Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 8 – No. 1, August 2025 – www.caeaccess.org

Machine Learning Models, Data Preprocessing Techniques and Suite of Metrics for Assessing Solar Power Forecasting: A Comprehensive Review

Asma
Department of EEE,
Sapthagiri College of Engineering,
Bengaluru

A.M. Nagaraja
Department of ECE,
Sapthagiri College of Engineering,
Bengaluru

ABSTRACT

As we know, the energy obtained from sun is intermittent in nature hence its generation is affected by rapidly changing weather conditions. Hence to balance the supply demand and to improve the accuracy and efficiency of the system, forecasting the solar energy highly necessitates. In view of this, different photovoltaic power forecasting techniques using machine learning models, different data pre-processing techniques and the evaluation metrices are discussed. Further to improve the system performance and efficiency quality data input which is pre-processed is highly required. This paper discusses the comprehensive review of different solar forecasting techniques along with traditional forecasting techniques and a comparison review of ML models, respective algorithms used and various techniques for preprocessing the data are presented. Further, different suite metrices for assessing the performance of solar power forecasting is also presented.

Keywords

Photovoltaic, Power forecasting, Data pre-processing, Machine learning, Metrics.

1. INTRODUCTION

The necessity of precise and trustworthy PV forecasting techniques arose from the growing integration of solar photovoltaic (PV) systems into contemporary energy infrastructures. Grid stability, energy market operations, and the best possible integration of renewable energy sources all depend on accurate forecasting. The most effective forecasting methods that is using Machine Learning models because of their capacity to represent the intricate, nonlinear relationships present in solar energy generation. Although a lot of research has been done on the creation and improvement in ML algorithms for photovoltaic power forecasting, data preprocessing plays an equally significant but frequently overlooked role. By converting noisy, unstructured, and incomplete data into an understandable format that can be analysed, data preprocessing serves as a link between the acquisition of raw data and model training.

In [1] demonstration of ANN16 model is done which shows ANN16 model yields the best MAE-Mean absolute Error, RMSE-Root Mean Square value, and coefficient of determination (R2) with values of 0.4693, 0.8816 W, and 0.9988, respectively. In [2] it is also been discussed that hybrid models that combines with ANN models with traditional linear regression, where RMSE, MAE, Mean Bias Error and correlation coefficient values of 2.74, 2.09, 0.01 and 0.932, respectively, performed better. A new ESDLS-SVR model efficiently manages seasonal influence by defining the decomposition data using the seasonal decomposition technique [3]. In [4], it is discussed that, a low-pass filter is constructed to model the annual cycle of the solar irradiation with power time series (i.e., the corresponding clear sky values) using a Fourier transformation because of smoothness, flexibility, and inherent periodicity. A predictive accuracy of roughly 90%, the ANN model in S1 currently yields the most accurate results for solar power. When predicting solar power separately at different time series, ANN shows the better performance compared to other traditional models [5]. The LSTM is more efficient and precise for the prediction process of power generation, and these are characterized by their generality and simplicity. In ideal weather, the former can predict the satisfactory power value at 15-minute intervals, while in non-ideal weather, it can precisely predict the power trend at 1-hour intervals [6]. Therefore, developing precise and reliable forecasting systems requires a methodical grasp of preprocessing techniques, which range from feature selection and time series handling to data transformation and cleaning [7]. It should to be noted that for the model training for forecasting, 70 and 30% of the data set are used for training and validation respectively [8]. Consolidating existing knowledge in this area, highlighting best practices, identifying common problems, and offering a framework for further study on data-driven solar forecasting techniques a detailed review of different methods used for solar power forecasting along with the different data preprocessing are presented. Further, to evaluate the performance of the solar power forecasting the review of different metrics are also discussed in the last part of this paper.

2. ML MODELS FOR SOLAR POWER FORECASTING

Over the years, numerous machine learning (ML)models are investigated and assessed for forecasting the solar energy. The forecasting capabilities, interpretability, data requirements, and complexity of these models vary. The various kinds of machine learning models for forecasting the solar power are summarized here in categories as follows in Table. 1

Table: 1 Summary of ML Models for Solar Power Forecasting

Sl. No	Author/Year	Type of Models	Benefits	Drawbacks	Application
1.	Mohammed Abuella & Badrul Choudhry, 2015 [9]	**Multiple Linear Regression Model**	Easy to understand and comprehend, Quick training, Better performance and is linearly separable.	In adequate results for nonlinear data Outlier-sensitive.	Short-term forecasting under clear skies; baseline forecasting.
2.	Valeriy V , Gavrishchaka, et. Al, 2001 [10]	**SVM** **Support Vector Machines**	In high-dimensional spaces, Effective Radial Basis Kernal-RBF performs better working with nonlinear data.	computationally demanding and calls for meticulous parameter adjustment. (e.g., kernel, C)	Using weather data to forecast the hours and days ahead.
3.	Neha Singh, Satyaranjan Jena, et al, 2022 [11]	**Decision Tree**	Simple to understand and picture, Managing nonlinear connections.	prone to overfitting, Poor performance in generalization on unseen data.	Rapid power output estimation in a span of weather conditions.
4.	Rathika Senthil kumar, P S Meeram et al, 2025 [12]	**Random Forest**	Better adaption capability of the algorithm, Strong against noise and overfitting, Addresses feature importance and missing values.	Slower prediction and training than single tree models, less comprehensible.	Accurate forecasting for the coming day based on historical power, irradiance, and weather data.
5.	Zhe Song, Fu Xiao, et al, 2025, [13]	**Gradient Boosting Machines (GBM)** • XGBoost • LightGBM • CatBoost	Efficient solar power management and dispatch	Needs more processing power than simpler models prone to overfitting if improperly adjusted.	Forecasting in the short and medium term using historical power and structured weather data.
6.	Nor Azuana Ramli, Nurul Haniz Azhan et al, 2019 [14]	**K-Nearest Neighbour (KNN)**	Better prediction results compared to other models, Easy to use and efficient with regional patterns	Sluggish when making predictions for big datasets, In high-dimensional spaces, poor performance.	Using historical data to forecast similar weather conditions.

7.	R. Asghar, F. R. Fulginei, et al, 2024 [15]	**Deep Learning Models and Neural Networks**	Can enhance PV power forecasting and improve grid stability by accessing data quality and model complexity and conducting rigorous validations. -Can learn complex patterns from large datasets - High accuracy in many tasks (e.g., vision, NLP) - Automates feature extraction - Scales well with data - Performs well on unstructured data	- Requires large amounts of labelled data - High computational cost (GPUs, TPUs) - Black-box nature (low interpretability) - Risk of overfitting - Requires expert tuning and architecture design	- **Natural Language Processing (NLP)**: Machine translation, chatbots, sentiment analysis - **Healthcare**: Disease detection, medical imaging diagnostics - **Finance**: Fraud detection, stock prediction - **Autonomous Vehicles**: Perception and decision-making - **Gaming and Robotics**: Strategy planning, motion control
8.	Nattha Thipwangmek , Nopparuj Suetrong et. Al, 2024 [16]	**Convolution Neural networks** **(CNN)**	-Robust model that excels in capturing complex patterns in solar PV generation data. -Excellent for image and spatial data processing - Raw data feature extraction automatically - Translation invariance in images - Works well with minimal pre-processing	- Requires large labelled datasets for training - Computationally intensive - Poor performance on non-spatial data - Susceptible to adversarial attacks - Limited understanding of internal workings	By extending the forecasting horizon, incorporating more granular weather data, assessing scalability and adaptability, and integrating emerging ML techniques can further advance the field of solar PV power generation forecasting.
9.	Meftah Elsaraiti and Adel Merab et al , 2022 [17]	**Recurrent** **Neural Networks (RNNs)**	More efficient operation of photovoltaic power plants in the future, promoting energy sustainability, decarburization, and the digitization of the electricity sector, Designed for sequential data (e.g., time series, text)	- Struggles with long-term dependencies (vanishing gradient problem) - Training can be slow and unstable - Limited parallelization due to sequential processing - Often outperformed by LSTM and Transformer models	- **Time Series Forecasting**: Stock prices, weather prediction - **Music Generation**: Creating sequences of musical notes - **Anomaly Detection**: In sequential or sensor data (e.g., IoT)
10.	R. Asghar, F. R. Fulginei, et al, 2024 [15]	**LSTM** **(Long Short-Term Memory)**	- Offers greater forecasting accuracy weather in standalone or hybrid - Captures long-range dependencies in sequences - Maintains memory over longer time steps - Effective on noisy and irregular	- Computationally more expensive than simple RNNs - Complex architecture (gates increase parameters) - Slower training and inference - Still less efficient than newer models	- **Time Series Forecasting**: Stock trends, electricity demand, weather prediction - **Healthcare**: Patient monitoring and diagnosis predictions - **Music & Video Analysis**: Sequence

			sequential data - Suitable for variable-length inputs	like Transformers in some tasks	modeling for rhythm, scene transitions
11.	Li, W. and Law, K.E., 2024 [18]	**Hybrid Deep Learning Models** **Ex: CNN+RNN, CNN+LSTM, Transformer+CNN, etc.**	- Leverage strengths of multiple models (e.g., CNN for spatial + RNN for temporal data) - Improved performance and accuracy on complex tasks - More flexible and adaptable to diverse data types - Can handle multimodal data (text, image, video, etc.) - Better generalization in many real-world applications	- Increased architectural complexity - Higher computational cost (more resources needed) - Difficult to train and tune effectively - Longer development and experimentation cycles - Requires large, diverse datasets to fully utilize potential	- **Autonomous Systems**: Combining vision (CNN) and decision-making (RNN or Transformer) - **Finance**: Market prediction using time-series + sentiment analysis
12.	Kate_Doubleday, Stephen Jascourt, et al. 2021 [19]	**Bayesian and Probabilistic Models** (Navie Bayes, Bayesian Networks, Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Probabilistic Graphical Models)	- **Uncertainty Quantification**: Provides probability estimates, not just binary decisions. - **Incorporates Prior Knowledge**: Can use domain expertise through priors. - **Robust to Overfitting**: Especially in low-data regimes. - **Interpretable**: Bayesian reasoning is transparent and often explainable. - **Flexible**: Can model complex, non-linear relationships probabilistically.	- **Computationally Intensive**: Bayesian inference (e.g., MCMC) can be slow for large datasets. - **Requires Prior Knowledge**: Poor or subjective priors can bias results. - **Complex Implementation**: Designing and tuning probabilistic models can be challenging. - **Not Always Scalable**: Inference in large Bayesian networks can be intractable.	- **Spam Detection** (Naïve Bayes) - **Sensor Fusion and Robotics** - **Risk Assessment and Forecasting** - **Anomaly Detection** - **Recommendation Systems** (e.g., Bayesian Personalized Ranking)
13.	N. Omri, J. Jamii, et al., 2023 [20]	**Gaussian Process Regression (GPR)**	- **Uncertainty Estimation**: Provides confidence intervals with predictions. - **Non-parametric Flexibility**: No fixed model structure needed; adapts to data. - **Strong Theoretical Foundation**: Based on Bayesian inference. - **Effective for Small Datasets**: Performs well with limited data. - **Kernel-Based**: Custom kernels allow	- **Scalability Issues**: Computational cost is **$O(n^3)$** for training and **$O(n^2)$** for prediction (n = number of samples). - **Memory Intensive**: Stores entire dataset for prediction. - **Requires Careful Kernel Selection**: Model performance heavily depends on the choice of kernel. - **Poor Performance in High Dimensions**: Can degrade with	- **Spatial Data Modeling** (e.g., geostatistics/Kriging) - **Surrogate Modeling** in engineering and simulations - **Hyperparameter Optimization** (e.g., Bayesian Optimization) - **Time-Series Forecasting** (short sequences)

			domain-specific modelling.	increasing input dimensionality.	
14.	Hussain, S., and AlAlili, A. 2016, [21]	**Bayesian Neural Networks (BNN)**	- **Uncertainty Quantification**: Outputs include confidence intervals or predictive distributions. - **Better Generalization**: Estimates the noise components in the data and achieves superior performance - **Improved Robustness**: Handles noisy or sparse data better than traditional NNs.	- **High Computational Cost**: Inference (e.g., via variational methods or MCMC) is slow and resource-intensive. - **Complex Implementation**: Requires advanced techniques for approximation and training. - **Scalability Issues**: Not easily scalable to very deep networks or massive datasets. - **Difficult Hyperparameter Tuning**: More parameters and distributions to manage.	-Forecasting - Weather and Climate Modeling - Active Learning - Bayesian Optimization - Physics-Informed Machine Learning
15.	AlKandari, M. and Ahmad, I., 2024 [22]	**Ensemble and Hybrid Models**	- **Improved Accuracy**: Typically outperform individual models by reducing variance (bagging), bias (boosting), or both. - **Robustness**: More resistant to overfitting and noise. - **Flexibility**: Can combine models tailored to different aspects of the problem. - **Handles Complex Patterns**: Hybrid approaches can capture both linear and non-linear relationships effectively.	- **Increased Complexity**: Harder to implement, interpret, and debug. - **High Computational Cost**: Training and prediction may require more time and resources. - **Difficult to Tune**: Requires careful selection of base models and parameters. - **Reduced Transparency**: Less interpretable than single-model approaches.	- **Classification & Regression Tasks** (e.g., Random Forest, XGBoost) - **Forecasting Problems** (e.g., weather, energy, finance) - **Image and Text Analysis** (e.g., combining CNNs and RNNs or LSTMs)
16.	M. J. Zideh, P. Chatterjee, et al., 2024 [23]	**Emerging ML Models** [Transformers, Graph Neural Networks (GNNs), Physics-Informed ML (PIML)]	**Transformers** - Excellent at capturing long-range dependencies. - Highly parallelizable, enabling efficient training. - Versatile: can be adapted for vision, text, audio. - State-of-the-art performance in NLP and CV.	- Requires massive datasets and compute. - Prone to overfitting in small-data regimes. - Difficult to interpret internal workings.	Natural Language Processing (NLP) - Computer Vision (CV) - Time-Series Analysis - Multi-modal Learning

			GNNs - Naturally models graph-structured data. - Learns relational dependencies between nodes/edges. - Efficient representation of non-Euclidean domains. - Scales well with sparse graphs.	- Struggles with very large or dense graphs. - Requires careful graph construction. - Limited interpretability.	- Social Network Analysis - Drug & Molecule Discovery - Recommendation Systems - Knowledge Graph Reasoning
			Physics-Informed ML (PIML) - Integrates physical laws into learning via constraints (e.g., PDEs). - Ensures physically consistent predictions. - Data-efficient in scientific domains. - Bridges gap between ML and traditional simulations.	- Requires expert knowledge to formulate constraints. - Solving PDEs within ML models can be computationally intensive. - Still maturing.	- Scientific Simulations - Engineering & Mechanics - Climate and Environmental Modelling - Renewable Energy Forecasting

3. DATA PREPROCESSING TECHNIQUES FOR SOLAR POWER FORECASTING

A critical first step in creating reliable and accurate solar power forecasting models is data preprocessing. Noise, missing values, inconsistencies, and redundant information are frequently present in primary data obtained from various sensing technologies, Photovoltaic (PV) systems, and weather sensors. These problems have the potential to severely impair machine learning (ML) algorithm performance in the absence of proper preprocessing. Preprocessing increases computational efficiency, feature relevance, and model generalizability in addition to improving data quality. The different data preprocessing techniques categories used for forecasting the solar power are reviewed in this section and presented in tabular column as follows.

3.1 Data Cleaning

Data cleaning fixes errors and discrepancies in the dataset. Typical difficulties consist of techniques like mean/median imputation, interpolation, or more sophisticated approaches like K-nearest neighbours (KNN) imputation handle missing data. Statistical techniques (e.g., Z-score, IQR) or machine learning-based anomaly detection (e.g., Isolation Forest) are used to identify outliers caused by anomalous weather occurrences or system malfunctions. Moving averages and other to minimize measurement noise and short-term fluctuations smoothing techniques are used.

3.2 Data Transformation

By converting raw features into appropriate formats Convergence and accuracy of model is enhanced. For distance-based and gradient-based ML algorithms, normalization and standardization are crucial because they bring features to a common scale. To deal with skewed distributions, especially in irradiance or power data, logarithmic or power transformations are utilized Using one-hot or label encoding, categorical encoding transforms weather or temporal indicators into numerical values.

3.3 Feature Engineering

It is greatly improved by creating pertinent features:

- Seasonal and diurnal patterns are captured by extracting temporal features.
- Temporal dependencies in PV generation are represented by lag features, such as past power output at t−1, t−24.
- Time features are subjected to cyclic encoding, which preserves their periodic character by applying sine and cosine transformations.

3.4 Reducing Dimensionality and Choosing Features

Reducing superfluous or unnecessary features lowers model complexity and helps avoid overfitting: Commonly employed techniques include

- Statistical methods (e.g., mutual information, correlation analysis) and machine learning.
- Autoencoders and Principle Component Analysis (PCA) are used to reduce dimensionality while keeping important information.

3.5 Techniques for Signal Decomposition

Decomposition techniques are used to more accurately model the intricate, nonlinear, and non-stationary nature of solar power time series: For better forecasting, the original signal is divided into several components using Wavelet Transform (WT), Empirical Mode Decomposition (EMD), and Variational Mode Decomposition (VMD).

3.6 Processing of Weather and Irradiance Data

Preprocessing weather data is essential since weather has a big impact on PV output Consistency with historical observations is ensured by bias correction and temporal alignment of weather forecast data (such as NWP outputs). By taking into consideration the movement and coverage of clouds in real time, sky image analysis and cloud detection algorithms improve estimates of solar irradiance. Irradiance measurements are frequently normalized using clear-sky models.

3.7 Aggregation and Resampling

To match data resolution across sources, temporal resampling is utilized (e.g., converting minute-level data to hourly). In order to smooth out local variability and more accurately represent grid-level effects, spatial aggregation averages the outputs from several PV plants.

3.8 Data Balancing and Reduction of Noise

Data Balancing works well with high-frequency data and noise reduction is particularly helpful: Some of the methods are Savitzky-Golay filters, denoising autoencoders, and low-pass filters. Resampling or synthetic data generation (like SMOTE) can be used to rectify imbalanced datasets (like those that skew toward midday peaks). By converting unprocessed, noisy, and incomplete data into high-quality inputs appropriate for ML models, Data preprocessing serves as a fundamental component of PV power forecasting. The forecasting horizon, data availability, and model type can be taken into care while selecting the right preprocessing methods. The accuracy and dependability of solar forecasts will be greatly enhanced by sophisticated preprocessing techniques, particularly those involving signal decomposition, feature engineering, and uncertainty handling, as the volume and complexity of solar data continue to increase.

4. GENERALIZED FLOW CHART OF SOLAR POWER FORECASTING

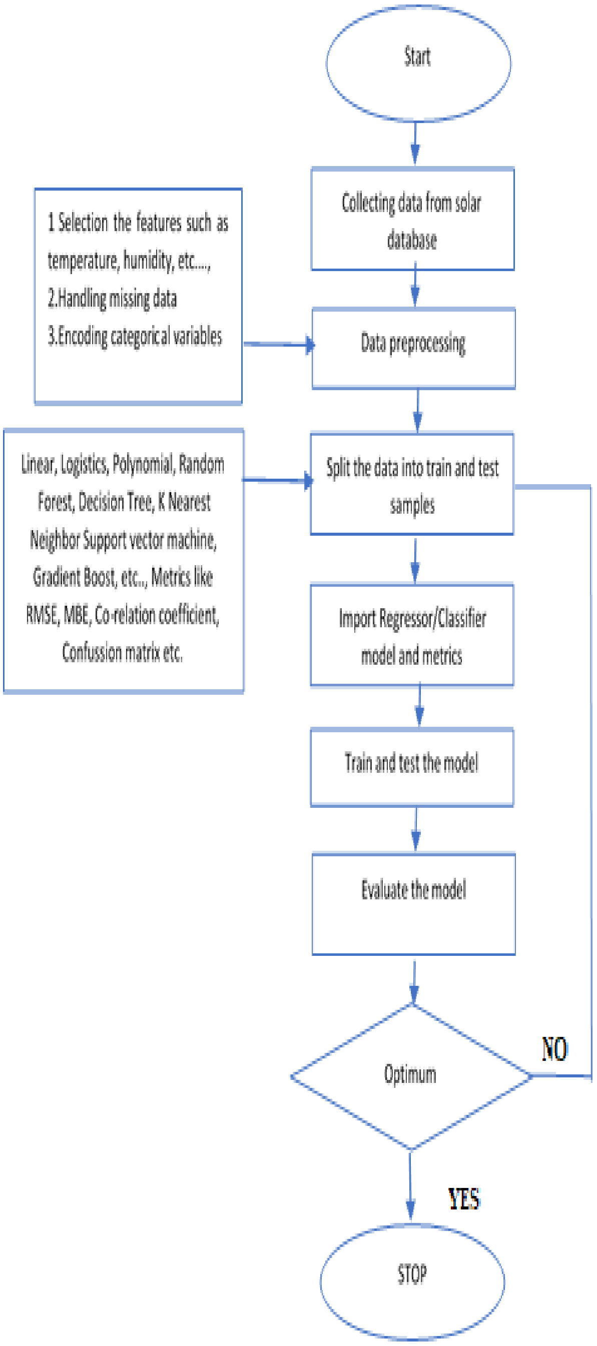
Start
Collecting data from solar database
1 Selection the features such as temperature, humidity, etc...., 2.Handling missing data 3.Encoding categorical variables
Data preprocessing
Linear, Logistics, Polynomial, Random Forest, Decision Tree, K Nearest Neighbor Support vector machine, Gradient Boost, etc.., Metrics like RMSE, MBE, Co-relation coefficient, Confussion matrix etc.
Split the data into train and test samples
Import Regressor/Classifier model and metrics
Train and test the model
Evaluate the model
Optimum
NO
YES
STOP

Fig.1. Generalized Flow chart for solar power forecasting plants

Table 2. Summary of different Data Preprocessing Techniques.

Technique	Advantages	Disadvantages	Applications
Normalization / Scaling (Min-Max, Z-score, etc.)	- Brings features to the same scale - Improves performance of distance-based models (e.g., KNN, SVM)	- Sensitive to outliers - Choice of method impacts model behavior	- Image processing - ML models (SVM, KNN, Neural Networks)
Encoding Categorical Data (Label, One-hot, Target Encoding)	- Converts non-numeric data into usable form - Preserves class information	- One-hot encoding increases dimensionality - Label encoding may introduce ordinal bias	- NLP - Customer segmentation - Recommendation systems
Missing Value Imputation (Mean, Median, KNN, MICE)	- Handles incomplete data - Maintains dataset size	- May introduce bias or noise - Complex methods are computationally expensive	- Medical data - Surveys - Environmental datasets
Outlier Detection & Removal	- Improves model robustness - Reduces skewed performance due to anomalies	- Risk of removing important rare cases - Requires domain knowledge	- Fraud detection - Sensor data cleaning
Feature Selection (Filter, Wrapper, Embedded)	- Reduces dimensionality - Improves training speed and avoids overfitting	- May discard useful features - Computationally intensive methods	- Genomics - Text classification - Financial forecasting
Feature Extraction (PCA, t-SNE, Autoencoders)	- Captures essential patterns - Enables visualization and noise reduction	- May reduce interpretability - PCA assumes linearity	- Image compression - Data visualization - Dimensionality reduction
Data Augmentation	- Increases dataset size - Reduces overfitting	- Risk of introducing unrealistic data - Domain-specific techniques required	- Image classification - NLP - Time-series modeling
Binning / Discretization	- Simplifies models - Useful for rule-based systems	- Loss of information - Can introduce artificial thresholds	- Decision trees - Rule-based systems - Demographic segmentation
Text Preprocessing (Tokenization, Stemming, Lemmatization)	- Converts unstructured text to structured form - Reduces dimensionality in NLP tasks	- May lose contextual meaning - Requires language-specific processing	- Sentiment analysis - Chatbots - Search engines

5. SUITE METRICS

Using a design-of-experiments methodology in combination from literature survey with response surface, sensitivity analysis, and nonparametric statistical testing methods, a detailed analysed framework suggested metrics are of three different kinds mainly to improve the PV forecasting technique.

Forecasting improvements can be divided into three categories as follows (i) uniform improvements when there is no ramp. (ii) improvements in the magnitude of ramp forecasting, and (iii) changes in the threshold for ramp forecasting. Forecasts for simulated and real solar power plants can be examined for both one hour and one day in advance. According to the results of the sensitivity analysis, (i) all of the suggested metrics could be used to demonstrate how the solar forecasts accuracy changes with uniform forecasting improvements, and (ii) the metrics of skewness, kurtosis, and Rényi entropy could be used to demonstrate how solar forecasts accuracy changes with ramp forecasting improvements and a ramp forecasting threshold. A brief description of metrics are described as follows.

5.1 Statistical Metrics

Distributions of forecast errors at various time frames and locations were studied to understand the differences in solar forecasts. The distribution of forecast errors shows the raw forecasting error data in a graphical form; this gives a clear picture of how well the forecasts perform over longer periods. Also, interval forecasts of solar power can help identify the reserve needs to cover forecast errors. This plays vital role in managing and scheduling generating units. Researchers have looked at several distribution types to measure the distribution of solar (or wind) power forecast errors. These include the hyperbolic distribution, kernel density estimation (KDE), normal distribution, and Weibull and beta distributions.

5.2 Kernel Density Estimation (KDE)

KDE is a method that estimates the probability density function of a random variable without assuming any particular parameters. The solar energy community has widely used KDE to characterize wind speed distribution [24,25] and for predicting wind and solar power [26, 27]. KDE is defined as [28].

$$\hat{f}(x;h)=\frac{1}{n}\sum_{i=1}^{n}K_h(x-x_i)$$
$$=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)$$

Where K has a kernel function K and a bandwidth h a smoothing factor.

5.3 Pearson Correlation Coefficient

Pearson's correlation coefficient measures the two variables or data sets. Pearson's correlation coefficient, ρ is represented by the given formula mathematically as,

$$\rho=\frac{cov(p,\ \hat{p})}{\sigma_p\sigma_\beta}$$

In this context, p and pˆ is the present value and predicted solar power output, respectively. Pearson's correlation coefficient serves as a global measure of error. A higher value of this coefficient suggests that the solar forecasting is more accurate. It reflects how closely the overall trends of the forecasts align with the actual values. Because of geographic smoothing, this metric tends to be more useful for assessing forecast accuracy at individual plants or in smaller clusters of plants, rather than across larger balancing authority areas or interconnections. This smoothing effect can lessen the distinction between a good forecast and a poor one.

5.4 RMSE, NRMSE, RMQE and NRMQE

The RMSE metric is a global error measurement, throughout the entire forecasting period defined by,

$$\mathrm{RMSE}=\sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\hat{\mathrm{p}}_\mathrm{i}-\mathrm{p}_\mathrm{i})^2}$$

Forecasts errors are effectively compared across different temporal and spatial scales and this can be normalized using the solar power plant capacity by RMSE. The RMSE, or NRMSE, is particularly sensitive to larger forecast. The RMQE is given by,

$$\mathrm{RMQE}=\left[\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\widehat{\mathrm{P}}-\mathrm{P}_\mathrm{I})^4\right]^{1/4}$$

5.5 MaxAE, MAE, MAPE and MBE

The MaxAE is evaluated to check how well we can predict the events occurring in power system during short-term. Lower the value of MaxAE, more accurate will be the forecast. This metric is particularly good at identifying the biggest forecast error during a specific period in power system. However, it tends to place too much emphasis on extreme events, making it more beneficial. The MaxAE is given by,

$$\mathrm{MaxAE}=max_{i=1,2,\ldots,N}|\ \hat{p}_i\text{-}p_i|$$

The evaluation metric used in regression problems of renewable energy forecasting can be done by using Mean Absolute Error (MAE) which can be defined by,

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|\hat{p}_i-\ p_i\ |$$

Thus, this metric serves as a measurement of global error, but it tends to be less harsh on extreme forecast events compared to the RMSE metric. When you see smaller MAE values, it generally means the forecasts are more accurate and these are expressed as,

$$\mathrm{MAPE}=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{p}_i-p_i}{p_0}\right|$$

$$\mathrm{MBE}=\ \frac{1}{N}\sum_{i=1}^{N}(\widehat{p_i}-\ p_i)$$

P_0 represents the capacity of the solar power plants. For comparison of forecast errors MAPE is usually used. Also, this metric also helps us understand the average bias in our forecasts. A larger MBE means there's more bias in the forecast.

MBE, also helps to enhance forecasts, but it doesn't really capture the full spectrum of forecast errors. For instance, the same MBE value might correspond to various error distributions, some of which could be more favorable than others.

5.6 Kolmogorov-Smirnov Test Integral and OVER

When comparing forecasts over longer time periods and attempting to determine how closely the distributions of those forecasts match the present value of the relevant period, the KSI metric is especially helpful. The major difference among two cumulative distribution functions (or CDFs) is called as the K statistics D [29].

$$D=\max|F(p_i)-\hat{F}(p_i)|$$

The cumulative distribution function of the present value and predicted generated data sets are denoted by the letters F and Fˆ in this context, respectively. The null hypothesis is defined as:

In the event that the D statics, which calculates how the distribution is going to be differed from the actual reference distribution. If this difference is less than the threshold value Vc, it suggests that the two data sets are having exactly similar distribution and are identical statistically. The critical value Vc is calculated by the number of points mentioned in the time series of forecast, which is computed at a 99% confidence level [29].

$$V_O=\frac{1.63}{\sqrt{N}},\quad N>35$$

For each interval, the difference in the actual CDFs and power forecast for each and every interval is defined by [29],

$$D_j=\max\ |F(p_i)-\widehat{F}(p_i)|,\ \ j=1,2\ldots,m$$
$$\text{where }\ \mathrm{p_i}\in[\mathrm{p_{min}}+(\mathrm{j}-1)\mathrm{d},\mathrm{p_{min}}+\mathrm{jd}]$$

The interval distance d, is given by, [28]

$$d\ =\frac{P_{max}-P_{min}}{m}$$

Where, P_{max} and P_{min} are respectively maximum and minimum values of solar power generated. The KSI is the integrated difference between the two CDFs and is given by [29],

$$\mathrm{KSI}=\int_{P_{min}}^{P_{max}}D_n\mathrm{dp}$$

KSIPer is evaluated by the given formula below which is used for the comparison of different spatial and temporal scles of forecasts errors,

$$\mathrm{KSIPer}(\%) = \frac{\mathrm{KSI}}{a_0} \times 100$$

The OVER metric is the difference between the CDFs of present actual and predicted solar Power. Unlike the KSI metric, which looks at all errors, the OVER metric focuses specifically on those significant forecast errors that go beyond a certain threshold, since these larger errors are crucial for power system management [30]. OVER is defined by,

$$\mathrm{OVER} = \int_{P_{\min}}^{P_{\max}} t\mathrm{d}p$$

$$\mathrm{OVERPer}(\%) = \frac{\mathrm{OVER}}{a_0} \times 100$$

Where t is given by,

$$t = \begin{cases} D_j - V_c & \text{if} \quad D_j > V_c \\ 0 & \text{if} \quad D_j \leq V_c \end{cases}$$

5.7 Kurtosis and Skewness

The MAE and RMSE metrics are insufficient, when we want to distinguish among two distributions which are having the identical mean and variance but different skewness and kurtosis. The third is skewness, which quantifies the symmetry of a probability distribution. A Positive skewness in these errors suggests an over forecasting tails, whereas a negative skewness suggests an under-forecasting tail.

5.8 Uncertainty quantification and Propagation Metrics

The two key metrics to help measure the solar power uncertainty are: (i) The Standard Deviation of errors in solar power forecast. (ii) The Renyi entropy of those forecast errors. Traditional forecasting metrics like RMSE and MAE only work well when the error distribution follows a Gaussian pattern. The Renyi entropy is defined by [29],

$$H_{\propto}(X) = \frac{1}{1-\propto}\log_2 \sum_{i=1}^{n} p_i^{\propto}$$

In this context, α (where α > 0 and α ≠ 1) represents the order of the Rényi entropy, which helps us create a range of Rényi entropies. The term P_i is the probability density of the i^{th} discrete section. Generally speaking, a higher Rényi entropy value suggests greater uncertainty in forecasting.

5.9 Swinging Door Algorithm

Different time frames and geographic locations can really impact how severe the ups and downs in solar power output are. By forecasting solar power, we can help minimize the uncertainty that comes with power supply. This is utilized to pinpoint the ramps across the various time intervals [30]. This works by extracting ramp periods from a series of power signals, identifying where each ramp begins and ends. The user sets a threshold that affects how sensitive the algorithm is to variations in ramps. The only adjustable parameter in the algorithm is the threshold, represents the width of a “Door”. The ε parameter directly influences how sensitive the threshold is to noise and minor fluctuations. If ε is smaller, the algorithm will catch many small ramps; if it’s larger, it will only pick up a few significant ramps [30].

5.10 Economic Metrics

Evaluating the economic value of enhancements in solar power forecasting is aided by an objective value of the measure, the reduction in the cost of extra operating reserves associated with the management of solar variability. In the power sector, for the assessment of load (or other) variability forecasts as to how many operating reserves users should procure uses the 95th percentile of the forecast errors [30].

The different metrics and their respective description is given in Table 3 below

Table.3: Different Evaluation metrics

	Abbreviation	Description	Relevance in Solar Forecasting
Mean Absolute Error	MAE	Average of absolute differences between predicted and actual values.	Easy to interpret; useful for measuring overall error magnitude.
Mean Square Error	MSE	Average of squared differences between predicted and actual values.	Penalizes large errors more heavily; good for the analysis of sensitivity.
Root Mean Squared Error	RMSE	Square root of MSE; expresses error in the same unit as the output variable.	Popular for comparing models; highlights large forecasting errors.
Mean Absolute Percentage Error	MAPE	Mean of absolute percentage errors between forecasted and actual values.	Useful for percentage-based error; not ideal when actual values are near zero.
Normalized RMSE	nRMSE	RMSE divided by the mean or range of actual values.	Useful for comparing models across different datasets.
R-squared (Coefficient of Determination)	R^2	Measures the proportion of variance in actual data explained by the model.	Indicates goodness of fit; ranges from 0 (no fit) to 1 (perfect fit).
Symmetric MAPE	sMAPE	Modified version of MAPE using symmetric percentage error.	Avoids issues with division by zero in MAPE.
Mean Bias Error	MBE	Mean of differences between forecasted and actual values.	Shows model bias (positive or negative).
Relative Absolute Error	RAE	Ratio of absolute error to the error of a naïve model.	Indicates how much better the model is compared to a simple average-based model.

F-Score	Skill Score	Compares forecast performance to a reference or baseline forecast (e.g., persistence).	Useful for benchmarking against traditional forecasting methods.

6. CONCLUSION AND FUTURE SCOPE

This work introduces the solar energy forecasting by using different machine learning models. As the ML models have found promising for forecasting solar energy from the literature survey done. This study identifies different ML models from traditional methods to emerging models for forecasting solar power generation, their benefits, drawbacks and applications which are presented in Table 1. Also, the different algorithms being used and the data preprocessing techniques along with their benefits, limitations and use cases are also summarized in Table 2. Furthermore, a detailed description of suite metrics required for the assessing quality of solar power forecast is been discussed in Table 3. Based on these it can be easier for the researchers to look into further future works to improve the reliability, accuracy and also efficiency of the system configuration. However, future work can be done by combining with suitable and task-specific preprocessing pipelines, their efficacy is greatly increased. Achieving high forecasting precision still depends on the cooperation of model selection and preprocessing design.

7. REFERENCES

[1] Essam Y, Ahmed A. N., Ramli R., Chau, K. W, Idris Ibrahim, M. S., Sherif, M. El-Shafie A. “Investigating photovoltaic solar power output forecasting using machine learning algorithms. Engineering Applications of Computational Fluid Mechanics”, 16(1), 2002-2034. https://doi.org/10.1080/19942060.2022.2126528.

[2] P.Ramsami, V. Oree, “ A hybrid method for forecasting the energy outpiut of photovoltaic systems”, Energy Converse. Manag., Vol. 95, pp. 406-413, May 2015, doi:10.1016/j.enconman.2015.02.052

[3] Lin, K.-p.,-F., “Solar power output forecasting using evolutionary seasonal decomposition least square support vector regression , Journal of cleaner Production (2015), http://dx.doi.org/10.1016/j.jclepro.2015.08.099

[4] Huang, Jing & Perry, Matthew, 2016, “A Semi-emperical approach using gradient boosting and k-nearest neighbors regreszsion for GEFCom 2014 probablistics solar power forecasting, “ International Journal of Forecasting, Elsevier, Vol. 32(3), pages 1081-1086. doi: 10.1016/j.ijforecast.2015.11.002

[5] H. Long, Z. Zhang, and Y. Su, “Analysis of daily solar power prediction with data-driven approaches, “Appl. Energy, Vol. 126, pp. 29-37, Aug. 2014, doi: 10.1016/j.apenergy.2014.03.084

[6] A Nespoli, S. Leva, M. Musetta and E. G. C Ogliari, “ A selective ensemble approach for accuracy improvement and computational load reduction in ANN-based PV power forecasting”, IEEE Access, Vol. 10, pp. 32900-32911, 2022, doi: 10.1109/ACCESS.2022.3158364

[7] Lopez Gomez, J.Ogando Martinez, A. Troncosco Pastoriza, F. Febrero Garrido, L. Granada Alvarez, E. Orosa Garcia, J. A, “ Photovoltaic Power prediction using Artificial Neural Networks and Numerical weather data”, Sustainibility 2020, 12,10295. https://doi.org/10.3390/su122410295.

[8] Keddouda A, Ihaddadene r, Boukhari A, Arici M, Lebbihiat N, et al., “ Solar Photovoltaic Power prediction using Artificial Neural Network and Multiple regression considering ambient and operating conditions”, Energy Converge Management, 2023, https://doi.org/10.1016/j.enconman.2023.117186. https://www.researchgate.net/publication/371094033

[9] M. Abuella and B. Chowdhury, "Solar power probabilistic forecasting by using multiple linear regression analysis," *SoutheastCon 2015*, Fort Lauderdale, FL, USA, 2015, pp. 1-5, doi: 10.1109/SECON.2015.7132869.

[10] Gavrishchaka, Valeriy & Ganguli, Supriya. (2001). Support vector machine as an efficient tool for high-dimensional data processing: Application to substorm forecasting. Journal of Geophysical Research. 106. 29911-29914. 10.1029/2001JA900118

[11] Neha Singh, Satyaranjan Jena, Chinmoy Kumar Panigrahi, “A novel application of Decision Tree classifier in solar irradiance prediction”, Materials Today: Proceedings, Volume 58, Part 1, 2022, Pages 316-323, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2022.02.198. (https://www.sciencedirect.com/science/article/pii/S2214785322008124).

[12] Rathika Senthil Kumar, P.S. Meera, V. Lavanya, S. Hemamalini, “Brown bear optimized random forest model for short term solar power forecasting”,Results in Engineering, Volume 25, 2025, 104583, ISSN 2590-1230, https://doi.org/10.1016/j.rineng.2025.104583. (https://www.sciencedirect.com/science/article/pii/S2590123025006619)

[13] Zhe Song, Fu Xiao, Zhe Chen, Henrik Madsen, “Probabilistic ultra-short-term solar photovoltaic power forecasting using natural gradient boosting with attention-enhanced neural networks”, Energy and AI, Volume 20, 2025, 100496, ISSN 2666-5468,https://doi.org/10.1016/j.egyai.2025.100496. (https://www.sciencedirect.com/science/article/pii/S266654682500028X)

[14] Nor Azuana Ramli, Mohd Fairuz Abdul Hamid, Nurul Hanis Azhan, Muhammad Alif As-Siddiq Ishak, “ Solar power generation prediction by using k-nearest neighbor method”, *AIP Conf. Proc.* 2129, 020116 (2019) https://doi.org/10.1063/1.5118124

[15] R. Asghar, F. R. Fulginei, M. Quercio and A. Mahrouch, "Artificial Neural Networks for Photovoltaic Power Forecasting: A Review of Five Promising Models," in *IEEE Access*, vol. 12, pp. 90461-90485, 2024, doi: 10.1109/ACCESS.2024.3420693.

[16] Nattha Thipwangmek , Nopparuj Suetrong, Attaphongse Taparugssanagorn , Suparit Tangparitkul , And Natthanan Promsuk, “ Enhancing Short-Term Solar Photovoltaic Power Forecasting Using a Hybrid Deep Learning

Approach", 16 August 2024. Digital Object Identifier 10.1109/ACCESS.2024.3440035

[17] Meftah Elsaraiti and Adel Merabet , "Solar Power Forecasting Using Deep Learning Techniques", March 17, 2022, date of current version March 25, 2022. Digital Object Identifier 10.1109/ACCESS.2022.3160484.

[18] W. Li and K. L. E. Law, "Deep Learning Models for Time Series Forecasting: A Review," in *IEEE Access*, vol. 12, pp. 92306-92327, 2024, doi: 10.1109/ACCESS.2024.3422528.

[19] Kate Doubleday, Stephen Jascourt, William Kleiber, Bri-Mathias Hodge, " Probabilistic Solar Power Forecasting Using Bayesian Model Averaging", IEEE Transactions on Sustainable Energy, Jan 2021, https://doi.org/10.1109/tste.2020.2993524

[20] N. Omri, J. Jamii, M. Mansouri and M. F. Mimouni, "Solar Irradiance Forecasting using Gaussian Process Regression Model," *2023 20th International Multi-Conference on Systems, Signals & Devices (SSD)*, Mahdia, Tunisia, 2023, pp. 99-103, doi: 10.1109/SSD58187.2023.10411196.

[21] Hussain, S., and AlAlili, A. (October 18, 2016). "Online Sequential Learning of Neural Networks in Solar Radiation Modeling Using Hybrid Bayesian Hierarchical Approach." ASME. *J. Sol. Energy Eng*. December 2016; 138(6): 061012. https://doi.org/10.1115/1.4034907

[22] AlKandari, M. and Ahmad, I., 2024. Solar power generation forecasting using ensemble approach based on deep learning and statistical methods. *Applied Computing and Informatics*, *20*(3/4), pp.231-250.

[23] M. J. Zideh, P. Chatterjee and A. K. Srivastava, "Physics-Informed Machine Learning for Data Anomaly Detection, Classification, Localization, and Mitigation: A Review, Challenges, and Path Forward," in *IEEE Access*, vol. 12, pp. 4597-4617, 2024, doi: 10.1109/ACCESS.2023.3347989.

[24] Choudhury, Souma & messac, Achille & Castillo, Luciano. (2011), " Multivariate and multimodal Wind Distribution model based on kernel density estimation", Renewable energy. Doi.org/10.1115/ES2011-54507

[25] Zhang J , Choudhury, Messac, Achille & Castillo, L., 2013a., 2013a., " A Multivariate and Multimodal Wind Distribution model", Renewable Energy, Vol 51, 436-447. http://dx.doi.org/10.1016/j.renene.2012.09.026

[26] Zhang. J, Hodge, B.-M., Florita, A., 2013e, Joint probability distribution and correlation analysis of wind and solar power Forecast Errors in the Western Interconnection. Journal of Energy Engineering.http://dx.doi.org/10.1061/(ASCE)EY.1943-7897.0000189_

[27] Juban, J., Siebert, N., Kariniotakis, G. N., 2007, " Probabilistic short-term wind power forecasting for the optimal management of wind generation. IEEE Power Eng. Society, Lausanne Power Tech Conf. proc., 683-688. IEEE Transactions on industry Applications, Vol. 49. No. 6, Nov/Dec2013.http://dx.doi.org/10.1109/PCT.2007.4538398

[28] Jones, M. Marron, J Sheather. S, "A brief survey of bandwidth selection for density estimation", American Statistical Association., 1996., vol 41, 401-407, http://www.jstor.org/stable/2291420

[29] Espinar B, Ramirez L, Drews A, Beyer H G, Zarzalejo L F, Polo J, martin L, Analysis of different comparisonparameters to solar radiation data from satellite and german radiometetric stations. Solar Energy, 2009, vol. 83, 2009, https://doi.org/10.1016/j.solener.2008.07.009

[30] Florita M, Hodge and K Orwig, "Identifying wind and Solar ramping Events", IEEE green technologies Conference (Green Tech), 2013, 147-142, doi: 10.1109/GreenTech.2013.30

[31] Zhu H, Li X, Sun Q, Nie L, Yao J, Zhao G. a power prediction method for photovoltaic power plant based on wavelet Decomposition anrtificial neural networks, energies, 2015, vol 9 https://doi.org/10.3390/en9010011.

[32] Grimaccia F, Leva S, Mussetta M, Olgiari E, "ANN sizing procedure for the day- ahead output power forecast of a PV plant", Appli science, Vol 7 (6):622, https://doi.org/10.3390/su151411144

[33] Bahera M K, Majumdar I , Nayak N," Solar Photovoltaic Power forecasting using optimised modified extreme learning machine technique", Engineering science and technology, An International Journal, Vol 21 Issue 3, 2018, 245 : 114569, https://doi.org/10.1016/j.jestch.2018.04.013

[34] Guermoui M, Bouchouicha K, bailek N, Boland J W, " Forecating Intra hour varience of photovoltaic power using a new integrated model", Energy Conversion Management, 2021, Vol 245:114569, https://doi.org/10.1016/j.enconman.2021.114569

[35] Jiang M, Ding K, Chen X, Cui L, Zhang J, Yang Z, et al, "Research on Time series based and similarity search based methods for PV power Prediction", Energy Conversion Management, 2024, Vol, 308: 118391. https://doi.org/10.1016/j.enconman.2024.118391

[36] Lin h, Gao L, Cui M, Liu H, Li C, Yu M, " Shprt term distributed photovoltaic power prediction based on temporal self attention mechanism and advanced signal decomposition techniques with feature fusion", Energy, 2025, Vol 315:134395, https://doi.org/10.1016/j.energy.2025.134395.

[37] M Gao, J li, F Hong and D Long, " Day-ahead power forecasting in a large scale photovoltaic plant based on weather classification using LSTM", Energy, Vol. 187, 2019, Art No. 115838, doi:10.1016/j.energy.2019.07.168.

[38] S Theocharides, G makrides, A Livera, M Theristis, P Kiamakis and G E Georghiou, " Day-ahead photovoltaic power production forecasting methodology based onm machine learning and Statistical postprocessing", Appli. Energy, Vol 268, Jun 2020, Art No. 115023, doi:10.1016/j.apenergy.2020.115023.

[39] S Pretto, E Ogliari, A Niccolai, A, Nespoli, "A New Probablistic Ensemble Methos for an enhanced day-ahead PV power forecast", IEEE J.Photovolt, Vol 12, pp 581-588, Mar 2022, doi: 10.1109/JPHOTOV.2021.3138223

[40] A Abdellatif, h Mubarak, S Ahmed, T Ahmed, G M Shafiullah, A Hommaoudeh, H Abdellatif, M M Rehman and H M Gheni, "Forecasting Photovoltaic Power generation with a stacking Ensemble Model", Sustainibility, Vol 14, no 17, p 11083, Sep 2022, doi: 10.3390/su141711083

[41] W Khan, S Walker, W Zeiler," Improved Solar Photovoltaic Energy Generation Forecast using Deep Learning based Ensemble Stacking Approach", Energy, Vol 240, Feb, 2022, Art No. 122812, doi: 10.1016/j.energy.2021.122812.