer an assessment of 24 ML models based on numerical weather predictions for forecasting day-ahead PV power using two years of data from Hungarian PV plants. This study highlights the significance of selecting input data and tuning of hyperparameters with Multilayer Perceptron emerging as the recommended model for practical applications due to its high accuracy and lower training time compared to Kernel Ridge. The investigation carried out in [7] provides a comprehensive review of ML approaches for PV power forecasting, covering factors like weather conditions, and forecasting horizons. It discusses how accurate weather forecasting can enhance PV power prediction considering variables like solar irradiance and ambient temperature. The study evaluates various ML algorithms and performance metrics, demonstrating that ML models outperform baseline methods in solar PV power forecasting, with Gradient Boost and Random Forest (RF) exhibiting particularly strong performance with default and tuned hyperparameters respectively. The research performed in [8] introduces a robust short-term load forecasting method for a university campus in Canada employing 19 regression

models to create load forecasting models evaluated through error indices. Among these models, the GPR models emerge as the top performers, showcasing their viability as load forecasting methodologies. GPR being nonparametric offers flexibility in capturing patterns without being constrained by specific functional forms making it adept at extrapolating data and providing predictive distributions with mean values and respective variances. The analysis done in [9] introduces a new low-cost portable datalogger tailored for monitoring PV systems, emphasizing its affordability and accessibility to researchers and users. By integrating a Linear Regression (LR) ML algorithm, the system accurately forecasts power generation with less than 10% error, utilizing real-time data and free software resources. The authors in [10] focussed on translating irradiance-to-power in PV power forecasting presenting and comparing ML, physical, and hybrid modelling approaches. Hybrid modelling, combining physically-calculated PV power output with ML predictions emerges as the most effective approach. The findings underscore the importance of optimizing model chains and selecting the appropriate forecasting directive with implications for enhancing the accuracy and reliability of PV power forecasts across various applications. The analysis carried out in [11] underscores the critical influence of training sets and seasonal variations on model performance in predicting PV power generation. The proposed Seasonal Clustering Forecasting Technique offers a robust approach by clustering historical climatological data resulting in improved accuracy particularly in regions with fluctuating climates. The study highlights the importance of addressing bias errors in solar irradiance data to ensure the reliability of PV power generation forecasts, suggesting avenues for further research to enhance data quality, and forecasting precision.

The research done in [12] introduces deep learning architectures, including stacked Long Short-Term Memory (LSTM) with Bayesian optimization and drop-out architecture for hourly day-ahead solar forecasting, demonstrating superior performance in handling univariate and multivariate data with lower error metrics. The optimized stacked BiLSTM/LSTM model, fine-tuned using Bayesian optimization achieves high forecasting accuracy by optimizing six hyperparameters during training. The flexibility of the proposed model allows for selecting architecture types and fine-tuning, enhancing its effectiveness for solar energy and power forecasting, especially when incorporating Plane of Array (POA) data. Despite its success, future research is recommended to improve POA measurement accuracy and consider module cover effects to further enhance forecasting precision. The study executed in [13] utilizes supervised learning algorithms for predicting PV power by means of SVMs, LRs, K-Nearest Neighbours and Artificial Neural Networks (ANN). By evaluating these methods against original data from a PV system, the research underscores the effectiveness of ML techniques in accurately estimating power generation from meteorological variables. Notably, ANNs demonstrated superior performance in PV power prediction, surpassing other techniques in terms of error metrics. The inspection conducted in [14] displays a comprehensive evaluation of Ensemble ML (EML) algorithms for predicting generated PV power using meteorological data, filling a gap in existing literature by providing serialized steps from data preparation to performance valuation. The study facilitates informed decision-making for greenfield solar projects offering insights into the most suitable prediction models. The findings highlight the effectiveness of voting and stacking algorithms in achieving high prediction accuracy suggesting the applicability of the proposed EML framework for extensive solar PV power plants in comparable climatological conditions with potential extensions to techno-financial analysis and optimization of algorithm performance. Several models for estimating module temperature, ranging from Nominal Operating Cell Temperature (NOCT)-based to energy balance approaches were compared using experimental data in [15]. The optimal model constructed on an energy balance and incorporating a heat transfer coefficient derived from experimental data achieved a relative Root Mean Square Error of 19.8%. Additionally, two models accounting for module temperature's influence on electrical efficiency were tested. The research paper [16] provides a wide-ranging literature review and comparative analysis of procedures for estimating cell temperature, global solar radiation, and PV prediction. Eleven decomposition models for evaluating daily global solar radiation were evaluated, with the Collares-Pereira and Rabl (CPR) model modified by Gueymard (CPRG) model showing overall accuracy. The study also examined seven models for assessing cell temperature with the Skoplaki models yielding the most accurate power generation forecasts and the Mattei models yielding next best results.

As the renewable energy sector continues to evolve, the integration of advanced analytics and ML-driven insights will play a key part in shaping a sustainable energy future. The following are the core contributions of this research paper:

- Calculate hourly global solar radiation on an inclined surface (G_γ) and the energy delivered by the PV panel (E_{PV}) from the obtained meteorological parameters.
- Propose optimized ML models to evaluate annual, monthly, and seasonal forecast based on hourly estimates of E_{PV}.
- Tune the hyperparameters with Bayesian Optimization.
- Evaluate and compare the performance accuracy of the optimized ML models for predicting E_{PV}.

The structure of the paper is as follows: Section II gives an account of the methodology followed in the research paper to forecast EPV. Section III explains the predicted results of EPV for the assessment criteria along with comparison and evaluation metrics. Section IV concludes the article with section V providing the future scope of work in the field of power forecasting.

II. METHODOLOGY

The flowchart depicted in Fig. 1 outlines a process for predicting PV power output using ML models. The process begins with data collection where data is gathered through the National Solar Radiation Database (NSRDB) Data Viewer. This data includes various parameters: Ozone (O), Solar Zenith Angle (SZA), Precipitable Water (PW), Temperature (T), Dew Point (DP), Diffuse Horizontal Irradiance (DHI), Direct Normal Irradiance (DNI), Global Horizontal Irradiance (GHI), Relative Humidity (RH), Surface Albedo (ρ), Pressure (P), Wind Direction (WD), and Wind Speed (v_w). Following data collection, data pre-processing takes place which involves removal of anomalous data and outliers. Next step is calculation of solar angles and multiplying factor to compute G_γ. In the next step, the cell temperature (T_c) is calculated, which is then used to compute E_{PV}. The subsequent stage comprises feature selection to identify the most relevant data points for the ML models. The data is then divided and normalized in the data division and data normalization stages respectively. Subsequently, ML models such as SVM, ET, GPR, and NN are deployed for PV power prediction. The final step is performance analysis, where the effectiveness of the ML models in predicting PV power output is evaluated.

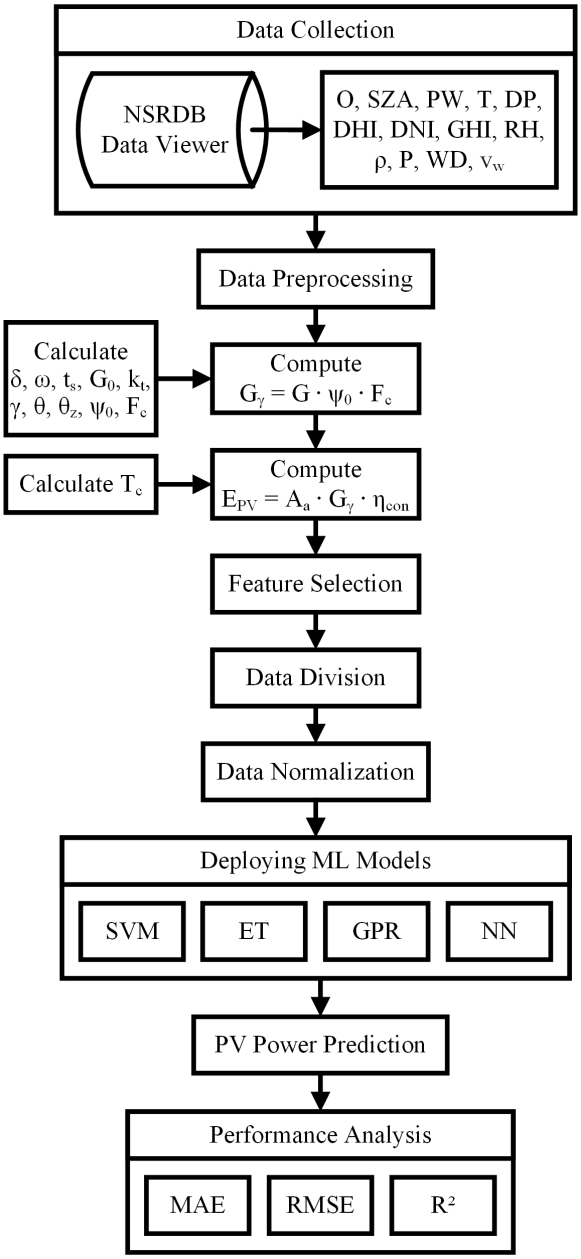

FIGURE 1. Methodology of the proposed system.

A. DATA COLLECTION AND DESCRIPTION

The geographical coordinates pinpointing the Vellore Institute of Technology, Chennai Campus (VITCC) is 12.8438° N and 80.1533° E. The pertinent dataset utilized in this investigation was procured from the U.S. National Renewable Energy Laboratory's NSRDB website [17], [18]. This data was obtained by providing the precise latitude and longitude coordinates of VITCC into the NSRDB Viewer webpage. The dataset was then retrieved covering a span of five consecutive calendar years, commencing from January 1st, 2016, and culminating on December 31st, 2020, with a granularity of one-hour intervals. The acquired dataset encompasses an array of meteorological parameters specific to the designated locale, encompassing ozone levels, solar zenith angle, precipitable water content, temperature readings, dew point values, DHI, DNI, GHI, relative humidity metrics, surface albedo, atmospheric pressure, wind direction and speed, as well as chronological date and time stamps.

B. DATA PRE-PROCESSING

Data pre-processing constitutes a pivotal phase aimed at orchestrating the input data, thereby increasing the likelihood of yielding competent predictive outputs. Within this critical stage, data captured during periods devoid of daylight, delineated as the interval between sunset and sunrise, is systematically eliminated. The utilization of box plots elucidates the configuration and dispersion patterns inherent within the dataset, serving as a diagnostic tool to discern the existence of outliers, should they be present. The depiction of hourly GHI distributions for each month from 2016 to 2020 is graphically rendered in the form of box plots, as shown in Fig. 2. The data is almost symmetrically distributed for all the months of the years from 2016 to 2020 with the median line positioned in the centre of the box for just about all the months.

C. CALCULATION OF HOURLY GLOBAL SOLAR RADIATION ON AN INCLINED SURFACE

Consider a PV panel with a maximal power rating of 250W [19]. Enclosed within Table 1 and Table 2 are the mechanical and electrical specifications corresponding to this PV panel respectively. Within the scholarly work referenced [20], the authors devised a model aimed at

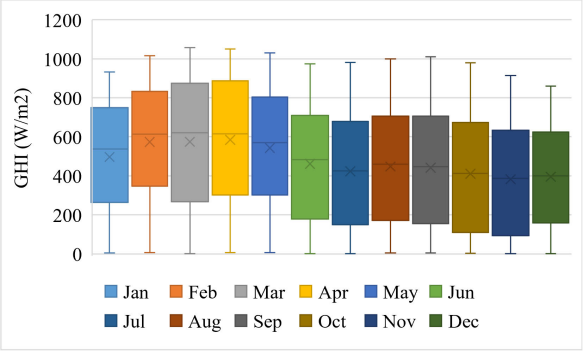

FIGURE 2. Hourly GHI distribution for each month from 2016 to 2020.

TABLE 1. Mechanical characteristics of the solar panel.

Mechanical Characteristics	Specification
Model Number	TS250
Manufactured by	Tata Power Solar
Solar panel type	Polycrystalline
Number of cells	60
Frame material	Anodized Aluminium
Module dimensions (mm)	1667x1000x33

TABLE 2. Electrical characteristics of the solar panel.

Electrical Characteristics	Specification
Maximum power (P_{max})	250 W
Power tolerance	$\pm$ 2.5 %
Voltage at maximum power (V_{mp})	30.7 V
Current at maximum power (I_{mp})	8.16 A
Open circuit voltage (V_{oc})	38.1 V
Short circuit current (I_{sc})	8.58 A
Maximum system voltage	1000 V
Standard Test Conditions (STC)	1000 W/m^2, AM = 1.5, 25°C
Module efficiency (η_{STC})	15.00 %
Temperature coefficient, β	-0.4383 %/°C
NOCT	47 °C

estimating G_γ under clear sky conditions. However, the authors proffered an innovative extension to this model, facilitating estimation under diverse sky conditions. While an array of models exists for the estimation of G_γ, these typically break down the solar radiation elements into direct and diffuse components. Notably, the model proposed herein eschews this decomposition, executing calculations holistically without fragmenting the components.

Within the scope of work, the following solar angle equations as elucidated in [16] and [21] are worked out in the process of calculating G_γ. Solar declination angle (δ) ranges between $\pm 23.45°$ where the maximum and minimum solar declination angles occur yearly on 22nd June and 21st or 22nd December respectively. To find this value, Cooper's equation as given in (1) is used.

$$\delta = 23.45° * (\sin((360/365)(284 + d)) \quad \text{(in degrees)} \quad (1)$$

where, d is the Julian day number i.e., 1st of January indicates d = 1 and 31st December indicates d = 365 or 366 depending on the year being a leap year or not. The value 23.45° is the Earth's axial tilt.

Solar hour angle (ω) is calculated by using (2).

$$\omega = (360°/24) * (t_s - 12) \quad \text{(in degrees)} \quad (2)$$

The value of ω varies with the position of the sun as,

$$\omega = \begin{cases} -ve, & \text{beforenoon} \\ 0°, & \text{noon} \\ +ve, & \text{afternoon} \end{cases}$$

The solar time (t_s) and local time are not same. t_s is necessary to estimate ω as given in (3).

$$t_s = t_{loc} \pm (4 * (L_{st} - L_{loc})) + E$$

Since India lies in the eastern hemisphere,

$$t_s = t_{loc} - (4 * (L_{st} - L_{loc})) + E \quad \text{(in minutes)} \quad (3)$$

where, t_{loc} is the local time signifying clock time, the value 4 implies the time taken in minutes by the sun to traverse 1° of longitude, L_{st} is the standard meridian of the local time zone. India's standard meridian is situated at 82° 3' E, L_{loc} is the longitude of the location and E is the equation of time correction as in (4).

$$\begin{aligned} E = {} & 229.2 * (0.000075 + (0.001868 * \cos B) \\ & - (0.032077 * \sin B) - (0.014615 * \cos 2B) \\ & - (0.04089 * \sin 2B)) \quad \text{(in minutes)} \end{aligned} \quad (4)$$

where, $B = (d - 1)(360/365)$

Hourly extra-terrestrial global solar radiation on horizontal surface is given by (5) as,

$$\begin{aligned} G_0 = {} & I_{sol} * (1 + (0.033(\cos((360d)/365))) \\ & * ((\cos\phi \ \cos\delta \ \cos\omega) + (\sin\phi \ \sin\delta)) \ (\text{in } W/m^2) \end{aligned} \quad (5)$$

where, I_{sol} is the solar constant (1367 W/m^2), the value 0.033 is the variation of extra-terrestrial radiation flux due to variation of the earth-sun distance, the variation between the distance of the sun and the earth causes variation in the extra-terrestrial radiation flux is 3.3% and ϕ is the latitude of the site.

Hourly clearness index (k_t) is the ratio between GHI and hourly extra-terrestrial global solar radiation on horizontal surface, as in (6).

$$k_t = GHI/G_0 \quad (6)$$

G_0 represents the overall solar energy falling per unit area at a particular site over a given duration of time. However, when a surface is inclined or tilted the solar angle of incidence affects the amount of solar energy received. The angle at which the PV panel is tilted is called surface inclination angle (γ) and is given in (7).

$$\gamma = \phi - \delta \ \text{(in degrees)} \quad (7)$$

The solar incidence and zenith angles are as in (8) and (9) respectively.

$$\begin{aligned} \theta = {} & \arccos((\sin\delta \ \sin(\phi - \gamma)) \\ & + (\cos\delta \ \cos(\phi - \gamma) \ \cos\omega)) \quad \text{(in degrees)} \end{aligned} \quad (8)$$

$$\theta_z = \arccos((\sin\delta\ \sin\phi) + (\cos\delta\ \cos\phi\ \cos\omega)) \quad \text{(in degrees)} \tag{9}$$

Both θ and θ_z vary from 0° to 90° as

$$\begin{cases} 0°, & \text{noon} \\ 90°, & \text{sunrise and sunset} \end{cases}$$

The function ψ_0, as in (10), accounts for the change in radiation received by the PV panel by translating the solar radiation received by an inclined surface from a horizontal surface.

$$\psi_0 = exp(-k_t * ((\pi\theta/180)^2 - (\pi\theta_z/180)^2)) \tag{10}$$

The multiplying factor (F_c) adjusts for any additional factors that may affect the solar radiation on the inclined surface, such as shading, reflections, or surface characteristics. This is given in (11).

$$F_c = 1 + \rho \sin^2(\theta/2) \tag{11}$$

The calculated data is then utilized to compute the hourly global solar radiation falling on an inclined surface by using (12).

$$G_\gamma = G * \psi_0 * F_c \quad (\text{in } W/m^2) \tag{12}$$

D. CALCULATION OF ENERGY GENERATED BY PV PANEL

The energy generated by the PV panel is calculated by means of the following equations. T_c of the PV panel is obtained from (13) by computing the second model suggested by M. Mattei [15] and [16]. It is measured in °C. The expression is given by,

$$T_c = \frac{(u_{PV}(v_w)T + G_\gamma \cdot (\tau \cdot \alpha - \eta_{STC}(1 + \beta \cdot T_{STC})))}{(u_{PV}(v_w) - \beta \cdot \eta_{STC} \cdot G_\gamma)} \tag{13}$$

where, u_{PV} is the heat exchange coefficient for the total surface of module and is given as,

$$u_{PV}(v_w) = 24.1 + 2.9(v_w) \quad (\text{in } W°C^{-1}m^{-2})$$

calculated at a wind speed, v_w of 1 m/s. τ is the transmittance of the cover system and α is the absorption coefficient of the PV cells.

Temperature affects the efficiency of PV cells in various ways. A popular method to estimate this is given by computing the conversion efficiency (η_{con}) [16] as in (14).

$$\eta_{con} = \eta_{STC}(1 - \beta(T_c - T_{STC})) \tag{14}$$

By multiplying the total area of the panel, irradiance received on an inclined surface and conversion efficiency, (15) is used to calculate the total energy delivered by the PV panel [15], [22], and [23]. This energy represents the amount of electrical power produced by the PV panel under certain climatic circumstances, such as solar irradiance levels and cell temperature.

$$E_{PV} = A_a \cdot G_\gamma \cdot \eta_{con} \quad (\text{in } Wh) \tag{15}$$

Fig. 3 shows box plot representation of the hourly distribution of E_{PV} for each month from the year 2016 to 2020. The median lies towards the bottom of the box with short lower whiskers representing a positively skewed data distribution. It is evident from the plot that an innate asymmetry lies within the dataset.

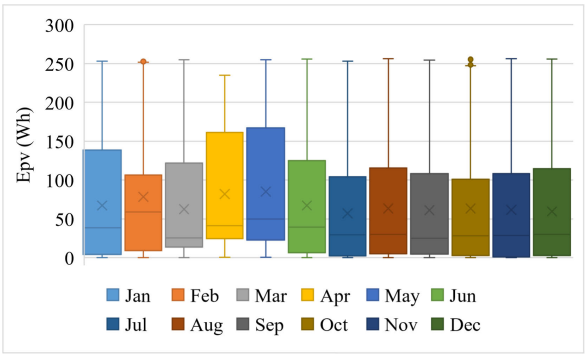

FIGURE 3. Hourly E_{PV} distribution in each month from 2016 to 2020.

E. FEATURE SELECTION

Amidst the plethora of meteorological variables, it is discerned that not all variables wield relevance in the forecasting of energy yield from PV panels. Hence, imperative to this endeavour is a correlation analysis aimed at elucidating the interrelationship between energy generation and each meteorological variable. Employing Pearson's Product-Moment Correlation Coefficient (PPMCC) [6], [13], denoted as 'r' and represented in (16), serves as a quantitative measure to ascertain these associations.

$$r = \frac{\sum_{i=1}^{N}((E_{PV,i} \cdot (M_i - \overline{M}))}{\sqrt{\sum_{i=1}^{N}(E_{PV,i} - \overline{E_{PV}})^2}\sqrt{\sum_{i=1}^{N}(M_i - \overline{M})^2)}} \tag{16}$$

where, N is the number of observations, i is the index, E_{PV} is the energy generated by the PV panel, M is one of the meteorological parameters, $\overline{E_{PV}}$ and $\overline{M}$ are the mean of E_{PV} and M respectively.

F. DATA DIVISION AND NORMALIZATION

The available dataset necessitates partitioning into subsets tailored for both the training and testing phases. Approximately 80% of the data spanning from January 1st, 2016, to December 31st, 2019 is allocated for the training of the ML models, whereas the residual 20% corresponding to the entirety of the year 2020 is reserved exclusively for model testing purposes. This partitioning of the data is visually depicted in Fig. 4 for clarity. Notably, the dataset comprises numerous meteorological variables characterized by disparate scales of measurement. To address this incongruity, Table 3 comprehensively encapsulates these variables alongside their respective ranges and units of measurement. It is imperative to rectify this incongruence through a process of data normalization, as prediction conducted amidst such disparity inevitably culminates in erroneous outcomes. Thus,

TABLE 3. Meteorological variables with range and units.

Meteorological Variable	Variable Range	Units
Ozone	0.221 - 0.293	N/A
Solar Zenith Angle	4.05 - 178.43	Degree
Precipitable Water	1 - 7.8	cm
Temperature	19.5 - 39.3	°C
Dew Point	12.4 - 28.5	°C
DHI	0 - 517	W/m^2
DNI	0 - 974	W/m^2
GHI	0 - 1057	W/m^2
Relative Humidity	30.86 - 100	%
Surface Albedo	0.12 - 0.19	N/A
Pressure	994 - 1018	mbar
Wind Direction	0 - 360	Degree
Wind Speed	0.2 - 11.3	m/s

the input data undergoes a pivotal stage of normalization to rectify this disparity and ensure the integrity of subsequent analyses.

FIGURE 4. Division of data.

Min-max method [24] is employed and the dataset is converted to [0, 1] range. Data denormalization is also attained using min-max method. The min-max equation is as in (17).

$$X_{norm} = (X - X_{min})/(X_{max} - X_{min}) \tag{17}$$

where, X_{norm} is the normalized value of the input variable, X is the original value of the input variable, X_{min} and X_{max} are minimum and maximum values of the corresponding input variable.

G. DEPLOYING ML MODELS

In the realm of Artificial Intelligence and Data Science, ML models play a pivotal role in extracting meaningful insights, making predictions, and driving decision-making processes across various domains. Among the myriad of ML algorithms, the ones that are relevant and fit the use case for energy predictions are SVM, ET, GPR and NN. These models were implemented on MATLAB version R2023b in a laptop with AMD Ryzen 3 7320U with Radeon Graphics processor and 8GB of DDR4 memory running on Windows 11 Pro 64-bit operating system.

1) SVM

SVM represents a category of supervised learning algorithms applied for regression and classification tasks. At its core, SVM aims to find the optimal hyperplane that best separates data points into different classes or predicts continuous outcomes. By maximizing the margin between classes or fitting the data with minimal error, SVM offers robustness against overfitting and generalizes well to unseen data. SVM handles linear and nonlinear relationships using kernel functions, making it versatile for various applications such as image classification, text categorization, energy prediction and financial forecasting.

2) ET

Ensemble learning techniques particularly ensembles of decision trees have gained prominence in the ML community due to their ability to improve predictive performance and reduce model variance. Bagging (Bootstrap Aggregating) and Boosting are two popular ensemble methods that leverage multiple decision trees to make collective predictions.

- Bagging as seen in RFs trains multiple decision trees on bootstrapped subclasses of the data and combines their forecasts through averaging or voting.
- Boosting exemplified by algorithms like Gradient Boosting Machines, sequentially trains weak learners to correct the errors of the previous models, resulting in a strong ensemble with superior predictive accuracy.

3) GPR

GPR offers a probabilistic approach to regression tasks, allowing for flexible modelling of complex relationships between inputs and outputs. Unlike traditional regression methods that assume a specific functional form, GPR models the entire distribution of possible functions, making it suitable for non-parametric regression and uncertainty estimation. GPR employs Gaussian processes to define a prior distribution over functions and updates this distribution based on observed data to obtain posterior predictions. This approach enables GPR to capture intricate patterns, handle noisy data, and provide valuable insights into prediction uncertainty.

4) NN

NN represent a classification of deep learning models motivated by the human brain's functioning and structure. These models consist of interconnected layers of artificial neurons, each performing simple computations and transmitting signals to subsequent layers. Through a process known as forward propagation and backpropagation, NNs learn to map input data to output predictions by adjusting the weights and biases of the connections between neurons.

H. ERROR METRICS

1) ROOT MEAN SQUARE ERROR (RMSE)

RMSE is an extensively used error metric for gauging the accuracy of forecast models, including those used in energy prediction for PV panels. RMSE as in (18), measures the square root of the average of the squared differences between predicted and observed values. It delivers a measure of the average magnitude of errors in the predictions with smaller RMSE values signifying superior model performance. RMSE is particularly useful when larger errors are penalized more heavily, providing a comprehensive assessment of model accuracy across the entire dataset.

$$RMSE = \sqrt{(1/N)\sum_{i=1}^{N}(Y_i - \widehat{Y})^2} \tag{18}$$

where, Y_i is the calculated value and $\widehat{Y}$ is the predicted value.

2) MEAN ABSOLUTE ERROR (MAE)

MAE is another commonly utilized error metric for assessing the performance of predictive models. Unlike RMSE, MAE quantifies the average absolute difference between predicted and observed values without squaring the errors as in (19). This metric provides a more straightforward interpretation of model accuracy representing the average magnitude of errors in the predictions. MAE is robust to outliers and gives equal weight to all errors, making it suitable for scenarios where the impact of large errors needs to be minimized. However, MAE does not penalize large errors as heavily as RMSE, potentially underestimating the influence of outliers on model performance.

$$MAE = (1/N)\sum_{i=1}^{N}|Y_i - \widehat{Y}| \tag{19}$$

3) COEFFICIENT OF DETERMINATION (R^2)

The R^2 coefficient, also known as the coefficient of determination, in (20), quantifies the proportion of variance in the dependent variable (i.e., the energy generated by PV panels) that is explained by the independent variables (i.e., the predictive features). R^2 ranges from 0 to 1 providing a measure of how well the model captures the inconsistency in the observed data, with values nearer to 1 representing a compelling bond between the predictors and the target variable. However, it does not provide information about the absolute magnitude of prediction errors and can be influenced by the number of predictors in the model, necessitating careful interpretation in conjunction with other error metrics like RMSE and MAE.

$$R^2 = 1 - \frac{\sum(Y_i - \widehat{Y})^2}{\sum(Y_i - \overline{Y})^2} \tag{20}$$

III. RESULTS AND DISCUSSION

A scatter plot matrix, also known as a pair plot, is a visualization technique used to explore the relationships between multiple variables in a dataset. It consists of a grid of scatter plots, where each plot in the matrix shows the relationship between two variables, with one variable upon the x-axis and the other upon the y-axis. Along the diagonal of the grid, each cell typically displays a histogram or kernel density estimate of the corresponding variable, allowing for the examination of individual variable distributions. The matrix format offers the advantage of concurrent comparison among all pairs of variables, facilitating the discernment of intricate patterns, correlations, and aberrations dispersed throughout the dataset. Particularly, scatter plot matrices emerge as invaluable tools for identifying the complex multivariate interdependencies among variables, thereby affording insights into potential associations or trends permeating the dataset. Depicted in Fig. 5 is a 15 × 15 scatter plot matrix, demonstrating the interplay among 15 distinct features. Each grid within the plot encapsulates a bivariate distribution pertaining to a specific pair of features. Within the upper right quadrant of the plot lies a magnified depiction, spotlighting the intricate distributions of temperature, GHI, EPV and G_γ.

Utilizing (16), the interplay between weather-related parameters and E_{PV} is meticulously scrutinized. Upon applying correlation analysis, it emerges that several parameters exhibit robust linear relationships, whereas others demonstrate varying degrees of moderate to weak linear correlations. Furthermore, certain parameters appear devoid of any discernible correlation. Encapsulating the PPMCC for each weather-related parameter in relation to E_{PV}, Table 4 delineates these correlations along with their corresponding interpretations. Noteworthy among these parameters earmarked for further investigation are the SZA, T, DHI, DNI, GHI, RH, v_w, G_γ, alongside E_{PV}.

The dataset comprising the curated selection of weather-related parameters undergoes rigorous training and testing utilizing the Regression Learner app within the MATLAB environment. The training phase entails a meticulous hyperparameter tuning process, meticulously refining these parameters iteratively until an optimal model configuration is attained. Subsequent scrutiny is then undertaken to evaluate the efficacy of these refined models. Enumerated within Table 5 are the various ML models alongside their

TABLE 4. PPMCC of the weather-related parameters with E_{PV}.

Weather-related Parameters	r value	Interpretation
O	0.1514	Weak positive linear correlation
SZA	-0.6921	Moderate negative linear correlation
PW	-0.1302	Weak negative linear correlation
T	0.3667	Weak positive linear correlation
DP	-0.0871	No correlation
DHI	0.7282	Moderate positive linear correlation
DNI	0.5753	Moderate positive linear correlation
GHI	0.8668	Strong positive linear correlation
RH	-0.3175	Weak negative linear correlation
ρ	-0.0371	No correlation
P	-0.0180	No correlation
WD	-0.0182	No correlation
v_w	0.1472	Weak positive linear correlation
G_γ	0.9570	Strong positive linear correlation

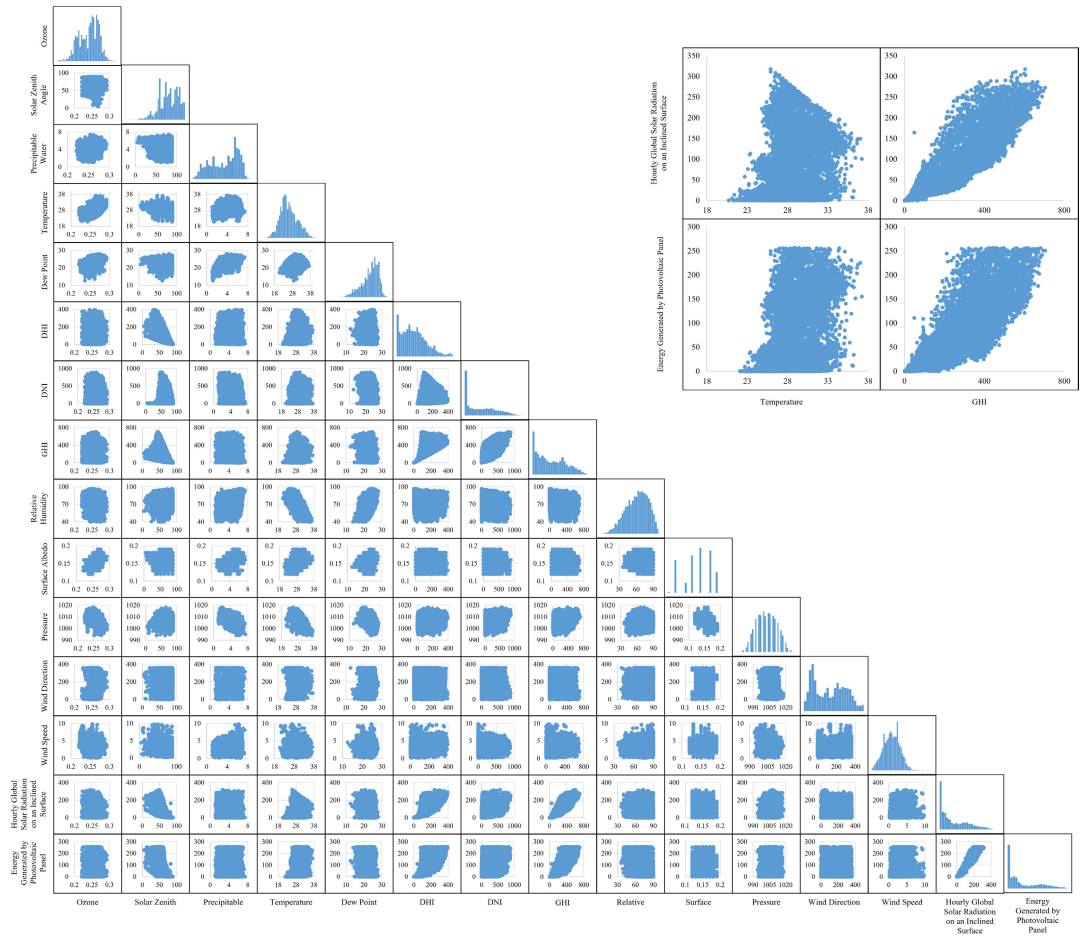

FIGURE 5. Scatter plot matrix between all the data-points.

corresponding hyperparameters earmarked for optimization. Furthermore, the table delineates the ranges within which these hyperparameters are explored and ultimately highlights the best-optimized hyperparameters for each ML model.

The procedural steps entailed during both the training and testing phases are expounded below:

- Step 1: Partition the dataset into distinct training and testing sets.
- Step 2: Establish a validation scheme predicated upon 5-fold cross-validation.
- Step 3: Elect Bayesian Optimization as the preferred optimizer for hyperparameter tuning.
- Step 4: Enumerate the hyperparameters slated for optimization.
- Step 5: Consecutively tune all designated hyperparameters for each model.
- Step 6: Assume the existence of 'n' models within the ensemble.
- Step 7: Each model undergoes subdivision into i = 5 iterations.
- Step 8: Each iteration further subdivides into k = 5 folds for cross-validation purposes.
- Step 9: Reserve a fold within every iteration as the cross-validation test set.
- Step 10: For iterations 'i' ranging from 1 to 5, compute the respective errors e_1 to e_5.
- Step 11: Calculate the average errors 'A_e' for models 1 to n, respectively.
- Step 12: Amongst the ensemble, identify the model with the lowest RMSE during the validation phase.
- Step 13: Document the optimized hyperparameters pertaining to this distinguished model.
- Step 14: Apply the optimized hyperparameters across all models during the training phase.
- Step 15: Subject the models to rigorous testing.
- Step 16: Iteratively repeat steps 4 through 15 for each ML model.
- Step 17: The model exhibiting the lowest RMSE during the testing phase emerges as the optimal ML model for the prognostication of E_{PV} utilizing the designated dataset.

The forecast outcomes undergo meticulous scrutiny through a tripartite lens. Initially, the annual forecast accuracy is rigorously assessed across all ML models. Subsequently, a granular examination is conducted on a month-by-month

TABLE 5. Bayesian optimization hyperparameter search range.

Model	Hyperparameters	Hyperparameter Search Range	Optimized Hyperparameters
SVM (M1)	Kernel function	Gaussian, Linear, Quadratic, Cubic	Cubic
	Box constraint	0.001 – 1000	0.13311
	Kernel scale	0.001 – 1000	1
	Epsilon	0.00038217 – 38.2173	0.0010993
	Standardize data	True, false	Yes
ET (M2)	Ensemble method	Bag, LSBoost	LSBoost
	Minimum leaf size	1 – 388	3
	Number of learners	10 – 500	114
	Learning rate	0.001 – 1	0.15668
	Number of predictors to sample	1 – 9	9
GPR (M3)	Basis function	Constant, Zero, Linear	Linear
	Kernel function	Nonisotropic Exponential, Nonisotropic Matern 3/2, Nonisotropic Matern 5/2, Nonisotropic Rational Quadratic, Nonisotropic Squared Exponential, Isotropic Exponential, Isotropic Matern 3/2, Isotropic Matern 5/2, Isotropic Rational Quadratic, Isotropic Squared Exponential	Nonisotropic Rational Quadratic
	Kernel scale	0.001 – 1000	0.36723
	Signal standard deviation		0.18845
	Sigma	0.0001 – 2.665	2.4114
	Standardize data	True, false	No
	Optimize numeric parameters		Yes
NN (M4)	Number of fully connected layers	1 - 3	2
	First layer	1 - 300	107
	Second Layer	1 - 300	59
	Third Layer	1 - 300	
	Activation	ReLU, Tanh, Sigmoid, None	ReLU
	Iteration limit		1000
	Regularization strength (γ)	1.287e-08 – 128.7001	3.2604e-05
	Standardize data	Yes, No	Yes

basis to discern the nuanced impact of individual months on the forecast accuracy of each ML model. Lastly, the predictive precision across diverse seasons is meticulously evaluated vis-Ã -vis the performance of the ML models.

A. ANNUAL FORECAST ASSESSMENT

The annual forecasting is conducted utilizing the refined methodologies of the four optimized ML models. The empirical findings regarding the power output generated by the PV panel are compiled within Table 6, with particular emphasis placed on highlighting the most proficient model denoted in bold. Notably, M2 emerges as the top performer amongst the quartet. This is substantiated by its MAE of 0.0310, representing a noteworthy reduction of 56% compared to M1, 43.21% in relation to M3, and a decrease of 27.16% vis-Ã -vis M4. Furthermore, the RMSE of M2, standing at 0.0454, exhibits a significant diminution of 43.48%, 41.74%, and 26.39% compared to M1, M3, and M4, respectively. The R^2 for M2 attains a value of 0.9740, thereby showcasing an improvement of 3.87%, 3.63%, and 1.89% when juxtaposed against M1, M3, and M4, respectively.

Table 7 delineates a comparative analysis of the training durations required for the annual forecast. M2 garners distinction for its minimal training time, clocking in at 17.212 seconds. Conversely, the training durations of the remaining models conspicuously surpass that of M2.

TABLE 6. Comparison of error metrics of test set for annual data.

Model	MAE	RMSE	R^2
M1	0.0552	0.0707	0.9370
M2	**0.0310**	**0.0454**	**0.9740**
M3	0.0482	0.0694	0.9393
M4	0.0408	0.0592	0.9557

TABLE 7. Comparison of training time for the annual data.

Model	Training Time (s)
M1	602.79
M2	**17.212**
M3	4182.7
M4	423.5

Moreover, Fig. 6 offers a visual representation of the prediction curve for E_{PV} spanning four consecutive days in December. Herein, M2 exhibits remarkable fidelity to the base curve, particularly evident on the 2nd and 3rd of December 2020. This fidelity is accentuated by the observable discrepancy in the GHI on December 2nd, 2020, which consequently yields a commensurate reduction in E_{PV} output.

B. MONTHLY FORECAST ASSESSMENT

Based on the annual performance metrics, the forecast undergoes analysis through a granular dissection of the dataset into monthly segments spanning the calendar year.

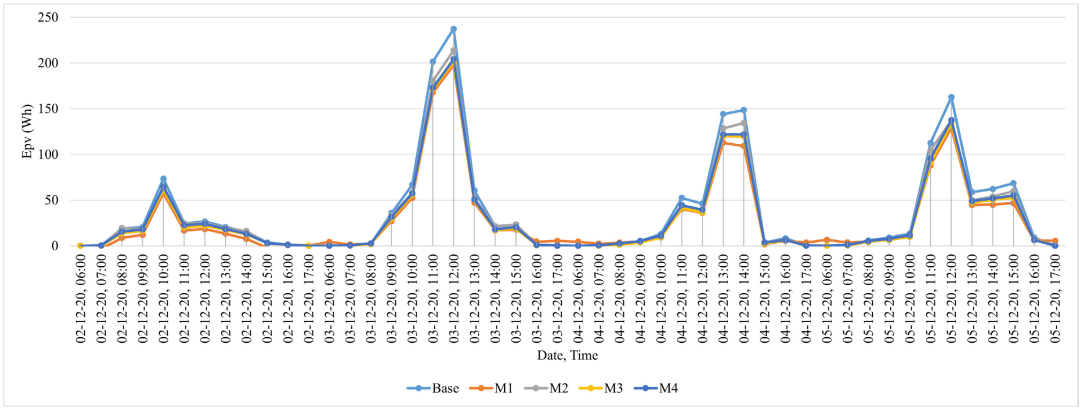

FIGURE 6. Prediction curve for 4 consecutive days of december.

This comprehensive approach entails the training and testing of the dataset across all four optimized ML models from January through December. Subsequently, an exhaustive investigation ensues, with the performance of each model scrutinized and tabulated for comparative analysis.

Table 8 presents a comprehensive overview, contrasting the error metrics derived from the monthly test sets across each model, whereas Table 9 provides a comparative analysis of the training durations required by the models for each month. Emphasizing the preeminent results, the most favourable predictions are underscored in bold for clarity.

Table 8 unveils a discernible pattern wherein MAE of M2 consistently outperforms its counterparts across eleven out of the twelve months, showcasing its remarkable predictive efficacy. However, it's noteworthy that MAE of M4 achieves a marginal superiority over M2, boasting a 17.033% reduction in MAE. Similarly, while M2 secures the lowest RMSE in ten of the twelve months, M4 exhibits a slight edge of 1.08% and 26.6% during May and December, respectively. Furthermore, M2 garners the highest R^2 in ten months, with May and December displaying marginal improvements of 0.05% and 4.64%, respectively.

Evidently, the comprehensive appraisal of the monthly data unequivocally positions M2 as the frontrunner in terms of predictive accuracy. Moreover, Table 9 corroborates these findings by highlighting the markedly shorter training durations associated with M2 relative to its counterparts.

In addition to tabular representations, Figs. 7, 8, and 9 offer insightful radar charts elucidating the monthly profiles of RMSE, MAE, and R^2, respectively. These radar charts, also known as spider charts or star plots, afford a graphical means to depict multivariate data within a two-dimensional space. Each variable is represented by a distinct axis radiating outward from the chart's centre, with data points plotted along these axes to form a polygon or area that encapsulates the dataset's profile. Such visualizations are invaluable tools for comparative analysis across multiple criteria, facilitating a comprehensive assessment of performance or characteristics across different entities.

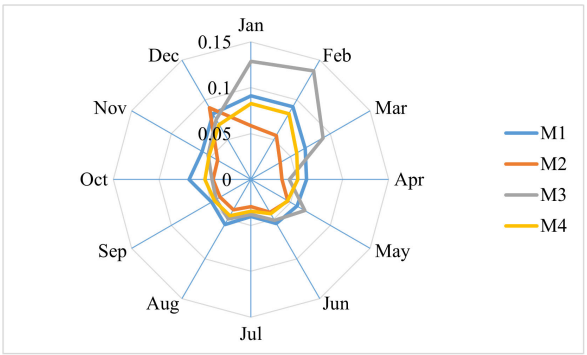

FIGURE 7. Radar chart representing monthly RMSE for test set.

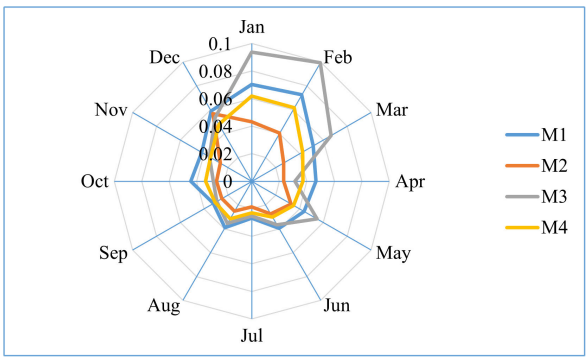

FIGURE 8. Radar chart representing monthly MAE for test set.

C. SEASONAL FORECAST ASSESSMENT

Drawing upon both annual and monthly performance analyses, a comprehensive exploration delves into the seasonal ramifications on the model's efficacy. Leveraging the Köppen climate classification system [25], Chennai is categorized as harbouring a tropical savanna climate, characterized by dry winters within the 'Aw' category. Table 10 shows the seasonal classification of Chennai, providing invaluable insight into the climatic nuances shaping the dataset. Complementing this analysis, Fig. 10 offers a visually striking depiction through a box plot showcasing E_{PV} across diverse seasons. The positively skewed distribution evident in the plot underscores

TABLE 8. Comparison of error metrics of test set month-wise.

	M1			M2			M3			M4		
	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2
Jan	0.0702	0.0911	0.9021	**0.0432**	**0.0584**	**0.9598**	0.0939	0.1286	0.8051	0.0619	0.0827	0.9194
Feb	0.0725	0.0914	0.9168	**0.0403**	**0.0551**	**0.9698**	0.0995	0.1365	0.8146	0.0616	0.0829	0.9316
Mar	0.0515	0.0683	0.9185	**0.0268**	**0.0368**	**0.9764**	0.0665	0.0906	0.8567	0.0427	0.0577	0.9419
Apr	0.0467	0.0605	0.9541	**0.0231**	**0.0337**	**0.9857**	0.0313	0.0420	0.9779	0.0371	0.0510	0.9674
May	0.0439	0.0580	0.9618	**0.0330**	0.0466	0.9753	0.0548	0.0680	0.9475	0.0350	**0.0461**	**0.9758**
Jun	0.0395	0.0556	0.9567	**0.0273**	**0.0414**	**0.9760**	0.0365	0.0513	0.9632	0.0299	0.0431	0.9740
Jul	0.0268	0.0402	0.9746	**0.0187**	**0.0296**	**0.9862**	0.0253	0.0379	0.9775	0.0230	0.0346	0.9813
Aug	0.0389	0.0568	0.9551	**0.0250**	**0.0381**	**0.9798**	0.0353	0.0498	0.9656	0.0315	0.0456	0.9711
Sep	0.0324	0.0492	0.9670	**0.0251**	**0.0387**	**0.9796**	0.0294	0.0435	0.9742	0.0308	0.0459	0.9713
Oct	0.0443	0.0673	0.9478	**0.0256**	**0.0414**	**0.9802**	0.0281	0.0424	0.9793	0.0336	0.0500	0.9712
Nov	0.0419	0.0615	0.9526	**0.0264**	**0.0419**	**0.9780**	0.0357	0.0538	0.9638	0.0346	0.0526	0.9653
Dec	0.0590	0.0827	0.9140	0.0567	0.0903	0.8973	0.0534	0.0758	0.9277	**0.0478**	**0.0691**	**0.9399**

TABLE 9. Comparison of training times for the monthly data.

	M1	M2	M3	M4
Jan	73.371	**15.317**	183.16	30.939
Feb	54.242	**10.507**	102.72	26.392
Mar	54.042	**9.5524**	93.688	27.108
Apr	52.634	**10.356**	94.692	27.303
May	55.215	**10.323**	107.83	27.991
Jun	59.811	**10.375**	131.18	27.994
Jul	66.584	**11.446**	206.69	35.937
Aug	60.609	**10.75**	143.94	32.224
Sept	61.371	**11.804**	153	33.014
Oct	72.257	**11.971**	230.54	43.659
Nov	76.454	**11.805**	265.42	47.378
Dec	73.042	**11.086**	265.24	40.977

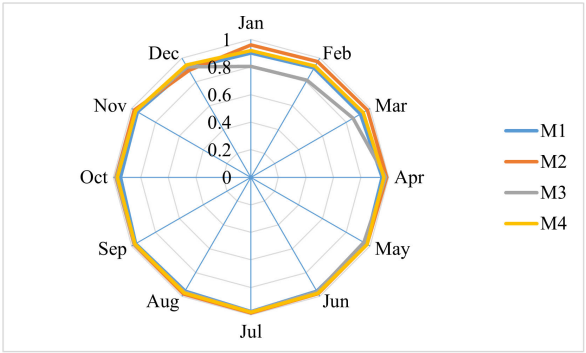

FIGURE 9. Radar chart representing monthly R^2 for test set.

TABLE 10. Seasons of Chennai grouped by months.

Seasons	Months of the year
Winter	Dec, Jan, Feb
Summer	March to June
Rainy	July to Nov

an inherent asymmetry within the dataset, further enriching our understanding of the seasonal dynamics at play.

The scrutiny of seasonal models precipitates a comprehensive assessment, with results compiled and shown within Table 11. This tabulation facilitates a comparative analysis of error metrics derived from seasonal test set data. Notably, M2 exhibits exceptional performance across all metrics, as exemplified by its MAE values of 0.0359, 0.0307, and 0.0261 for the winter, summer, and rainy seasons, respectively. Correspondingly, RMSE values of 0.0520, 0.0429, and 0.0398 underscore its consistent efficacy

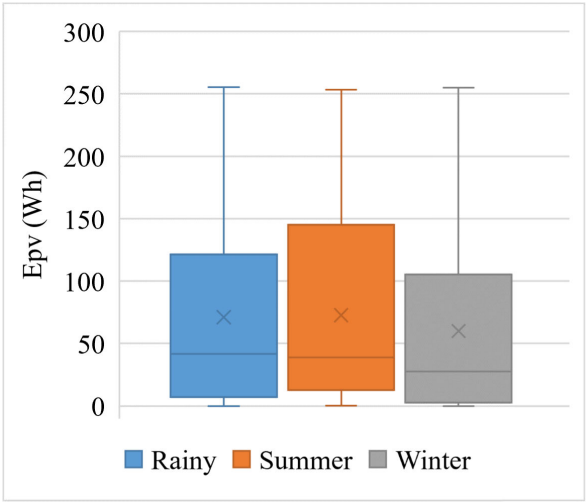

FIGURE 10. Seasonal distribution of E_{PV} from 2016 to 2020.

across these seasons. Moreover, R^2 attains impressive levels, standing at 96.91%, 97.56%, and 97.91% for winter, summer, and rainy seasons, respectively.

Substantiating the superlative performance of M2, Table 12 provides a comparative analysis of the training durations requisite for the seasonal test sets. M2 emerges as the frontrunner yet again, boasting markedly shorter training times relative to its counterparts. Noteworthy among these figures are the training durations of 13.987 seconds, 6.2223 seconds, and 8.9362 seconds for the winter, summer, and rainy seasons, respectively.

Fig. 11 offers a visual depiction, illustrating the prediction curve for a rainy day in October. Notably, the presence of clouds precipitates a gradual decline in irradiation levels commencing from 10:00 AM, consequently diminishing E_{PV} output. The nadir of E_{PV} generation is observed at approximately 11:00 AM, after which the gradual dispersal of clouds precipitates a resurgence in irradiation levels, thereby augmenting E_{PV} output.

From the above analysis it is evident that the dataset was captured hourly for three cases: annual, seasonal and monthly. The SVM, ET, GPR and NN algorithms were deployed on the above three distinct datasets which provides

TABLE 11. Comparison of error metrics of test set season-wise.

Model	Metrics	Winter	Summer	Rainy
M1	**MAE**	0.0715	0.0491	0.0429
	RMSE	0.0922	0.0635	0.0618
	R^2	0.9031	0.9466	0.9494
M2	**MAE**	**0.0359**	**0.0307**	**0.0261**
	RMSE	**0.0520**	**0.0429**	**0.0398**
	R^2	**0.9691**	**0.9756**	**0.9791**
M3	**MAE**	0.0735	0.0414	0.0332
	RMSE	0.1013	0.0554	0.0505
	R^2	0.8829	0.9594	0.9663
M4	**MAE**	0.0585	0.0362	0.0286
	RMSE	0.0796	0.0495	0.0443
	R^2	0.9277	0.9676	0.9740

TABLE 12. Comparison of training times for the season-wise data.

	M1	M2	M3	M4
Winter	207.08	**13.987**	2804.2	98.369
Summer	231.98	**6.2223**	3857.9	116.73
Rainy	303.24	**8.9362**	2561.2	170.3

appreciable results while forecasting the power in stand-alone photovoltaic systems. The granularity of the dataset if increased may lead to higher computational time.

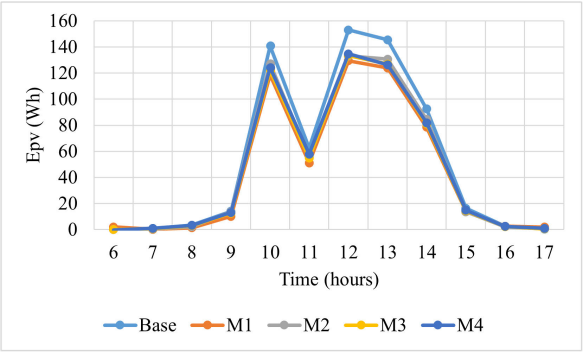

FIGURE 11. Seasonal distribution of E_{PV} from 2016 to 2020.

IV. CONCLUSION

The study unfolds through the collection of weather-related parameters by pinpointing the geographical coordinates of VITCC from the NSRDB webpage. Dataset for five consecutive years, 2016 to 2020, was downloaded for hourly intervals. Subsequently computation of G_γ, alongside the determination of EPV is carried out. Power forecast for three distinct dataset groupings is led namely: annual, monthly and seasonal. The study unfolds through the collection of weather-related parameters and the subsequent computation of G_γ, alongside the determination of E_{PV}. The forecasting of E_{PV} is orchestrated through the strategic deployment of four optimized ML models: SVM denoted as M1, ET as M2, GPR as M3 and NN as M4. This optimization journey is facilitated through the employment of Bayesian Optimization meticulously fine-tuning the ML models by calibrating their hyperparameters to achieve optimal performance. The key focus from the work done are as follows:

- The ensuing analysis is conducted across three distinct categories: annual, monthly, and seasonal, offering a comprehensive panorama of performance metrics across various temporal scales based on hourly estimates of E_{PV}.
- Each model is subjected to rigorous scrutiny with M2 emerging as the epitome of forecasting prowess across all categories.
- Evidentiary support from the annual dataset category highlights a substantial reduction in MAE of M2 by 56%, 43.21%, and 27.16% relative to Models 1, 3, and 4, respectively.
- Likewise, for RMSE, M2 exhibits a reduction of 43.48%, 41.74%, and 26.39% when compared to Models 1, 3, and 4, respectively.
- Furthermore, R^2 manifests noteworthy enhancements of 3.87%, 3.63%, and 1.89% over Models 1, 3, and 4, respectively.
- A nuanced examination of the monthly dataset category underscores the pre-eminence of M2, which delivers the most accurate predictions for 10 out of 12 months of the year.
- Similarly, across seasonal datasets, M2 consistently garners superlative results in all seasons.
- This resounding superiority of the optimized ET is further accentuated by its expedited training time, substantiating its efficacy as the model of choice across all categories.

V. FUTURE SCOPE OF WORK

The examination conducted in this research paper was carried out by pinpointing the geographical coordinates of VITCC which is located in the city of Chennai. Chennai is categorized as harbouring a tropical savanna climate, characterized by dry winters within the 'Aw' category. Equivalent investigation can be carried out by selecting diverse sites falling under wide-ranging climatic category. The study shows that there is a huge scope of work in this particular area. As the world progresses the demand for energy also rises. Over the course of the next couple of decades, the demand for energy will rise exponentially, as the speed of growth of developing nations accelerates. India is a prime example of the above phenomenon. With India's ambitious renewable energy targets and increasing investments in wind, solar and other non-conventional sources, there is a pressing need for accurate prediction of renewable energy generation. Innovative prediction techniques, together with ML algorithms and meteorological models, can enhance the reliability and efficiency of renewable energy integration into the grid, enabling better planning and management of renewable energy resources. The deployment of smart grid technologies, including smart meters, sensors, and advanced monitoring systems, presents new opportunities for power prediction and management. These technologies enable real-time data collection, analysis, and communication, facilitating more accurate and responsive power prediction models. By leveraging data analytics and ML algorithms, smart grid systems can increase energy distribution, enhance grid

resilience, and support the integration of distributed energy resources.

REFERENCES

[1] *India Could Become the World's 3rd Largest Economy in the Next 5 Years. Here's How | World Economic Forum*. Accessed: Apr. 22, 2024. [Online]. Available: https://www.weforum.org/agenda/2024/01/how-india-can-seize-its-moment-to-become-the-world-s-third-largest-economy/

[2] *Pradhan Mantri Sahaj Bijli Har Ghar Yojana—Saubhagya| National Portal of India*. Accessed: Apr. 22, 2024. [Online]. Available: https://www.india.gov.in/spotlight/pradhan-mantri-sahaj-bijli-har-ghar-yojana-saubhagya

[3] *Rural Electrification in India: Problems, Progress, and the Power of Citizen's Media—AIF*. Accessed: Apr. 23, 2024. [Online]. Available: https://aif.org/rural-electrification-in-india-problems-progress-and-the-power-of-citizens-media/

[4] *Sustainable Development Goals | United Nations Development Programme*. Accessed: Apr. 23, 2024. [Online]. Available: https://www.undp.org/sustainable-development-goals

[5] *1 Crore Households To Get Rooftop Solar Under Pradhan Mantri Suryodaya Yojana | Prime Minister of India*. Accessed: Apr. 23, 2024. [Online]. Available: https://www.pmindia.gov.in/en/news-updates/1-crore-households-to-get-rooftop-solar-under-pradhan-mantri-suryodaya-yojana/

[6] D. Markovics and M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," *Renew. Sustain. Energy Rev.*, vol. 161, Jun. 2022, Art. no. 112364, doi: 10.1016/j.rser.2022.112364.

[7] 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, doi: 10.1109/ACCESS.2023.3270041.

[8] M. Madhukumar, A. Sebastian, X. Liang, M. Jamil, and M. N. S. K. Shabbir, "Regression model-based short-term load forecasting for university campus load," *IEEE Access*, vol. 10, pp. 8891–8905, 2022, doi: 10.1109/ACCESS.2022.3144206.

[9] A. Patel, O. V. G. Swathika, U. Subramaniam, T. S. Babu, A. Tripathi, S. Nag, A. Karthick, and M. Muhibbullah, "A practical approach for predicting power in a small-scale off-grid photovoltaic system using machine learning algorithms," *Int. J. Photoenergy*, vol. 2022, pp. 1–21, Feb. 2022, doi: 10.1155/2022/9194537.

[10] M. J. Mayer, "Benefits of physical and machine learning hybridization for photovoltaic power forecasting," *Renew. Sustain. Energy Rev.*, vol. 168, Oct. 2022, Art. no. 112772, doi: 10.1016/j.rser.2022.112772.

[11] N. Omar, H. Aly, and T. Little, "Seasonal clustering forecasting technique for intelligent hourly solar irradiance systems," *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 2520–2529, Mar. 2023, doi: 10.1109/TII.2022.3177746.

[12] N. Elizabeth Michael, S. Hasan, A. Al-Durra, and M. Mishra, "Short-term solar irradiance forecasting based on a novel Bayesian optimized deep long short-term memory neural network," *Appl. Energy*, vol. 324, Oct. 2022, Art. no. 119727, doi: 10.1016/j.apenergy.2022.119727.

[13] L. Gutiérrez, J. Patiño, and E. Duque-Grisales, "A comparison of the performance of supervised learning algorithms for solar power prediction," *Energies*, vol. 14, no. 15, p. 4424, Jul. 2021, doi: 10.3390/en14154424.

[14] D. Chakraborty, J. Mondal, H. B. Barua, and A. Bhattacharjee, "Computational solar energy–ensemble learning methods for prediction of solar power generation based on meteorological parameters in eastern India," *Renew. Energy Focus*, vol. 44, pp. 277–294, Mar. 2023, doi: 10.1016/j.ref.2023.01.006.

[15] M. Mattei, G. Notton, C. Cristofari, M. Muselli, and P. Poggi, "Calculation of the polycrystalline PV module temperature using a simple method of energy balance," *Renew. Energy*, vol. 31, no. 4, pp. 553–567, Apr. 2006, doi: 10.1016/j.renene.2005.03.010.

[16] Ö. Ayvazoğluyüksel and Ü. B. Filik, "Estimation methods of global solar radiation, cell temperature and solar power forecasting: A review and case study in Eskisehir," *Renew. Sustain. Energy Rev.*, vol. 91, pp. 639–653, Aug. 2018, doi: 10.1016/j.rser.2018.03.084.

[17] *NSRDB*. Accessed: Apr. 23, 2024. [Online]. Available: https://nsrdb.nrel.gov/data-viewer

[18] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (NSRDB)," *Renew. Sustain. Energy Rev.*, vol. 89, pp. 51–60, Jun. 2018, doi: 10.1016/j.rser.2018.03.003.

[19] *TS250 Series 60-cell Multi-Crystalline Solar Photovoltaic Modules*. Accessed: Nov. 30, 2023. [Online]. Available: https://www.tatapowersolar.com

[20] F. J. Olmo, J. Vida, I. Foyo, Y. Castro-Diez, and L. Alados-Arboledas, "Prediction of global irradiance on inclined surfaces from horizontal global irradiance," *Energy*, vol. 24, no. 8, pp. 689–704, 1999. [Online]. Available: https://www.elsevier.com/locate/energy

[21] J. A. Duffie and W. A. Beckman, *Solar Engineering of Thermal Processes*. Hoboken, NJ, USA: Wiley, 2013, pp. 3–42.

[22] M. E. Ropp, M. Begovic, and A. Rohatgi, "Determination of the curvature derating factor for the Georgia tech aquatic center photovoltaic array," in *Proc. Photovoltaic Specialists Conf.*, 1997, pp. 1297–1300.

[23] J. V. Paatero and P. D. Lund, "Effects of large-scale photovoltaic power integration on electricity distribution networks," *Renew. Energy*, vol. 32, no. 2, pp. 216–234, Feb. 2007, doi: 10.1016/j.renene.2006.01.005.

[24] Y. Yu, G. Hu, C. Liu, J. Xiong, and Z. Wu, "Prediction of solar irradiance one hour ahead based on quantum long short-term memory network," *IEEE Trans. Quantum Eng.*, vol. 4, pp. 1–15, 2023, doi: 10.1109/TQE.2023.3271362.

[25] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, "World map of the Köppen-geiger climate classification updated," *Meteorologische Zeitschrift*, vol. 15, no. 3, pp. 259–263, Jul. 2006, doi: 10.1127/0941-2948/2006/0130.

AADYASHA PATEL received the B.E. degree in electrical and electronics engineering from Rajalakshmi Engineering College, Chennai, India, in 2010, and the M.Tech. degree in power electronics and drives from Hindustan Institute of Technology and Science, Chennai, in 2013. She is currently pursuing the Ph.D. degree in electrical engineering with Vellore Institute of Technology, Chennai.

O. V. GNANA SWATHIKA (Senior Member, IEEE) received the B.E. degree in electrical and electronics engineering from Madras University, Chennai, Tamil Nadu, India, in 2000, the M.S. degree in electrical engineering from Wayne State University, Detroit, MI, USA, in 2004, and the Ph.D. degree in electrical engineering from Vellore Institute of Technology, Chennai, in 2017. She was a Postdoctoral Researcher with the University of Moratuwa, Sri Lanka, in 2019. Her current research interests include microgrid protection, power system optimization, embedded systems, and photovoltaic systems.