Enhancing Solar Power Plant Efficiency through Forecasting and Anomaly Detection Using Deep Learning

ECE657 Probabilistic Machine learning, Spring 2025
Yu-che Cheng
G01478897
ycheng26@gmu.edu

1. Introduction

As the world increasingly shifts toward renewable energy, solar power has become one of the most promising and widely adopted sources. However, despite its growing presence, managing solar power generation isn't without its challenges. The output from solar panels can vary greatly depending on weather conditions, the cleanliness of the panels, and the performance of the equipment. This variability makes it difficult for operators to forecast energy production and maintain consistent system performance.

In this project, I aim to tackle these issues by applying machine learning techniques to real-world solar plant data. Our study uses operational data collected from two solar power plants in India over a span of 34 days. The dataset includes inverter-level power generation data and sensor readings that reflect overall environmental conditions.

I focus on three key areas to improve the efficiency and reliability of solar power plants:

I. Power Forecasting – I test three deep learning models—MLP, LSTM, and Transformer—to predict short-term solar power output over the next two days. Accurate forecasting is crucial for better grid integration and energy planning.
II. Maintenance Detection (Panel Cleaning Needs) – Solar panel efficiency can drop when dust or debris accumulates. By analyzing discrepancies between irradiation and actual output, I aim to identify when panels likely need cleaning—without waiting for visible signs.
III. Faulty Equipment Identification – I also look for signs of malfunctioning equipment by spotting consistent mismatches between expected and actual AC power output. This helps operators detect problems early and avoid prolonged inefficiencies.

Ultimately, our goal is to develop tools that support solar plant operators in making more informed decisions. By improving predictions and proactively identifying issues, these methods can reduce maintenance costs, extend equipment life, and make solar power more reliable and cost-effective.

2. Background and Related work

Solar power has gained prominence as a sustainable energy source, but its variability presents challenges for grid integration and operational efficiency [1]. This section examines relevant research in power forecasting, maintenance needs detection, and equipment monitoring for solar installations.

I. Power Forecasting Techniques

Solar power forecasting has evolved from physical models based on numerical weather predictions to advanced machine learning approaches [2]. Traditional statistical methods often struggle with the non-linear relationships between weather variables and power output [3]. Deep learning has emerged as the state-of-the-art approach, with several architectures showing promise. Multi-Layer Perceptrons (MLPs) perform well for point forecasting but lack inherent capabilities for sequence modeling [4]. Long Short-Term Memory (LSTM) networks have become widely adopted for their ability to capture temporal dependencies in generation patterns [5]. More recently, Transformer models have been applied to energy forecasting, leveraging self-attention mechanisms to model global dependencies without sequential constraints [6].

II. Maintenance Detection and Anomaly Identification

Panel soiling and equipment degradation can significantly impact system efficiency, with studies showing soiling alone can reduce performance by 5-30% [7]. Traditional maintenance relies on scheduled cleaning and visual inspections, which may not optimize resource allocation [8].Data-driven approaches have shown promise for more efficient maintenance scheduling. Unsupervised learning techniques, particularly clustering algorithms like DBSCAN, have gained attention for identifying anomalies without requiring labeled training data [9].

These methods establish baseline performance patterns and flag deviations that may indicate maintenance needs or equipment issues. Temperature correlation analysis has emerged as an important dimension in fault diagnosis, helping distinguish between environmental effects and actual equipment problems [10]. Similarly, frequency analysis of anomalies—examining their temporal distribution and recurrence—provides valuable context for maintenance prioritization [11].

III. Integrated Approaches

Recent research has begun to explore integrated frameworks that combine forecasting with anomaly detection for comprehensive plant management [12]. These approaches enable not only prediction of future generation but also proactive identification of maintenance needs and equipment issues. Our work builds upon these foundations by integrating state-of-the-art forecasting models with a novel hybrid anomaly detection framework. Unlike previous approaches that rely solely on clustering or statistical thresholds, our methodology combines these techniques with temperature correlation and frequency analysis to provide a more nuanced understanding of system health and maintenance requirements.

3. Statistical Problem

I. Power Generation Forecasting

- Objective:
  Quantify the predictive relationship between environmental conditions and short-term solar power output at the inverter level.
- Statistical Model:
  I treat this as a multivariate time series regression problem:
  $$P_t = f(I_t, T_t, T_{model}, P_{t-1}, \dots, P_{t-k}) + \varepsilon_t$$
  Where P_t is the AC power output at time t, I_t, T_t and T_{model} represent solar irradiation, ambient temperature, and module temperature respectively; and ε_t captures model error.
- Hypotheses:
  H_0: No significant difference in predictive accuracy across model types (MLP, LSTM, Transformer).
  H_a: Model performance differs significantly, with sequential models (LSTM, Transformer) outperforming MLP due to their ability to model temporal dependencies.
- Evaluation Metrics:
  Model performance is assessed using Mean Absolute Error (MAE) and R^2. These act as test statistics to compare predictive fit and explanatory power across model types.

II. Maintenance Detection via Anomaly Identification

- Objective:
  Detect inverter-level underperformance likely due to issues such as panel soiling, using deviations from expected power output.
- Statistical Model:
  This is framed as a semi-supervised anomaly detection task:
  $$\delta_{i,t} = \frac{P_{i,t} - \hat{P}_{i,t}(I_t)}{\sigma_i}$$
  Where $\delta_{i,t}$ is the standardized deviation for inverter i at time t, $\hat{P}_{i,t}(I_t)$ is expected power conditioned on irradiation, and σ_i is the standard deviation under normal operation.
- Hypotheses:
  H_0: Deviations are normally distributed and centered around expected output, with no clustering of negative anomalies.
  H_a: Deviations exhibit statistically significant negative clustering, suggesting non-random underperformance.
  I apply DBSCAN to identify statistically dense clusters of low $\delta_{i,t}$, with hyperparameters ε and min_samples acting as decision thresholds.

III. Equipment Performance Monitoring

- Objective:
  Identify inverter-specific equipment faults through anomaly frequency and environmental correlation.
- Statistical Model:
  This is approached via temporal event analysis and correlation testing:
  $$F_i(d) = \sum_{t \in d} 1\big(\delta_{i,t} < -\tau\big),$$
  $$C_i = corr(\delta_{i,t}, |T_t - \bar{T}|)$$
  Here, $F_i(d)$ is the daily frequency of significant negative deviations for inverter i, and C_i captures the correlation between deviations and ambient/module temperature departures from the mean.
- Hypotheses:
  H_0: Anomalies are randomly distributed and uncorrelated with temperature.
  H_a: Statistically significant temporal clustering $(F_i(d) \geq 3)$ and/or correlation with temperature deviations indicate equipment faults.

IV. Unified Statistical Framework

Collectively, these components form an integrated probabilistic framework. Each task—forecasting, anomaly detection, and equipment diagnosis—is

modeled via statistical constructs that enable both explanatory insight and operational decision-making. This layered approach allows us to distinguish:

- Random variance (e.g., sensor noise, transient fluctuations),
- Systematic but recoverable issues (e.g., panel soiling),
- Persistent equipment failures (e.g., inverter degradation).

By combining residual analysis, clustering methods, and correlation tests, I build a statistically principled methodology for proactive solar plant management.

4. Dataset

I. Overview

This study is based on operational data collected from two solar power plants located in India, referred to as Plant 1 and Plant 2. The datasets encompass both power generation metrics and environmental sensor readings, allowing for a comprehensive analysis of system performance, environmental influence, and potential equipment anomalies. These data are particularly valuable for developing predictive models and diagnostic tools to support solar plant management.

II. Data Collection Methodology

Two main types of data were collected using different strategies:

- Power Generation Data: Measured at the inverter level, where each inverter is connected to several solar panel arrays. This high-resolution data provides insight into localized inverter performance and makes it possible to detect operational inefficiencies or faults.
- Environmental Sensor Data: Collected via a centralized sensor array installed at each plant. These sensors measure ambient conditions—such as temperature and solar irradiance—that directly affect energy generation.

III. Dataset Details

Plant 1:

- Time Range: May 15, 2020 – June 17, 2020
- Number of Inverters: 22
- Power Generation Data
- Sampling Frequency: Every 15 minutes
- Key Features:
  - DATE_TIME: Timestamp of each measurement
  - PLANT_ID: Plant identifier
  - SOURCE_KEY: Unique ID for each inverter
  - DC_POWER: Amount of DC power generated by the inverter
  - AC_POWER: Alternating current power output (kW)
  - DAILY_YIELD: Energy generated on the current day (kWh)
  - TOTAL_YIELD: Total energy generated since inverter installation (kWh)
- Weather Sensor Data
- Key Features:
  - DATE_TIME: Timestamp of each measurement
  - PLANT_ID: Plant identifier
  - SOURCE_KEY: Sensor array identifier
  - AMBIENT_TEMPERATURE: Ambient air temperature (°C)
  - MODULE_TEMPERATURE: Surface temperature of the panels (°C)
  - IRRADIATION: Solar irradiance (W/m²)

Plant 2

- Time Range: May 15, 2020 – June 17, 2020
- Number of Inverters: 22
- Power Generation Data
- Sampling Frequency: Every 15 minutes
- Key Features: Same as for Plant 1
- Weather Sensor Data
- Key Features: Same as for Plant 1

IV. Data Characteristics and Considerations

The two datasets differ in their time spans: Plant 1 covers nearly 11 months, capturing seasonal variations in environmental conditions and generation performance, while Plant 2 covers about one month, offering more targeted insight during a specific period.

Notably, the generation data is sampled more frequently than the environmental sensor data. This reflects the relatively stable nature of atmospheric variables compared to the dynamic changes in power output, and helps balance the trade-off between data resolution and storage/processing demands.

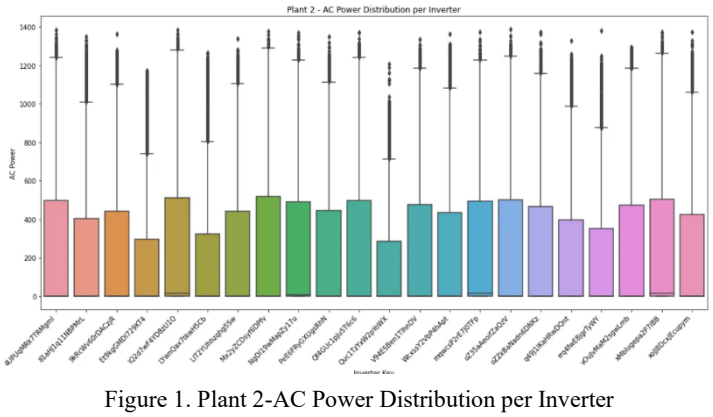

Figure 1. Plant 2-AC Power Distribution per Inverter

V. Relevance to Research Objectives

This dataset is well-suited to the study's three research objectives:

- Short-term Power Forecasting: The high-frequency historical data provides the basis for building and evaluating time-series forecasting models such as MLP, LSTM, and Transformer.
- Maintenance Needs Detection: By comparing irradiation levels with AC power output across inverters, I can detect anomalies that may indicate soiling, degradation, or other issues requiring maintenance.
- Equipment Performance Monitoring: Inverter-level granularity enables side-by-side comparison of similar units under similar environmental conditions, facilitating early detection of equipment underperformance.

5. Exploratory Data Analysis

To better understand the operational dynamics of the two solar power plants, I performed an exploratory data analysis (EDA) focusing on inverter-level performance and the temporal relationship between solar irradiation and power generation. This analysis aimed to identify operational inconsistencies, assess equipment performance, and generate insights that would later inform our modeling approach.

I. Inverter-Level Power Distribution

Our analysis began with an examination of the distribution of AC power output across inverters at each plant. Boxplots were generated to visualize performance patterns and variability. For Plant 2, the boxplot (Figure 1) reveals a notable degree of variation among inverters. Although most inverters reach a similar peak AC power—typically between 1300 and 1400 kW—there is significant divergence in their median outputs. Some inverters register median power around 500 kW, while others fall closer to 300 kW. One inverter in particular, labeled "EISXo6m9U7Z6H4," shows a considerably lower median output compared to the rest, raising the possibility of equipment malfunction or differing operational constraints. Additionally, the interquartile ranges differ noticeably between inverters, suggesting varying levels of consistency in performance throughout the day.

In contrast, the inverter data from Plant 1 (Figure 2) exhibits a much more uniform distribution. Median power output across inverters generally falls within the 550 to 650 kW range, and maximum power values consistently hover around 1400 kW. This suggests a high level of standardization in equipment behavior. However, one inverter—identified as "ibdGMCBZB4Q5Try"—appears to underperform

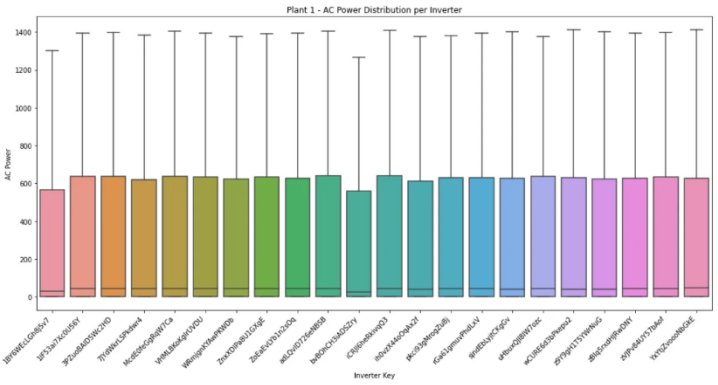

Figure 2. Plant 1-AC Power Distribution per Inverter

relative to its peers, with a noticeably lower median output. Even so, the overall uniformity in Plant 1 indicates more homogeneous operating conditions or perhaps better maintenance practices compared to Plant 2.

The contrasting performance profiles between the two plants suggest that while Plant 1 operates under relatively stable and consistent conditions, Plant 2 may be facing equipment-level issues or environmental inconsistencies affecting its overall efficiency. This variability highlights the importance of granular inverter-level monitoring to identify potential underperformance early and address maintenance needs proactively.

II. Temporal Patterns of Power Generation and Irradiation

In addition to inverter performance, I investigated the temporal behavior of AC power generation in relation to solar irradiation. This comparison provides insight into how closely the power output of each plant aligns with incoming solar energy and reveals patterns that may indicate operational anomalies.

Figure 3 illustrates the daily power and irradiation profile for Plant 2 on May 27, 2020. As expected, power generation begins around sunrise at 06:00, peaks near noon, and declines toward sunset. While the overall trend matches the irradiation curve, the power output shows a distinctive saw-tooth pattern during the peak

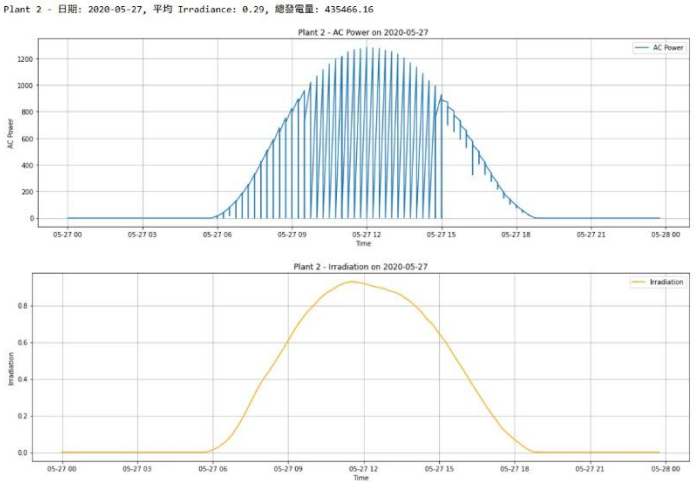

Figure 3. Plant 2-AC Power and Irradiation on 2020-05-27

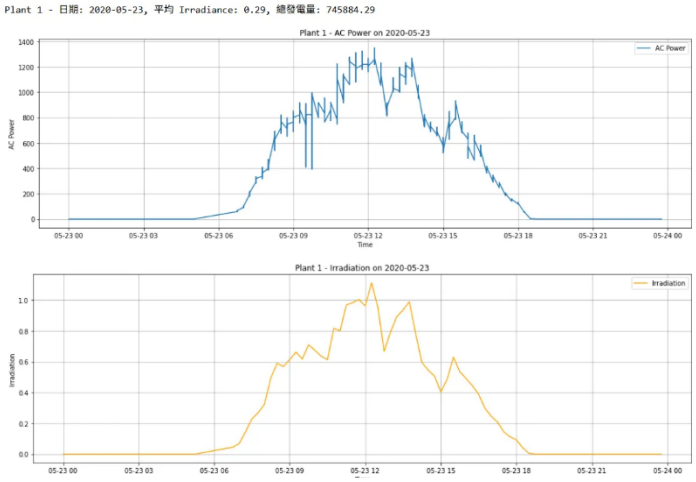

Figure 4. Plant 1-AC Power and Irradiation on 2020-05-23

hours. These regular, sharp fluctuations—characterized by sudden drops followed by quick recoveries—do not correspond with any variations in the irradiation data, which remains smooth and bell-shaped throughout the day. The total power generated on this day was 435,466.16 kW, with an average irradiation of 0.29. The mismatch between the steady irradiation and the fluctuating power output raises concerns about possible inverter-level control issues, intermittent faults, or system inefficiencies.

In comparison, the daily profile from Plant 1 on May 23, 2020 (Figure 4), reveals a closer relationship between solar irradiation and AC power generation. Although the irradiation curve exhibits fluctuations—likely due to cloud cover—the power output closely mirrors these variations. Notable dips around 09:30 and 12:00 align in both irradiation and power, confirming the influence of atmospheric conditions on system performance. Remarkably, despite having the same average irradiance of 0.29 as the Plant 2 example, Plant 1 generated significantly more energy, with a total output of 745,884.29 kW. This stark difference suggests that Plant 1 is converting solar energy more efficiently, further supporting the idea that Plant 2 may be experiencing performance degradation or suboptimal operation.

III. Insights and Working Hypotheses

Based on these observations, I formulated several hypotheses regarding system behavior and equipment performance. The consistent and predictable patterns in Plant 1's data suggest that it operates under normal conditions, serving as a reliable baseline for modeling and performance evaluation. On the other hand, the irregular power generation patterns in Plant 2, especially the midday oscillations not reflected in the irradiation data, suggest that this plant may be experiencing system-level issues. These could include soiled panels, inverter malfunctions, or misconfigured control systems.

Moreover, the observed differences in inverter-level output between the two plants reinforce the notion that Plant 2 suffers from higher performance variability. The presence of inverters with significantly lower median output implies localized problems that could potentially be resolved through targeted maintenance. The power-irradiation relationship also appears weaker in Plant 2, further highlighting potential inefficiencies. Despite experiencing identical irradiation levels on the observed days, Plant 1 produced substantially more energy, pointing to better energy conversion efficiency.

IV. Implications for Modeling Strategy

The insights gained from this exploratory analysis directly informed the design of our modeling approach. Since Plant 1 data demonstrates stable and consistent behavior, it was selected as the sole training dataset for our forecasting models, which include a Multi-Layer Perceptron (MLP), Long Short-Term Memory network (LSTM), and Transformer model. This decision ensures that the models are trained on data reflecting optimal operating conditions rather than learning from anomalous or potentially faulty behavior.

For maintenance detection, I incorporated data from both plants to capture a wide range of operating scenarios. By analyzing deviations in the expected power-to-irradiation relationship, I aimed to detect signs of equipment soiling, degradation, or failures. Plant 2's fluctuating patterns provided particularly useful examples for developing and testing these algorithms. Similarly, the inverter-level data from both plants was used to build an equipment monitoring framework, allowing us to detect underperforming units by comparing their outputs under similar environmental conditions.

6. Methodology

This section outlines the methodological approaches employed to address our three primary research objectives: power generation forecasting, maintenance needs detection, and equipment performance monitoring.

I. Data Preprocessing

Comprehensive data preprocessing was performed to ensure data consistency and quality across datasets. First, I aligned all timestamps to datetime format and synchronized the time series between inverter-level power generation and plant-level weather sensor data. These datasets were then merged based on timestamps and plant identifiers to create a unified dataset combining environmental and generation variables.

To enhance temporal pattern recognition, I engineered features such as hour of day, day of week, and day of year. Lag features of previous AC and DC

power readings were also constructed to capture autoregressive properties. The dataset was then split into training (80%) and testing (20%) sets while maintaining a stratified distribution of inverters. Feature scaling was performed using MinMaxScaler, and categorical inverter identifiers were one-hot encoded to help models learn inverter-specific behaviors. For sequence-based models, input sequences with a sliding window of 4 time steps were constructed.

Given that Plant 1 exhibited normal operational characteristics based on EDA, its data was used exclusively for training power forecasting models. In contrast, data from both plants were used for anomaly detection and equipment monitoring to maximize diversity and capture irregular patterns.

II. Power Generation Forecasting

To forecast short-term power output, I implemented three deep learning architectures: a Multilayer Perceptron (MLP), a Long Short-Term Memory (LSTM) network, and a Transformer model. Each was designed to predict both DC and AC power using historical data and environmental variables.

- Multilayer Perceptron (MLP):

  The MLP is a feedforward neural network designed to model non-linear relationships. Given input vector $x \in \mathbb{R}^n$, the MLP maps it to a predicted power output $\hat{y}$ through two hidden layers:

$$h_1 = ReLU(W_1 x + b_1)$$
$$h_2 = ReLU(W_2 h_2 + b_2)$$
$$\hat{y} = W_3 h_3 + b_3$$

  The model was trained using the Mean Squared Error (MSE) loss:

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

  with the Adam optimizer for 300 epochs.

- Long Short-Term Memory (LSTM)

  The LSTM model captures temporal dependencies through gated memory units. At each time step , the LSTM cell computes:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$$
$$i_t = \sigma(W_i x_i + U_i h_{t-1} + b_i)$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$h_t = o_t \odot tanh\ (c_t)$$

  This structure allows the model to remember or forget information selectively. The model was trained over 50 epochs using the MSE loss and the Adam optimizer.

- Transformer Model

  The Transformer model utilizes self-attention to model global dependencies. For input sequences $Q, K, V \in \mathbb{R}^{n \times d}$, attention is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

  Our Transformer Encoder included multi-head attention (2 heads), layer normalization, and position-wise feedforward networks. Outputs were aggregated via global average pooling and passed through dense layers for final power prediction. This architecture enables the model to dynamically focus on relevant time steps.

III. Maintenance Needs and Equipment Performance Monitoring

For maintenance and performance monitoring, I implemented an anomaly detection framework based on clustering, statistical deviation, and temporal correlation.

I applied the DBSCAN algorithm to detect outliers in the irradiation vs. power space. DBSCAN groups data points into clusters where points within a neighborhood radius $\varepsilon = 0.1$ and a minimum number of points $\min\_samples = 3$ form dense regions. Points not assigned to any cluster are flagged as anomalies.

To correct over-detection, I implemented a recovery step using linear regression to model expected power given irradiation. If a point's power exceeded the regression prediction by more than 2 standard deviations, it was reclassified as normal.

Since high power generation is not inherently problematic, a hybrid recovery mechanism was implemented using both linear regression and standard deviation approaches. This process involves fitting a linear regression model to the normal data points and identifying points that exceed expected power output by a certain threshold, then reclassifying them as non-anomalous.

Anomalies are correlated with temperature patterns by examining ambient and module temperatures within a 1-hour window around each anomaly. This helps identify whether anomalous power output coincides with temperature fluctuations that might indicate equipment issues or environmental factors.

To identify persistent issues rather than transient anomalies, frequency analysis identifies dates with anomaly concentrations. Days with at least three anomalies are flagged as having frequent anomalies, potentially indicating systematic issues rather than random fluctuations.

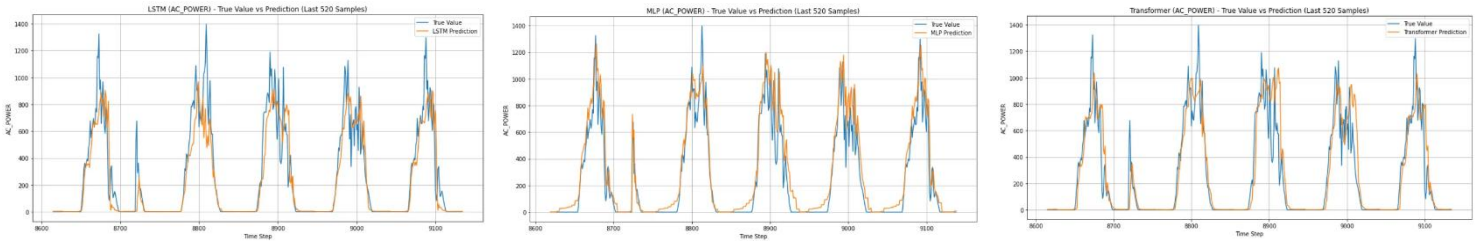

Figure 5. True Value vs Prediction (Last 520 Samples) of LSTM, MLP and Transformer

Final significant anomalies are identified as either frequent anomalies or those correlated with temperature anomalies. This combined approach helps distinguish between different types of issues requiring attention.

This comprehensive anomaly detection approach enabled us to distinguish between different types of anomalies. Panels requiring cleaning appear as consistent underperformance across multiple days, detected through the frequency analysis component. These anomalies typically show gradual degradation patterns without accompanying temperature anomalies. Faulty equipment manifests as persistent anomalies with distinct patterns in the irradiation-power relationship, often accompanied by temperature anomalies. These issues are identified through the combination of clustering-based detection and temperature correlation analysis.

IV. Evaluation Metrics

Model performance was evaluated using:

- Mean Absolute Error (MAE): $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$ A direct measure of average prediction error.
- R-squared score: $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$ This quantifies the proportion of variance explained by the model.

Anomaly detection was assessed using:

- Visual Validation: Through scatter plots to confirm spatial and temporal alignment of anomalies.
- Anomaly Counts: Categorizing the types and recurrence of anomalies to identify patterns and prioritize interventions.

These methodologies formed the foundation for reliable forecasting and actionable maintenance insights.

7. Results

I. Results and Visual Analysis of Power Forecasting Models

	MAE	R^2
LSTM	78.1781	0.8008
MLP	83.2889	0.8433
Transformer	86.0071	0.7661

Table 1. Model Performance Metrics for AC Power Prediction

The evaluation of our power generation forecasting models was conducted using both quantitative metrics and visual inspection to provide a holistic understanding of model performance. Table 1 presents the core metrics—Mean Absolute Error (MAE) and coefficient of determination (R^2)—while Figures 1 through 6 offer a visual comparison of predicted versus actual values through time series plots and scatter distributions.

Table 1 presents the performance metrics of the three deep learning models implemented for power generation forecasting. The evaluation was conducted on the test dataset, which comprised 20% of the available data from Plant 1. Two metrics were used to assess model performance: Mean Absolute Error (MAE) and coefficient of determination (R^2).

The LSTM model demonstrated the best overall performance with an MAE of 78.1781 and an R^2 value of 0.8008. While this MAE value indicates an average deviation of approximately 78 units from the actual power values, the model successfully explains about 80.08% of the variance in the power generation data.

The MLP model showed slightly lower performance with an MAE of 83.2889 and an R^2 of 0.8433. Interestingly, despite having a higher MAE than the LSTM model, the MLP achieved a better R^2 value, suggesting that while its absolute prediction errors might be larger on average, it captures the overall trend and variability of the data more effectively in some cases.

The Transformer model exhibited the weakest performance among the three architectures, with the highest MAE of 86.0071 and the lowest R^2 of 0.7661, explaining approximately 76.61% of the variance in the power generation data.

Visualizing the predicted and actual power values over time reveals additional patterns not immediately evident through metrics alone. The LSTM model demonstrates strong temporal alignment with the actual generation curve. It effectively captures the characteristic daily power generation cycle, including the morning ramp-up, midday peak, and evening decline. However, the model tends to slightly underestimate the true peak values, particularly during periods of maximum sunlight, and smooths over some

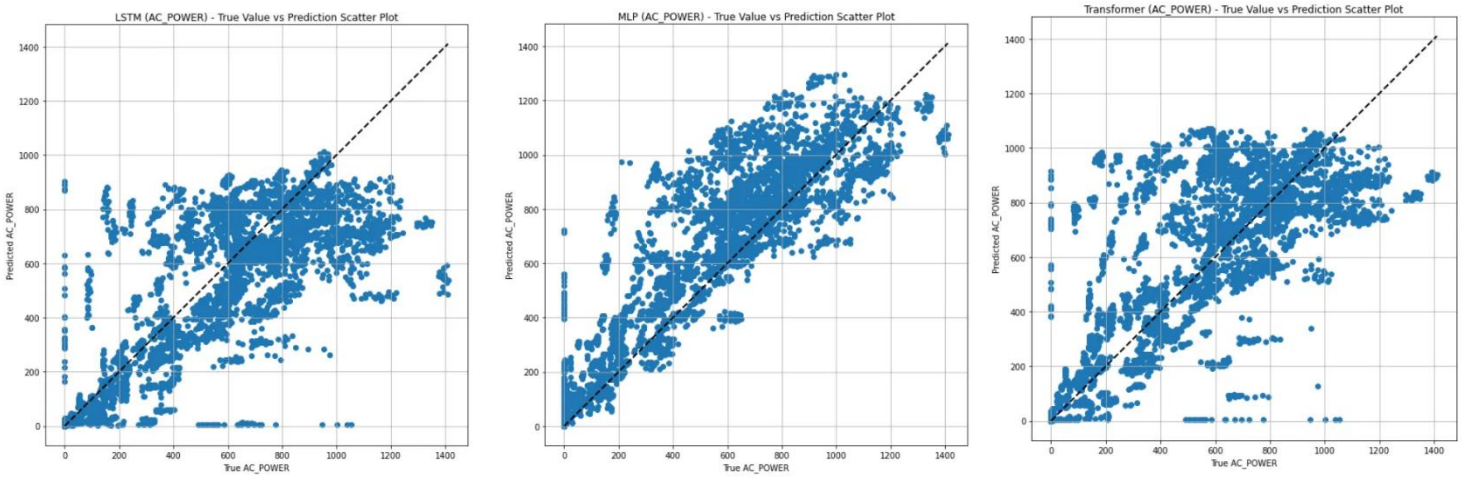

Figure 6. True Value vs Prediction Scatter Plot of LSTM, MLP and Transformer

of the short-term fluctuations—a consequence of its recurrent structure.

The MLP model shows a more reactive behavior in contrast. While it broadly follows the daily power pattern, it sometimes overshoots during peak periods, suggesting a greater sensitivity to sudden changes. This higher responsiveness occasionally results in overprediction or false positives, especially during time intervals when actual generation is close to zero, such as during early morning hours.

The Transformer model, though architecturally designed to capture complex dependencies through attention mechanisms, presents a more conservative prediction profile. It generally aligns with the correct timing of daily generation events but, like LSTM, tends to underestimate power during peak hours. Moreover, the model appears less capable of capturing sharp fluctuations in power output, leading to smoother and sometimes overly stable forecasts.

Taken together, these visual and quantitative evaluations highlight the distinct behaviors and trade-offs of each model. The LSTM offers the best balance in terms of average prediction accuracy and overall pattern tracking, though it smooths out short-lived variations. The MLP captures variability well and reacts swiftly to changes, but with a cost to stability and precision during non-peak periods. The Transformer model, despite its potential for modeling long-range dependencies, showed weaker performance under the specific constraints of this dataset and prediction task, likely due to the relatively short time window and highly structured solar generation patterns.

Overall, the combination of metric-based evaluation and visual inspection not only validates the superiority of the LSTM and Transformer in this scenario but also underscores the value of each architecture's unique strengths. These insights may be valuable for future hybrid modeling approaches, where models with complementary capabilities could be combined to improve both accuracy and adaptability in real-world solar power forecasting applications.

II. Anomaly Detection for Maintenance Needs and Equipment Performance Monitoring

To better understand the operational health of the two solar plants, I employed a hybrid anomaly detection framework based on an enhanced DBSCAN algorithm. This method effectively revealed distinct behavioral patterns between the two plants, offering valuable insights into potential equipment faults and maintenance priorities. In this section, I highlight both the quantitative and qualitative findings, focusing on how these patterns reflect the underlying performance differences between Plant 1 (figure 7) and Plant 2 (figure 8).

For Plant 1, the anomaly detection was conducted on Inverter ZnxXDlPa8U1GXgE. The scatter plot of its output shows a strong linear correlation between solar irradiation and AC power, with most data points closely aligned along the regression line. This consistency indicates that the inverter is operating under stable conditions with predictable performance. While some anomalies were detected, they appeared infrequently and were concentrated in specific regions—particularly in the mid-irradiation range (0.4 to 0.6), where power output occasionally dipped below expectations, and in the high-irradiation range (0.8 to 1.0), where a few outliers diverged from the expected trend. These deviations suggest occasional inefficiencies but not systemic issues. One of the strengths of the hybrid method was its ability to recover a number of initially flagged anomalies. Many of these recovered points—now identified as high-performance outputs—were located above the regression line, especially under high irradiation. This recovery process helped prevent misclassification of strong-performing data, ensuring that temporary fluctuations were not mistaken for system faults. Additionally, temperature-related anomalies were minimal, suggesting that environmental factors like heat had little influence on inverter behavior. The scarcity of frequent anomalies—those occurring in clusters over several days—further supports the

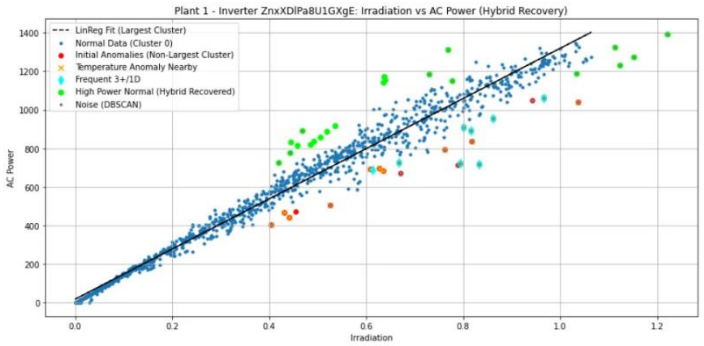

Figure 7. Anomaly Detection of Inverter from Plant 1

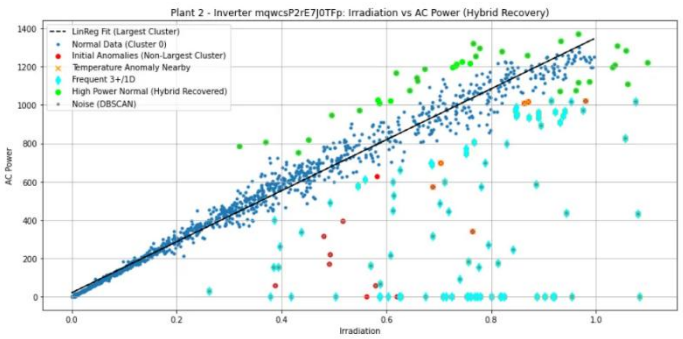

Figure 8. Anomaly Detection of Inverter from Plant 2

interpretation that Plant 1 is performing reliably, with no major signs of persistent or recurring issues.

In contrast, Plant 2 displayed a markedly different profile. Anomaly detection results for Inverter mqwcsP2rE7j0TFp revealed significantly more irregularities. Unlike the scattered and occasional anomalies seen in Plant 1, Plant 2 showed dense clusters of anomalous points, especially in the mid-to-high irradiation range (0.6 to 1.0). Many of these points reported zero or near-zero AC output despite high solar input—a striking indicator of operational failure. Rather than small deviations, Plant 2 exhibited entire groups of data points that were completely detached from the main performance trend. A particularly concerning pattern was observed between 0.4 and 0.8 irradiation, where output levels were consistently lower than expected. This could point to widespread soiling of panels or potential degradation across multiple modules. Interestingly, the hybrid recovery process also identified more high-output points in Plant 2 compared to Plant 1. These recovered points, mainly found at high irradiation levels, suggest that the system still has the capacity for optimal performance under the right conditions. However, the recurring presence of zero-output points across various time periods and sunlight conditions points toward possible inverter malfunction, wiring issues, or internal component failures.

These contrasting patterns between the two plants offer clear and actionable guidance for maintenance teams. In Plant 1, the relatively small number of anomalies and their limited impact indicate that only light maintenance—such as routine cleaning or minor tuning—may be necessary. The system overall appears healthy, with no urgent signs of deterioration. Plant 2, however, presents a more urgent case. The frequent and severe performance drops, especially those accompanied by temperature anomalies, suggest that the inverter or other key components may be failing. This is not a case of gradual degradation but rather recurring, possibly systemic, faults that warrant immediate investigation. These findings underscore the need for targeted inspections, potential component replacement, and perhaps a broader diagnostic check of the plant's electrical subsystems. The ability of the hybrid detection model to differentiate between true anomalies and exceptional high-performing outputs is particularly valuable here. It not only helps avoid unnecessary alarms but also uncovers opportunities to study and replicate optimal operating conditions elsewhere in the system.

8. Conclusions

This study tackled three major challenges in solar power plant management—power forecasting, maintenance detection, and equipment performance monitoring—using machine learning. By analyzing data from two operational solar plants in India, I demonstrated how these tools can improve operational efficiency and reduce costs.

Among the forecasting models tested, LSTM provided the most balanced performance, while MLP showed stronger responsiveness to sudden changes. The Transformer model, though theoretically robust, was less effective in this case. These insights suggest that hybrid models may offer further benefits by combining the strengths of different architectures.

Our anomaly detection approach, based on an enhanced DBSCAN framework, proved effective in identifying both performance issues and signs of equipment failure. Plant 1 showed relatively stable behavior, with only minor anomalies. In contrast, Plant 2 exhibited persistent and severe anomalies—particularly zero-output periods under high irradiation—highlighting possible inverter faults or system degradation that require urgent attention.

Through inverter-level analysis and temperature correlation, I was able to distinguish between transient issues and recurring faults, offering valuable guidance for prioritizing maintenance efforts. This multi-layered view enables plant operators to move from reactive fixes to proactive management.

While the study offers practical tools for solar plant operations, its short time span and lack of confirmed anomaly labels limit the scope of validation. Future work should include longer-term datasets, integration of weather forecasts, and feedback loops linking detected anomalies with actual maintenance actions.

Overall, this research highlights the value of machine learning in improving the reliability and

performance of solar energy systems—paving the way for smarter, more sustainable energy management.

9. References

[1] Yang, D., et al., "History and trends in solar irradiance and PV power forecasting: A comprehensive review," Solar Energy, vol. 177, pp. 118-142, 2019.

[2] Antonanzas, J., et al., "Review of photovoltaic power forecasting," Solar Energy, vol. 136, pp. 78-111, 2016.

[3] Diagne, M., et al., "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," Renewable and Sustainable Energy Reviews, vol. 27, pp. 65-76, 2013.

[4] Wolff, B., et al., "Comparing support vector regression, random forests, and artificial neural networks for short-term PV power forecasting," Energy Conversion and Management, vol. 156, pp. 279-288, 2018.

[5] Abdel-Nasser, M. and Mahmoud, K., "Accurate photovoltaic power forecasting models using deep LSTM-RNN," Neural Computing and Applications, vol. 31, no. 7, pp. 2727-2740, 2019.

[6] Li, P., et al., "Transformers in time series: A survey," arXiv preprint arXiv:2202.07125, 2022.

[7] Maghami, M.R., et al., "Power loss due to soiling on solar panel: A review," Renewable and Sustainable Energy Reviews, vol. 59, pp. 1307-1316, 2016.

[8] Jiang, Y., et al., "Towards preventive maintenance of photovoltaic systems using machine learning techniques," Renewable Energy, vol. 163, pp. 1742-1751, 2021.

[9] Deng, Z., et al., "An improved DBSCAN algorithm for anomaly detection in time series of solar irradiation," Energy, vol. 238, pp. 121798, 2022.

[10] Liu, H., et al., "Fault detection and diagnosis of photovoltaic system using correlation coefficient and time domain reflectometry," Solar Energy, vol. 159, pp. 670-681, 2018.

[11] Triki-Lahiani, A., et al., "Fault detection and monitoring systems for photovoltaic installations: A review," Renewable and Sustainable Energy Reviews, vol. 82, pp. 2680-2692, 2018.

[12] Harrou, F., et al., "An integrated monitoring approach using statistical and machine learning techniques for fault detection in photovoltaic systems," IEEE Transactions on Industrial Electronics, vol. 68, no. 1, pp. 606-614, 2021.