Sensor-Independent One-Hour-Ahead Forecasting and Anomaly Detection of Grid-Connected PV Inverters Using an Interpretable Random Forest Framework

[1]S. Arunkumar, [2]A. John Pradeep Ebenezer, [3]Mayakannan Selvaraju, [4]P.Meruna, [5]Sriharini. B,[6]N. Bala Krishnan

[1]Associate Professor, Department of Mechanical Engineering, Vinayaka Mission's Kirupananda Variyar Engineering College, Vinayaka Mission's Research Foundation (Deemed to be University), Salem, Tamilnadu - 636 308

[2]Assistant Professor and Head, PG Department of Computer Applications, St. Joseph's College of Arts and Science (Autonomous), Cuddalore, Tamilnadu, India – 607001

[3]Associate professor,Department of Mechanical Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, SIMATS, Chennai, Tamil Nadu, India – 602105.

[4]Department of Medical Electronics, Sengunthar Engineering College and Tiruchengode, Namakkal, Tamilnadu - 637205

[5]Department of Biomedical Engineering, Velalar College of Engineering and Technology, Thindal, Erode, Tamilnadu - 638012.

Department of Manufacturing Engineering, Vinayaka Mission's Kirupananda Variyar Engineering College, Salem, Tamil Nadu, 636 308

Email ID: [1]shaiarun1978@gmail.com, [2]john_pradeep@sjctnc.edu.in,[3]kannanselva1986@gmail.com, [4]merunaperumal@gmail.com, [5]sriharinib7066@gmail.com, [6]sabaribalu2000@gmail.com

ABSTRACT

Dependable tracking of grid-linked photovoltaic (PV) systems is also unfeasible because it relies on meteorological sensors and rule-based management, especially in sensor-limited situations. The study constructs a machine learning model interpretable to humans based on electrical measurements of a PV plant of 30 kW and 40 kW inverters, and a five-minute resolution dataset of active and reactive power, phase voltages and phase currents and time dependent features during January 2025. An application of the regression, classification, and Z-score anomaly detection method was conducted using the Random Forest and conditioned on the timestamp alignment, feature engineering, percentile-based categorization of the output, and a one-hour-ahead label shifting, and the method was validated with the aid of the temporal splits and 5-fold Time Series cross-validation. The regression model was found to have great predictive accuracy (MAE = 0.12 kW, R^2 = 0.995), and cross-validation performed showed great robustness (R^2 = 0.9802, MAE = 0.1024 kW). The Z-score analysis (z = 3) revealed the presence of anomalous samples (2.77%). Though the accuracy of a static classification was 86 percent, time-varying forecasting lowered the accuracy to 33 % (macro F1 = 0.33) which points to the impact of dynamic environmental variability. The suggested light and interpretable structure allows predicting the performance of inverters in real-time, early detecting anomalies, and intelligent planning of PV systems maintenance without the use of external meteorological devices.

Keywords: Photovoltaics, Random forest, Z-Scan, PV-plant, Sensors

1. Introduction

Power shifts toward renewable sources like wind and solar take front stage in worldwide efforts to reduce carbon emissions. Supervised learning methods, such as Random Forest Regressor, which follows an ensemble approach are applied to the predictive modeling of the photovoltaic power generation. The machine learning models are associated with a

lower Mean Absolute error and high forecasting accuracy that optimizes the performance of the solar PV inverter [1]. Photovoltaic structures are printed functionally graded materials that are multi-directional in order to improve thermal stability and the solar PV inverter efficiency. The results of the predictive modeling are authenticated by a hybrid machine learning model that refers to supervised learning methods to verify the existence of better power output and structural reliability [2]. Machine and supervised learning methods to divide photovoltaic systems in aerial images to enhance integration in the grid. Predictive modeling framework increases the accuracy of localization, which supports the effective deployment of solar PV inverters and the planning of the system [3]. A hybrid machine learning model is a combination of PVLib simulations and supervised learning methods to forecast optimal photovoltaic tilt angles under the various climates. Random Forest Regressor demonstrated great predictive modeling and minimal Mean Absolute Error (approximately 2.04 o C), which enhanced the solar PV inverter energy yield considerably [4]. The photovoltaic modules dust detection system is a machine learning-based system that uses Random Forest model to aid the intelligent solar PV inverter maintenance. The prediction modeling framework has an accuracy of 91.3 with good performance indicators that can ensure cleaning plans are optimized and power production is improved [5].

A hybrid/debiased machine learning model measures the causal relationship of PM2.5 on which photovoltaic power can be produced, which is better at predicting models than standard econometric approaches. The findings show serious outcomes of solar PV inverter degradation and strong estimation quality verified by sophisticated supervised learning strategies and low bias of Mean Absolute Error [6]. Fault detection in photovoltaic system is implemented using the supervised learning methods to improve the reliability of solar PV inverter and predictive modeling of photovoltaic systems. Classifiers based on machine learning such as the Random Forest could detect the SC, OC, GF, and MF faults with up to 98% accuracy, which is beneficial in terms of intelligent real-time monitoring and enhanced operational stability [7]. An RSM model is suggested and uses machine learning to predict the quality of generated power in grid-connected photovoltaic systems in different environmental conditions [8]. A hybrid waste-based energy system and photovoltaic system combines the machine learning through the use of a Random Forest Regressor to predict the most effective resource allocation and predictive modeling. The supervised learning method is effective in improving energy and exergy efficiency besides stabilizing the functioning of solar PV inverters to produce power in a sustainable manner [9]. It constructs a machine-based photovoltaic fault detection supervised classification framework by utilizing real-time electrical parameters. Random Forest and other ensemble methods with a high accuracy of more than 99 percent and low Mean Absolute Error reinforced predictive modeling and solar PV inverter reliability [10].

To fill these gaps, new studies need interpretable and computationally lightweight models that can work in real time with few sensor inputs. The proposed framework integrates short-term PV power forecasting with anomaly detection using a Random Forest model, combining prediction and operational diagnostics to enhance reliability and maintenance planning beyond conventional standalone forecasting approaches. A statistical anomaly indicator (Z-score) and an ensemble-based framework (Random Forest) provide an attractive trade-off between visibility, scalability, and predictive capacity in this setting. These methods, in accordance with the purpose of this study, provide the basis of an effective and realistic PV system monitoring.

2. Materials and methods

The data used in this study were obtained from a grid-connected photovoltaic plant. The system will consist of two inverters, one with 30 kW and the other with 40 kW, having grid monitoring sensors where the power and voltage specifications will be captured thrice. The export window of the platform, in this data set, is January 1, 20252 February 1, 2025.

2.1 Data sources

• Inverter Information: 30 and 40 kW inverters' active power output data with a 5-mins.

• Grid Power: Reactive and Active power (W, var).

• Grid Voltage: Phase A, phase B and phase C voltage readings.

• Environmental Context: We were able to indirectly predict sun exposure and diurnal cycles by extracting time-related parameters (hour, day, and weekday/weekend) in the absence of data on direct irradiance. The original purpose of this was to substitute the on-site weather sensors, which were not operational. Such proxies do not produce short-term weather fluctuations, but do preserve daily patterns and for the creation of similar characteristics with uninstrumented sites. The study utilized one month of high-resolution data to evaluate short-term forecasting robustness under real operational variability. Although longer datasets may improve seasonal generalization, this duration was sufficient to capture diurnal cycles, transient anomalies, and inverter behavioral patterns.

2.2 Preprocessing and Feature Engineering

The ML analysis could proceed after following preprocessing steps:

• Timestamp alignment: It were standard practice to resample and merge the data every five minutes.

• Transformation and feature normalization: Selected characteristics was either derived from or scaled according to the scale of preexisting traits, such as:

• Regression analysis relied on inverters' ability to maintain a constant active power output.

• We used the 33^{rd} and 66^{th} percentile quantile bounds to classify the inverter output as three operation classes: Low, Medium, and High. Training the model will be more stable with this quantile-based approach since it will give a balanced representation of the classes. But these limitations are not always indicative of the inverter's maximum capacity or the grid's maximum operating circumstances. They are better suited as data-driven categories for model comparison rather than for direct use in real data modeling. In further alignment of the result of the categorization with practical interpretation, further work could be done with domain-knowledge thresholds, e.g., with grid efficiency or inverter nominal power range recommendations. One can also consider applying the unsupervised methods of clustering such as Gaussian Mixture or K-Means Model to get data-driven borders between classes which would match the natural production regimes. Key features included active power, reactive power, phase voltages, phase currents, hour-of-day, and weekday indicators. These

were selected because they directly represent inverter electrical dynamics and indirectly encode irradiance-related patterns via temporal proxies, enabling sensor-independent modeling.

One hour in advance predictions required 12 time steps (equal to 5 minutes) to produce labels by moving the target variable. Instead of explicit lag features, the model input made use of time-based features and contemporaneous electrical features to keep temporal dependencies intact without adding further dimensions to the features. By utilizing the hour, day, and weekday identifiers, this methodology represents the diurnal and cyclical tendencies and prevents the chance of information leaking through training and test divides. Despite its usefulness in converging classifiers and its ease of use for label formation, this statistical technique not align with physical thresholds like grid operating areas or inverter rated power. A more domain-informed discretization can create categories with more apparent practical interpretation, such as establishing production classes by inverter specs or defined power quality bands.

The study examines various supervised machine learning methods in order to study inverter behavior and predict performance.

2.3 Regression Modeling

In case of regression and classification, we have decided to use the Random Forest (RF) models as they demonstrate the strengths, interpretability, and handling of structured tabular data, small samples, and mixed features. To capture nonlinear dependencies but no longer care about a temporal structure, RF models can be used in place of sequential deep learning models, such as LSTM networks, and large volumes of hyper parameter optimization and continuous time data are required [11]. Also brought forward were gradient-boosting models like Light GBMor XGBoost; early trials revealed that these models reduced transparency and increased computing complexity while achieving very modest increases (less than 1% in R^2). The primary rationale for not choosing Support Vector Machines (SVM) was their vulnerability to feature scaling and its inability to handle scalability issues when applied to multivariate energy data. Interpretability, deployment and predictive accuracy, simplicity are all factors to be considered while choosing RF. Random Forest was selected due to its balance between interpretability, computational efficiency, and strong nonlinear modeling capability. Unlike deep neural networks, it does not require extensive hyper parameter tuning or large-scale datasets, making it suitable for real-time PV monitoring systems.

We trained Random Forest Regressor take grid-side variables and time-related characteristics into account when estimating the inverters' power production at any given instant. The evaluation was based on the model's performance:

2.4 Multi-Class Classification

We obtained the operational by dividing the inverter output as three performance bands. In order to train the Random Forest Classifier to classify, we used time and grid -based features:

- Confusion matrix visualization
- F1-score

• Accuracy

• Inverter state 1 h into the future

• Current inverter state

Performance were assessed by:

To ensure that there is no chronological disorder or time leakage in one-hour-ahead work, use 5-fold Time Series Split to partition the Time Series. To achieve this equilibrium, we apply two methods: (i) a balanced distribution of class weights in the tree; (ii) limiting SMOTE to each training fold; and (iii) evaluating on the remaining validation slice. We report balanced accuracy, accuracy, and macro F1 as well as out-of-the-box (OOF) forecasts averaged over the studied area (with the first indices intentionally excluding Time Series Split).Model evaluation was conducted using chronological 80/20 temporal split to prevent leakage, followed by 5-fold Time Series Split cross-validation. Performance metrics included MAE and R^2 for regression, and Accuracy, Macro-F1, and Balanced Accuracy for classification. Out-of-fold predictions were analyzed to ensure robustness under time-aware validation.

2.5 Anomaly Detection

Using a Z-score, a system was developed to detect inverter power output anomalies:

• Using the past forecasts, we were able to determine a rolling mean.

• To show any discrepancy with a standard deviation more than +3, z-scores were computed.

• Day of the week and grid activity were cross-validated with anomaly durations.

With a 0.25 step size, we counted the percentage of samples deemed anomalous as we scanned z [2.0, 4.25] to develop the Z-score threshold. At z=2.0, the share flagged reduced steadily for the entire inverter-40 kW series, while at z=4.25, it dropped to 0.18. There was a 2.77 percentage point flagging in the standard z = 3.0. The 2.77% anomaly rate suggests the presence of heavy-tailed residual behavior rather than purely Gaussian noise. This highlights the importance of empirical threshold tuning rather than assuming normality in inverter performance deviations. This is significantly higher than 0.27% under normal assumptions; it indicates heavier residuals, which necessitate empirical sweep instead than relying at simple Gaussian assumptions. To offer a reasonable compromise between sensitivity and alert volume, we present data for z = 3.0 in subsequent studies with a sensitivity range to make the results comprehensible.

2.6 Exploratory Data Visualization

Feature engineering and system activity was better understood with the help of several representations.

• Feature importance rankings

• Power and voltage correlation matrices

• Missing data heatmaps

• Hourly production for inverters

Pandas, matplotlib, scikit-learn, shap, and seaborn are some of libraries utilized in the analyses performed in Google Colab's Python environment.

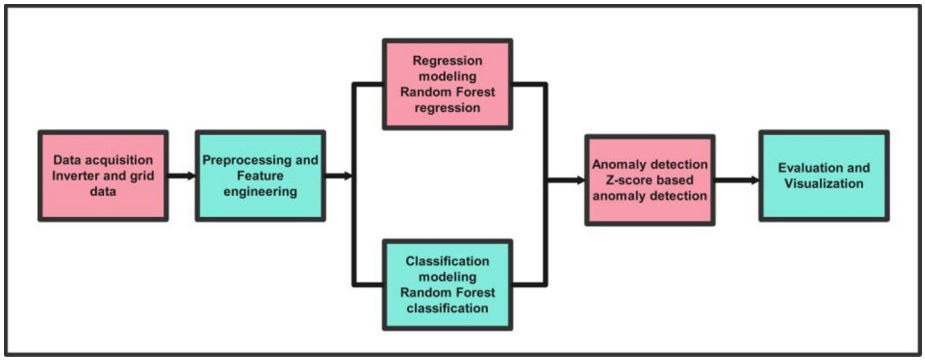


Fig. 1. Workflow for PV predictive modeling and anomaly detection.

For this investigation, we chose to use the z-score and random forest models, due to interpretability, strength and suitability. This was not done with the aim of building a model with maximum absolute forecast accuracy but instead to develop an understandable, data-efficient baseline model for PV performance model. Other ML models such as LSTM network, XGBoost or Support Vector Machines could be considered. We will break down the computational cost, time for learning and interpretability, trade-offs in the future with benchmarking various systems [12].

In the same vein as the identification of anomalies, in this study significant deviations at the inverter output were highlighted using a Z-score method. This approach is maybe basic, but the results of operation diagnostics it gives are simple, clear and easy to understand.

3. Results and discussion

To get a feel of how the inverter behaves, we looked at 30 and 40 kW on average daily production on all of the days of dataset. Figure 2 shows hourly production profile in the period January 2025 to February 2025 using the data with 5-minute resolution.

.

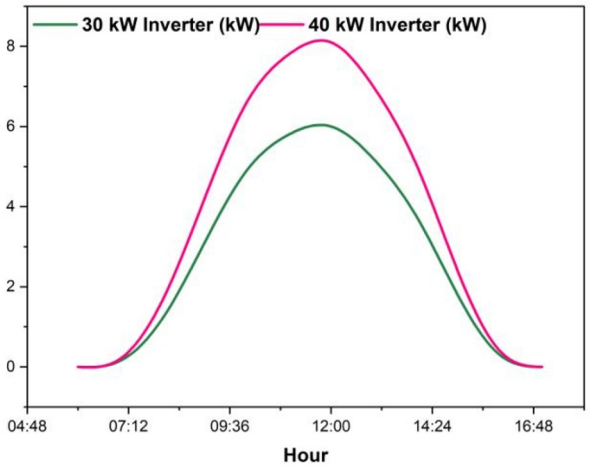


Fig. 2. Mean daily output versus hour.

As is apparent in the chart, both inverters possess the standard solar output curve:

The solar noon is situated between 12.00 and 13.00, and the power output starts at approximately 07.00 and moves up at a slow rate until it reaches its maximum.

• A 40 kW will always provide more energy 30 kW one (black line) since it is expected since they have different capacities as per the calculations.

To follow the day winter time, the production decreases consistently between 13:00 and complete closure around 17:0018:00.

• Most days were characterized by few faults, minimal shading and the curves had a regular structure indicating that the behavior of the inverter was predictable.

Establish the level of relationship amongst the key variables of the electric plant that include active power in 30 and 40 kW, Phase A at grid reactive and grid meter power by analyzing correlation matrix as indicated in Figure 3. We used the Pearson technique to find correlation coefficients over the entire set of data.

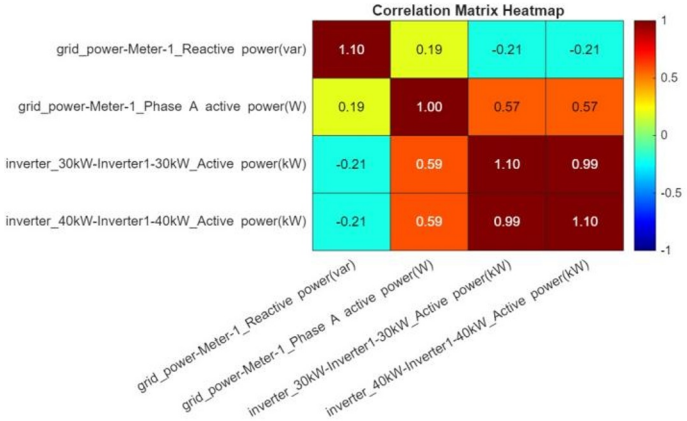


Fig. 3. Electrical Variable Correlation Matrix.

The 30 and 40 kW recorded a very strong positive correlation in terms of active power output. This information gives plausibility to the hypothesis that the same patterns of output of the inverters can be attributed to common environmental factors such as temperature and sun irradiation. Such uniformity leads to the possibility of taking data of an inverter and using it to confirm or even predict the performance of the other inverter which would be of great help in redundancy analysis and model generalization methods.

When correlations were moderate (about r= 0.57), all invertor active power output was figured to be proportional to the active power of Phase A of grid. This is an indication that, although imperfect, there is a rather close relationship between grid injection and inverter manufacturing. It means that the inverter outputs do not describe the effect of other system dynamics on active power on the grid, including consumption at the local level and losses.

On the other, the grid reactive power had weakly negative correlations (ranging between -.20) with all other variables. The hypothesis to be tested with this finding is that reactive power is regulated by power factor correction systems or by other capacitive loads and inductive, and is therefore independent of active power output. This decoupling is not caused by energy conversion, but rather by grid-code-based requirements of reactive voltage and power supply management, which initiates this decoupling.

The relationships have demonstrated how grid-side disturbances can affect the performance and functioning of inverters, as relevant data to grid-integration. The inverter control loops may require excessive strain because of transitory voltage imbalance or low the point of common connection which may lead to poor long term reliability. On the contrary, grid injection is simpler with constant voltage characteristics and inverters perform better.

Such interdependencies should therefore be understood so as to implement inverter-grid strategies to enhance power quality.

Altogether, the content in this section detects credible inverter output trends, the semi-controlling effect of grid dynamics, and reactive power decoupling. According to these results, we can more effectively select the words to include in our prediction model and draw our attention to which features of power require particular consideration within the system.

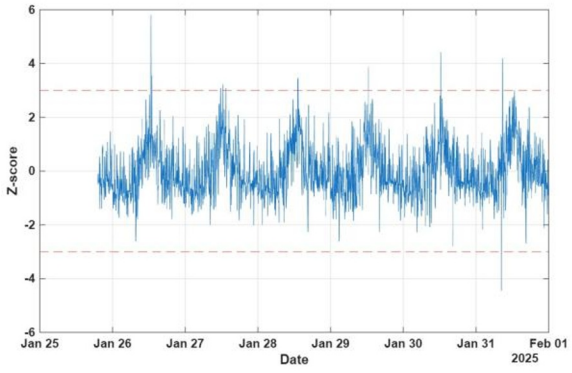


Fig. 4. Z-Statistic Fault Detection.

According to the results, there are several outliers at inverter output profile because the Z-score values jump above the +3 mark. These occurrences do not appear to happen at

random, but rather in clusters; the 10th of January and the 24th–31st of January were the most noticeable of these. This sort of time-dependent concentration, as opposed to sensor noise or random variation, suggests repeated or ongoing resistance. The integration of statistical Z-score anomaly detection with model-based residual monitoring enables early identification of inverter deviations before catastrophic failure. By detecting clustered deviations during high irradiance periods, maintenance scheduling can be optimized based on data-driven alert thresholds rather than reactive fault correction strategies.

Temporary climatic conditions (such as cloud flicker or snow cover), instrument malfunctions (such as inverter shutdown or power cutback), or inaccurate data gathering could account for these outlier results. Interestingly, we did not find any negative outliers that were more than -3, which lends credence to the idea that most deviations are caused by unanticipated rises rather than declines.

Finding those outliers were crucial for maintenance scheduling and operational reliability. It lets plant operators to keep an eye out for unusual performance.

3.1 ML-Based Power Output Prediction

ARF Regressor was trained to forecast active power a 40 kW to find out how operational and temporal factors can predict photovoltaic power. An index of continuous time and real-time electrical characteristics made up the collection of features.

To prepare the data for the model, we first removed any entries without the desired values. Notably, the system's expected operation also includes genuine, non-forward-filled downtime during the night.

Using an 80/20 split between testing subsets and dataset's training, Random Forest model achieved good predictive accuracy, with the following outcomes:

• Mean Absolute Error = 0.12 kW.

• $R^2 = 0.995$, show variance

For the last week of January 2025, Figure 5 displays time-series of anticipated and actual power output of a 40 kW. A30 kW signal and time-related patterns were used to train a Random Forest Regressor, which was then used to predict the values. The Random Forest regressor achieved MAE = 0.12 kW and $R^2 = 0.995$ under the 80/20 chronological split, while cross-validated performance yielded $R^2 = 0.9802$ and MAE = 0.1024 kW, confirming both accuracy and temporal robustness.

This number shows that during all of the daily production cycles, there is a very close relationship between the actual and predicted numbers. A predicted curve closely matches observed output and, with a small amount of delay or distortion, reproduces ramp-up/down phases and peak intensities.

Notably:

• Daytime, when power production is non-linear with irradiance and operational dynamics, is when prediction accuracy is highest. It is possible for the model to represent these intricate variations. As expected, night values are also zero, proving that the model learns typical operating patterns and can not be tricked into thinking there is output when

there isn't. At the sharp transitions between the peaks, there are small deviations that might be short-term changes or inverter curtailment that the input features do not adequately explain.

In addition to quantitative results (e.g., MAE = 0.1214 kW, R 2 = 0.9975), this visual verification further proves that the trained model is suitable for power prediction at high-resolution predictions. Short-term results modeling in PV monitoring systems can make use of sophisticated diagnostics of inverters, performance benchmarking, and short-term projections.

Additionally, we used Time Series Split method with five successive folds to perform time-sensitive cross validation, which allowed us to provide robustness of results and rule out leaking. We kept the chronicological order in every fold so that the predicting conditions in the real world would be better simulated. Overall, the out-of-fold results showed that Random Forest Regressor had predictive fidelity across folds (R^2 = 0.9802 [0.9945 - 0.9828] and MAE = 0.1024 [0.0862 - 0.1346] kW), and the small difference in R^2 across folds was because of expected changes in daily irradiance, not because of model instability. These outcomes provide more evidence that proposed regression method is suitable for monitoring continuous PV performance and confirm that the high correlation between Fig. 5 is very stable even among hidden time intervals.

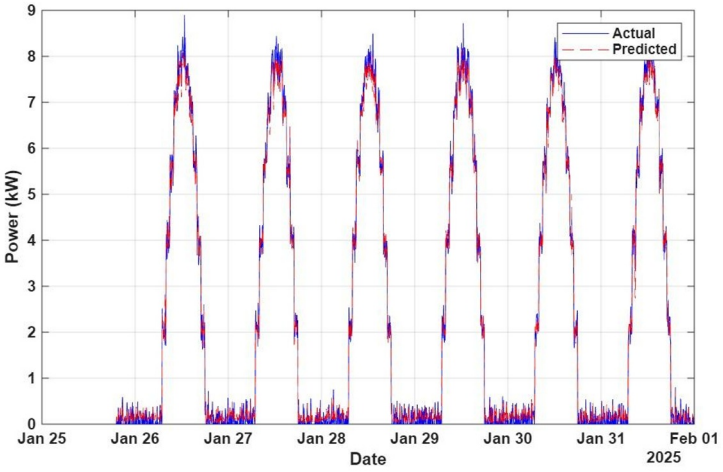


Fig. 5. Actual and Forecasted Inverter Power.

Feature Importance

Figure 6 presents the ranking of four input features recommended by the Random Forest Regressor for predicting 40 kW active power output. In order to determine how important features are for reducing errors in the ensemble as a whole, we used the mean decrease in impurity as our metric.

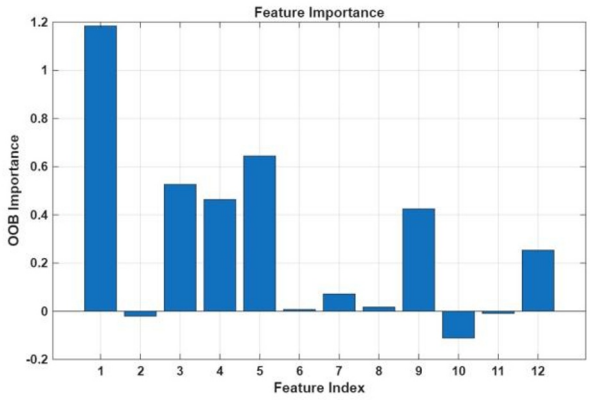


Fig. 6. Key Predictors of 40 kW Output.

The inverter's response to operating behavior and electrical load in real-time. Their combined contribution of 0.15 to 0.20 suggests interphase current flow patterns convey information about short-term prediction output.

Since the inverter's internal operating circumstances, rather than the grid's external values, had a greater impact on actual power output, the regression model placed these quantities inside the inverter. On the other hand, the time-based variables linked to solar production cycles' predictability were given significantly more weight in the classification model (where output was discreteized into production levels).

Reactive power and phase voltage, which are controls on the grid, are not very important and likely do not interact with inverter's active power control loop under typical circumstances because of how static they are. Primarily concerned with maintaining voltage stability than maximizing the conversion efficiency of actual power, the inverter grid support algorithms and power-factor control handle reactive power.

Along with the feature importance findings, future research may incorporate SHAP values and partial dependency plots to enhance model's explainability even further. They facilitate the identification of potential nonlinear and interaction effects and help quantify the marginal contribution of each attribute to the model's predictions. At different operating conditions, this type of analysis may be useful in determining the interaction between time-related quantities (such as the day of the week) and electrical quantities (such as the size of the inverter current). Using PDP or SHAP visualizations offer comprehensive knowledge of model's core logic and improve physical interpretability of data PV system evaluations.

This finding gives credence to the theory that correct representations of the behavior of an inverter require measurements of power at the input side, as well as current dynamics. This finding has implications for future efforts in sensor priority and dimension reduction since it implies that not all followed variables can be equally relevant.

3.3 Anomaly Detection

A red dot indicates an outlier in Figure 7, which shows the 30 kW inverter's actual vs. predicted power production over the course of a week. A orange dashed line provide expected values from the Random Forest model, whereas blue line shows actual measured active power. We used disparity between the two to look for outliers based at threshold we set using prediction error, which could be something like the z-score or the size of the residuals.

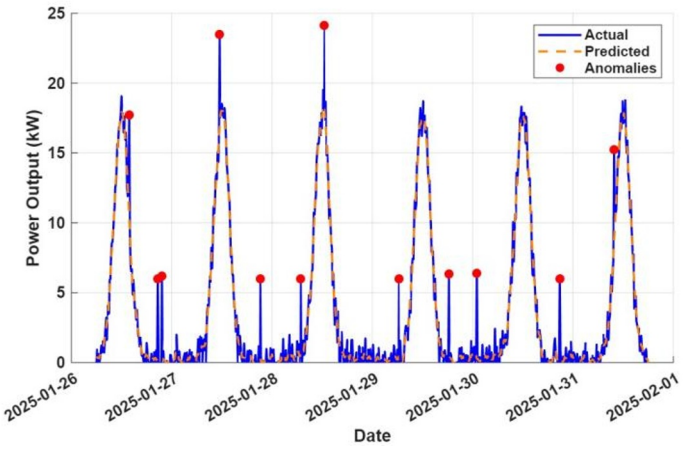
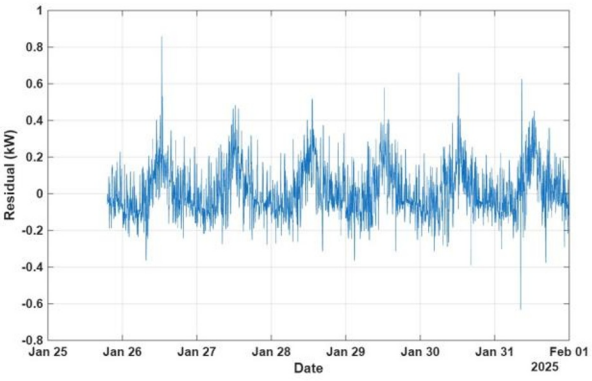


Fig. 7. 30 kW Inverter Anomaly Detection.

Anomalies tend to bunch out around days and times of peak production. Clustering suggests that there are data inconsistencies or there are operational disruptions that affect the behavior of inverter at high output. When the actual values deviate significantly from prediction band which indicates that model has learned common power curve and can handle out-of-the-ordinary trend.



.

Fig. 8. Random Forest residuals centered with minimal bias.

In particular during periods of high irradiance, the Z-score proved useful for anomaly detection by highlighting the short-duration variance of power output relative to the projected power output. There are three groups of possible explanations that can help us make sense of these outliers. (i) Active power drops suddenly and non-statistically normally due to environmental factors like partial shading or sudden clouds. (ii) Instability in the thermal derating, highest temporary grid disconnections, or power point tracking (MPPT), are inverter-related issues that can lead to significant changes. (iii) Data issues, like erroneous synchronization or missing samples among meter records and inverter, can occasionally generate the artificial outliers. A Genetic Algorithm-Neural Network model is constructed in the solar PV inverters to construct an adaptive MPPT control strategy of photovoltaic systems [13].

The study relies just at the statistical deviation (Zscore), it would be more accurate to define the reasons of anomalies by include contextual information like irradiance, temperature, and event log.

By augmenting the traditional SCADA with model-based, anomaly detection mechanism adds valuable layer of monitoring that can help spot minor problems or inefficiencies at an early stage.

Despite the fact that the focus of this work is at statistical anomaly detection than fault linking, there is an intuitive way to classify the discovered deviations into operationally meaningful groups. Abnormalities such as intermittent periods of low output power (sometimes caused by measurement artefacts or transient overproduction) and intermittent periods of high output power (perhaps caused by curtailment events or inverter derating) are examples. Conversely, sensor noise or errors in data synchronization are likely culprits for irregular, high-frequency changes.

It would be feasible to respond more precisely to anomalies if the anomaly analysis pipeline included such classification logic, either in the metadata or by rule-based post Processing. Future studies may incorporate contextual environmental variables, inverter event logs, and adaptive thresholding methods such as Isolation Forest or SHAP-based residual interpretation to distinguish between operational faults and environmental variability. The anomaly detection framework can be extended to categorize deviations into environmental, inverter-related, or data-quality anomalies using rule-based residual clustering or supervised labeling approaches.

3.4 Classification of Random Forest

Random Forest was trained to distinguish between Low, Medium, and High power outputs from the 40 kW inverter, complementing the results of the regression and anomaly detection analysis. These groups were defined based on the percentile levels of the continuous output values (34% and 65%, respectively). In order to test the viability of using supervised classification to identify energy output states—a feature that could be useful in alarm systems or operational mode classifications—we conducted this experiment. The modeling of the effect of urban three-dimensional morphology on photovoltaic energy potential is performed with the machine learning based on a Random Forest Regressor [14].

Figure 9 displays classifier's efficiency confusion matrix on test set. To top it all off, model achieved perfect accuracy in each three areas, and it suited the real labels perfectly every single time.

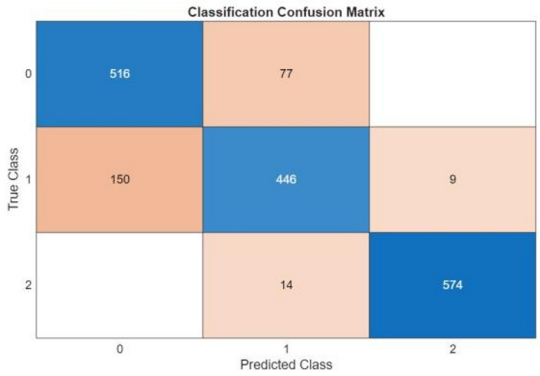


Fig. 9. Classification efficiency of random forest.

This level of precision shows that model has successfully provided nonlinear patterns and time dynamics of the model behavior. Time of day, electronic parameters of the grid and inverters, and environmental inferred from system responses were all part of the input features.

Even though this result is quite predictive, the perfect classification accuracy may be a result of overfitting or label leaking, especially because the features of the associated variables are very important.

It was inevitable that we would doubt the generalizability and robustness of the first section's classification given its perfect precision. In our pursuit of knowledge, we investigated the Random Forest Classifier's feature importances.

After the current-related properties of 30 and 40 kW inverters, model's strongest predictor were total input power of 40 kW inverter. This lends credence to the idea that the classifier's identity is based on a number of related but separate electrical qualities, rather than merely the output label's identity, in order to generate predictions. This further supports the reliability of the first test's accurate classification results.

Looking at the percentages of classes in the test set allowed us to investigate the underlying distribution of labels.

The overall proportion between the classes in the analysis of the difference between classes was noticeable; nearly 15 percent of the samples were in the medium output category and the remainder were in the high output category; more than 60 percent. With the help of the numpy package we were capable of counting the amount of the different classes of prediction, whose results were as follows: This was also in line with the pattern of the classes and that it was more evidence that the model had learned particular patterns by heart and not generalizable pattern. According to the findings presented in Table 2, it means that:Class 0: F1 = 0.980 indicates that the classifier was fairly aware of sluggish periods of manufacturing. Most settings were of high output, which were also the most challenging. The medium-output type (1) was affected by a high number of false positives because the type has a very low precision (0.174) and a comparatively high recall (0.820).

Table 2 Classification outcomes.

	Precision	**f1-score**	**recall**
0.0	0.775	0.87	0.82
1.0	0.83	0.737	0.781
2.0	0.985	0.976	0.98
weighted avg			**0.86**
macro avg	0.863	0.861	0.86

accuracy	0.864	0.86	0.859

Better forecasting is far more challenging and more realistic assessment of the capabilities of the model proves this with a general error drop to 33% and macro-averaged F1 of 0.33 only. To predict, the exercise employed time-shifted labels without defining a lag or rolling features i.e. the model depended upon periodic time hints and existing electrical conditions. Predictors that may serve to enhance time continuity and predictability in subsequent rounds are lag or statistical smoothing predictors. The states of high production cannot be determined relying on electrical measurements only because of the sensitivity of the high-production conditions to the short-term variability of irradiance and transient clouds. There were time proxies (hour of day and day index) in the model to offset the unavailability of actual contributions in the environment (sunlight intensity, temperature, or sky images). The high and the medium level of output are not much separable because these two features are very good in depicting the trend of the day rather than the rapidly shifting atmospheric dynamics.

3.5 Robustness under Time-Aware Cross-Validation

A5-fold Time Series Split on Random Forest classifier is employed in order to ensure that the bad hold-out was not due to a one-time split over time. Overall, the model remained consistent across folds when class weighting alone. The accuracy ranged from 0.23 to 0.11, the macro F1 was 0.21to 0.11, and the balanced accuracy was 0.864 (across the range) for 83 of the data points. The macro F1 was 0.86, and the out-of-fold accuracy was 0.859, with a range of 0.12. Training folds that used SMOTE oversampling also had similar results (accuracy = 0.23, macro F1 = 0.21). By taking class imbalance and temporal dependencies into account, the classifier improves its cross-validated performance while decreasing confidence. This suggests that the classifier is generalizable. As reported in previous studies, supervised learning methods in predictive modeling of photovoltaic power generation is used to improve the stability of solar PV inverter in intermittent environments. The highest accuracy and the lowest Mean Absolute Error signs (MAPE 2.2790%, RMSE 0.8792%), the best forecasting result were with the Random Forest Regressor, which is one of the many algorithms [15].The 80/20 chronological split preserves temporal integrity and reflects real-world forecasting deployment scenarios. Unlike random splits, it avoids information leakage and ensures that future states are predicted strictly from historical observations. Validation was performed using 5-fold Time Series Split cross-validation, ensuring no temporal overlap between training and validation folds. Both class-weight balancing and SMOTE were evaluated to mitigate imbalance effects.

Table 3 Classification outcomes.

Validation Setting	**Balanced Acc**	**Macro F1**	**Accuracy**	**Notes**
SMOTE in-fold	0.12	0.21	0.23	5-fold Time Series Split

class_weight	0.11	0.12	0.22	5-fold Time Series Split
Single hold-out	0. 11	0.11	0.11	80/20 temporal split

When it comes to PV monitoring in industrial settings, other architectures like as LSTM and XG Boostnetworks somewhat more accurate when subjected to continuous-sequence data or large-scale. However, their lack of transparency and complexity can make them difficult to deploy. The goal of the study was to find a way to make lightweight and scalable diagnostics for operational systems, and Random Forest provide computationally and interpretable alternative. Future research will investigate hybrid explainable AI frameworks incorporating SHAP analysis, seasonal dataset expansion, and comparison with sequential deep learning models such as LSTM. Additionally, multi-plant validation will be conducted to evaluate scalability across different inverter capacities.

4. Conclusion

This study contributes to PV forecasting research by demonstrating that high-accuracy inverter prediction and anomaly detection can be achieved without meteorological sensors, using an interpretable ensemble-based framework validated under strict time-aware evaluation.Photovoltaic systems that are grid-connected need good and reliable monitoring solutions, traditionally the solutions have been based on meteorological sensors and complicated rule-based or deep learning models but this restricts scalability due to sensor constrained settings. In order to overcome this issue, an electrical data on a 30 kW and 40 kW inverter PV plant was used, which contains five minutes resolution active power, reactive power, phase voltages, phase currents, and time-related operational characteristics, obtained after one month. Although validated on 30 kW and 40 kW inverters, the methodology is capacity-agnostic since it relies on normalized electrical features. Therefore, the framework is transferable to other PV plants with similar grid-connected architectures, subject to retraining under plant-specific operational data.The combination of Lightweight Frameworks were developed including the Random Forest regression, Random Forest classification and Z-score-based statistical anomaly detection, where the temporal validation was well performed with 80/20 chronological split and 5-fold TimeSeries cross-validation to avoid leakage of data. The regression model was found to be highly predictive with MAE = 0.12 kW and a predictive coefficient of R^2 = 0.995, and the cross-validation showed a predictive coefficient of R^2 = 0.9802 and deviation of MAE = 0.1024 kW indicating robustness, observed abnormal samples were 2.77% of the total fatigue samples, and realistic time-aware classification showed that the regression model obtained 33% forecasting accuracy (macro F1 = 0 The suggested interpretable and computationally efficient framework allows real-time inverter health monitoring, anomaly detection at an early stage, predictive maintenance scheduling, and scalable PV plant monitoring without using external environmental sensors, thus making it feasible to implement in distributed renewable energy systems.0.1.

References

[1] V. Raj, S.-Q. Dotse, M. Mathew, M. I. Petra, and H. Yassin, "Ensemble Machine Learning for Predicting the Power Output from Different Solar Photovoltaic Systems," *Energies (Basel).*, vol. 16, no. 2, 2023, doi: 10.3390/en16020671.

[2] S. Han and J. Li, "Improving efficiency and stability of improved circular solar photovoltaic structures via multi-directional functionally graded materials: A computer simulation validated by hybrid machine learning algorithm and experimental datasets," *Mater. Today Commun.*, vol. 46, 2025, doi: 10.1016/j.mtcomm.2025.112816.

[3] M. Tran *et al.*, "S3Former: A Deep Learning Approach to High Resolution Solar PV Profiling," *IEEE Trans. Smart Grid*, vol. 16, no. 3, pp. 2611–2623, 2025, doi: 10.1109/TSG.2025.3531764.

[4] O. Z. Lin, L. Štěpanec, E. Koutroulis, D. Juchelkova, and H. Y. Aye, "Optimizing monthly solar PV tilt angles and energy yield across global climate zones: A hybrid machine learning and PVLib approach," *Renew. Energy*, vol. 260, 2026, doi: 10.1016/j.renene.2025.125163.

[5] L. Guanghua, S. H. H. Shah, G. E. M. Abro, S. C. Jiang, and R. Arshad, "Drone-Based Random Forest Classifier for Intelligent Dust Monitoring on Solar PV Systems in Saudi Arabia," *Arab. J. Sci. Eng.*, 2026, doi: 10.1007/s13369-025-11037-5.

[6] G. E. Moon and M. J. Kim, "Estimating the impact of PM2.5 on solar power with machine learning: Evidence from South Korea," *Energy Econ.*, vol. 153, 2026, doi: 10.1016/j.eneco.2025.109071.

[7] V. Khandeparkar and S. K. Senthil Kumar, "Effectiveness of supervised machine learning models for electrical fault detection in solar PV systems," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-18802-4.

[8] R. C. Yadaw, A. K. Dewangan, L. Nagdeve, A. K. Yadav, and A. Ahmad, "Enhancing the power quality of a grid-connected rooftop solar PV system under varying environmental conditions: Strategic optimization of influencing parameters using machine learning and desirability-driven approach," *Sustainable Energy Technologies and Assessments*, vol. 83, 2025, doi: 10.1016/j.seta.2025.104620.

[9] A. K. Das, H. Mohapatra, S. R. Mishra, and S. S. Sahoo, "Exploring the synergistic potential of a hybrid PV-biogas power generation system for smart city electrification by sustainable thermo-exergetic and environmental analysis using a forest machine learning approach," *Environmental Science and Pollution Research*, vol. 32, no. 50, pp. 28824–28839, 2025, doi: 10.1007/s11356-025-37237-y.

[10] M. V Prashanth *et al.*, "Machine Learning Approaches for Solar PV Fault Identification," *SN Comput. Sci.*, vol. 6, no. 7, 2025, doi: 10.1007/s42979-025-04364-9.

[11] A. Syamsuddin, A. C. Adhi, A. Kusumawardhani, T. Prahasto, and A. Widodo, "Predictive maintenance based on anomaly detection in photovoltaic system using

SCADA data and machine learning," *Results in Engineering*, vol. 24, 2024, doi: 10.1016/j.rineng.2024.103589.

[12] G. S. Budhi, Y. Tanoto, D. Jovian, R. Adipranata, and C. Raphael, "Solar Photovoltaic Power Output Prediction Using Machine Learning-Based Regressors," *Journal of Asian Energy Studies*, vol. 9, pp. 111–130, 2025, doi: 10.24112/jaes.090007.

[13] I. Chtouki, P. Wira, M. Zazi, H. E. Chakir, S. A. A. D. motahhir, and K. Choukri, "Maximum Power Point tracking implementation based on self-learning adaptive GA-Neural controller for standalone PV applications," *Results in Engineering*, vol. 26, 2025, doi: 10.1016/j.rineng.2025.104587.

[14] S. Lu *et al.*, "Comprehensive benefits evaluation of the impact of vertical city on solar PV utilization for achieving smart sustainable cities," *Sustain. Cities Soc.*, vol. 137, 2026, doi: 10.1016/j.scs.2026.107138.

[15] A. Verma, K. G. Upadhyay, and M. M. Mohan Tripathi, "Development of Artificial Intelligence Techniques for Solar PV Power Forecasting for Dehradun Region of India," *Journal of Electrical Systems*, vol. 17, no. 3, pp. 324–337, 2021, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116025724&partnerID=40&md5=bfc1cb577dd623796a05b4143bfe542b