

SCADA-driven variable similarity-based model for fault detection and predictive maintenance in photovoltaic systems

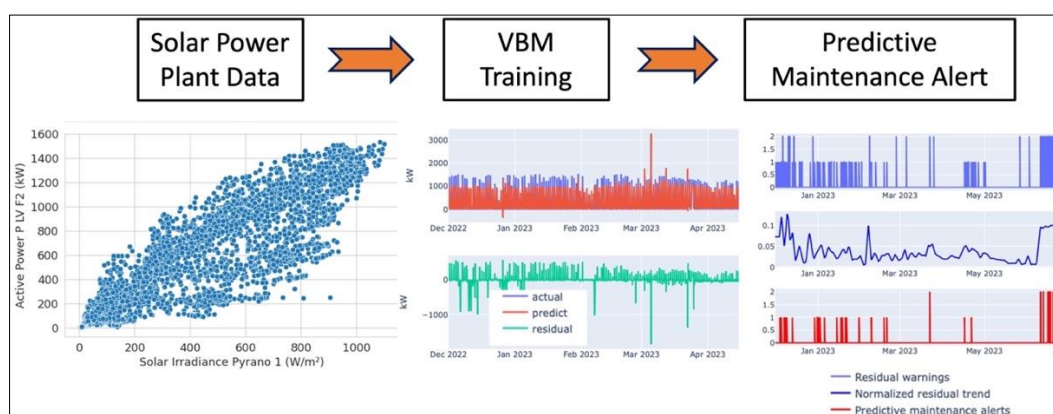
Achmad Widodo^{1*}, Toni Prahasto¹, Agussalim Syamsuddin², Andrew Cahyo Adhi²,
 Amie Kusumawardhani³

¹ Department of Mechanical Engineering, Universitas Diponegoro, Semarang 50275, Indonesia

² PT. PLN (Persero) Research Institute, Jl. Duren Tiga 102, Jakarta 12760, Indonesia

³ Department of Management, Universitas Diponegoro, Semarang 50275, Indonesia

✉ achmadwidodo@lecturer.undip.ac.id



Highlights:

- A predictive maintenance approach is proposed for solar power plants to improve operational reliability.
- A data-driven model is developed based on inverter performance parameters.
- Variable similarity-based model (VBM) is applied for accurate anomaly detection in photovoltaic (PV) systems using SCADA data.
- The system generates early predictive maintenance alerts to prevent failures and optimize maintenance actions.

Article info

Submitted:
 2025-07-24

Revised:
 2026-03-06

Accepted:
 2026-03-30



This work is licensed under
 a Creative Commons
 Attribution-NonCommercial 4.0
 International License

Publisher

Universitas
 Muhammadiyah
 Magelang

Abstract

Global efforts to mitigate the rise in average global temperatures have increasingly emphasized the adoption of renewable energy sources. Among these initiatives, the deployment of solar power plants has emerged as a promising solution, utilizing photovoltaic (PV) systems to convert solar energy—an abundant, cost-free, and year-round resource—into electricity. Solar power plants present a viable alternative to fossil fuels traditionally used in thermal power generation. The reliability of PV systems, which directly influences their performance, functionality, safety, and economic viability, is a critical factor in realizing this potential. This article presents the development of a predictive maintenance framework for PV systems, incorporating anomaly detection and fault diagnosis based on supervisory control and data acquisition (SCADA) data. The methodology employs a variable similarity-based model (VBM) to identify anomalies and diagnose faults, while generating predictive alerts to inform operators of potential issues, thereby enabling proactive maintenance scheduling. The proposed framework is validated using real SCADA data collected from an operational solar power plant. The results demonstrate that the method effectively detects anomalies with reasonable accuracy, underscoring its practicality for application in solar power plant operations.

Keywords: Photovoltaic; Predictive maintenance; Anomaly detection; Fault diagnosis; Machine learning; Similarity based model

1. Introduction

The global shift toward sustainable and renewable energy sources has been driven by the rising energy demand across industrial and domestic sectors [1]. Increasing awareness of green energy and environmental concerns has prompted governments worldwide to introduce policies supporting renewable energy deployment. Among various alternatives, solar energy is one of the most promising renewable resources, as it is abundant, freely available throughout the year, and highly scalable. According to Islam *et al.* [2], PV systems have emerged as a leading technology for harnessing solar energy, converting photon energy into electricity and storing it within battery systems. Despite its growing adoption, the rapid expansion of PV-based electricity generation presents several operational challenges, including concerns regarding system performance, efficiency, and functional reliability. These challenges must be addressed to ensure the effective integration of PV systems into fossil fuel replacement programs in thermal power plants and in hybrid configurations with other renewable energy sources, such as wind farms and hydropower plants.

The increasing utilization of PV systems has also amplified concerns regarding their maintenance and operational reliability. Although modern PV technologies generally require minimal maintenance and are often guaranteed by manufacturers for operational lifetimes exceeding 20 years [3], various factors can lead to performance degradation and failure. These include climatic and meteorological influences [4], atmospheric particulates contributing to surface contamination [5], [6], and defects arising from manufacturing processes, all of which elevate the risk of PV system malfunctions [7], [8]. As classified in [9], failures in PV systems typically originate from four primary factors: design, installation, operation, and maintenance deficiencies. Each of these can adversely affect power generation performance, leading to operational inefficiencies or failures [10], [11].

To mitigate these issues, the adoption of advanced maintenance strategies capable of early failure detection and monitoring is essential. Predictive maintenance (PdM) has emerged as a promising solution, offering capabilities for condition monitoring, fault diagnosis, and failure prognosis in PV systems. As reported in Fernandes *et al.* [12], PdM approaches have demonstrated the potential to improve PV system performance by employing predictive models to prevent array-level failures and enhance operational efficiency. Furthermore, PdM frameworks facilitate the prediction of time-to-failure, enabling timely preventive or corrective actions before actual breakdowns occur [13].

Despite its advantages, the practical implementation of PdM in solar power systems faces several challenges. The approach typically necessitates a large number of sensors to monitor system conditions and provide real-time data, resulting in elevated data acquisition costs [14]. Moreover, existing PdM frameworks for solar applications often involve complex architectures that require extensive datasets for training and validation. Sensor-derived input data are also subject to significant fluctuations and variability compared to output measurements, which complicates model robustness [15]. Additionally, PdM models are frequently site-specific, limiting their transferability across different PV installations due to variations in environmental conditions and data characteristics [16].

In response to these challenges, this study proposes a predictive maintenance framework for solar power plants that incorporates anomaly detection, diagnostic, and prognostic functionalities. The methodology leverages real operational data collected via a SCADA system, with input parameters including solar irradiance, ambient temperature, humidity, wind direction, and wind speed, while the primary calculated output is active AC power. Historical datasets of these parameters are utilized to train a VBM, which is subsequently employed to estimate inverter-generated active AC power. A residual analysis module is then applied to evaluate estimation accuracy and detect anomalies, thereby facilitating early fault detection and enhancing overall system reliability.

2. Related Works

Previous research on anomaly detection and predictive maintenance in PV systems has been widely explored. In De Benedetti *et al.* [17], anomaly prediction in PV systems was addressed using artificial neural networks (ANN), where the model was trained with solar irradiance and temperature data to predict system anomalies. Similarly, Durrani *et al.* [18] investigated PV power generation forecasting by predicting irradiance and temperature parameters to estimate active

power output. A related study by Lee *et al.* [19] introduced a solar energy prediction framework employing convolutional neural networks (CNN) and long short-term memory (LSTM) networks, demonstrating empirically that solar energy output could be accurately predicted using meteorological data obtained from the National Meteorological Center, even when such data were roughly estimated.

The application of machine learning (ML) for predictive maintenance in PV systems has also gained considerable attention. A comprehensive review in Ramirez-Vergara *et al.* [20] examined state-of-the-art approaches linking solar radiation and ambient temperature forecasting techniques with predictive maintenance models, discussing model suitability for PV systems based on prediction horizons, computational complexity, and potential improvements for future deployment. More recently, Rangelov *et al.* [21] investigated the effectiveness of three artificial intelligence (AI) techniques—random forest (RF), deep neural networks (DNN), and LSTM—for short-term PV output prediction. Using meteorological and performance datasets from an operational PV yard, RF and DNN demonstrated strong performance, producing accurate forecasts and effectively managing abrupt shifts in system behavior.

Fault detection studies have also emerged as a significant research direction. In Voutsinas *et al.* [22], a machine learning approach based on logistic regression with cross-validation was employed to detect DC-side PV faults using irradiance and temperature data. Although effective in identifying most faults, this work was limited to simulated models rather than operational large-scale power plants. For such plants, Xu *et al.* [23] proposed a fault detection technique using real-time maximum power point estimation (MPPE) via curve fitting. While conceptually simple, the method achieved reliable fault detection without additional hardware or measurements. Building on this, Jiao *et al.* [24] developed a real-time health identification-based fault diagnosis scheme for large-scale PV systems. By analysing both short-term and long-term operational characteristics, the approach aimed to minimize disruptions and enhance robustness, with its effectiveness validated through simulation and experimental testing. A bibliometric review in Sepúlveda-Oviedo *et al.* [25] revealed that since 2010, over 650 studies have been published on AI-based fault detection in PV systems, emphasizing the rapid expansion of this research domain.

Beyond PV-specific studies, similarity-based modeling has been applied in various power system contexts. Tobar *et al.* [26] implemented a similarity model to characterize power plant process behavior, integrating principal component analysis (PCA) to analyze model residuals for process change detection and baseline condition assessment. In Hu *et al.* [27], a similarity clustering framework for anomaly detection was introduced, comprising a data log collector, analyzer, storage, and behavior detector, though it was not deployed in real industrial settings. Further advancements include a dual similarity-based model for power distribution fault detection [28], combining temporal similarity with operational mode similarity constructed via k-means clustering and Euclidean distance. Similarly, Marins *et al.* [29] proposed an improved similarity-based model for rotating machinery, generating diagnostic features for classification using RF algorithms.

Although similarity-based models have been developed for nearly two decades and successfully applied in diverse domains, their application to anomaly detection in PV systems remains limited. To address this research gap, the present study proposes the implementation of a similarity-based modeling framework for real-world anomaly detection in PV systems.

Table 1 summarizes the additional discussion that highlights the methodological differences between commonly used approaches such as ANN, CNN-LSTM, and regression-based models, and the proposed Variable Similarity-Based Model (VBM) used in this study.

Table 1.

A comparative summary describing the differences between existing approaches for photovoltaic anomaly detection

Study	Method	Input Data	Dataset Type	Advantages	Limitations
De Benedetti <i>et al.</i> [17]	ANN	Irradiance, Temperature	Real PV data	Good anomaly detection accuracy	Requires large training data
Lee <i>et al.</i> [19]	CNN + LSTM	Meteorological data	Simulated / forecast data	Strong predictive capability	High computational complexity
Voutsinas <i>et al.</i> [22]	Feedforward Neural Network	Irradiance, Temperature	Experimental dataset	Good classification ability	Limited generalization
Xu <i>et al.</i> [23]	MPPE curve fitting	PV electrical parameters	Real system	Simple implementation	Limited fault classification
The present work	Variable Similarity-Based Model (VBM)	SCADA data (irradiance, temperature, AC power)	Real 3.5 MWp PV plant	Interpretable, low computational cost, suitable for real-time monitoring	Requires high-quality historical healthy data

3. Method

3.1. Similarity Based Model (SBM)

SBM is a nonparametric empirical modeling technology that uses pattern recognition from historical data to generate estimates of the current value of each variable in a set of modeled data sources [30]. In the field of condition monitoring, the data source is historical parameter data obtained from sensors installed on devices or equipment used to monitor industrial systems or processes. SBM consists of a training data set consisting of matched parameter data from each data source. In most cases, the training data is selected to adequately characterize the healthy or desired operating conditions of the monitored equipment. Unlike parametric modeling methods (such as polynomial regression), no assumptions are made about the format of the solution. Using local interpolation calculations, this model can reasonably, if not accurately, estimate the behavior of suitable devices in varied applications. Another important feature of SBM is that it can model any multivariate data stream as long as there is a consistent relationship between the individual data streams (linear and/or nonlinear). Data suitable for modeling with SBM, such as vibration, pressure, temperature, and flow, are typically considered as time series.

The main requirements for SBM technology are:

- Excellent historical data that accurately represents equipment with free form faults;
- Sufficient historical data to adequately capture patterns of interaction among the variables

These two requirements are characteristic of all empirical techniques. Furthermore, the developed method should propose a technique for cleaning the data to meet requirement 1, as well as an adaptive bootstrapping method for implementations where the available data are insufficient to meet requirement 2. Unlike traditional techniques, the SBM adaptation process is intermittent and only occurs when no anomalies are detected in the monitored data sources. Such coordination significantly reduces the possibility of coordination during failures.

SBM technology is based on applying similarity operations on pairs of observation sensor parameters and on the state matrix (D) containing the historical training data mentioned above. Thus, the number of columns of the state matrix corresponds to the number of representative training data (M), and the number of rows corresponds to the number of data sources included in training data (L). Define measurements taken at a particular time n_j as an example of a training data $X(n_j)$ as following Eq. (1).

$$X(n_j) = [x_1(n_j)x_2(n_j)x_3(n_j) \dots x_L(n_j)] \quad (1)$$

where $x_i(n_j)$ is the measurement from data source i at time n_j , then the state matrix D is given by:

$$D = [X(n_1)X(n_2)X(n_3) \dots X(n_M)] \quad (2)$$

The result of a similarity operation between two observation vectors is a similarity value (scalar). The similarity operation is similar to a dot product and can be extended (nonlinearly) to a matrix operation, with a scalar similarity for each combination of a row from the matrix and a column from the second matrix. Therefore, if it is given an input vector X_{in} containing individual measurements from each of the L data sources, the vector of corresponding estimated data source values is determined from the equation X_{est} .

$$X_{est} = D \cdot w \quad (3)$$

The estimate consists of a linear combination of vectors stored in D. Here, w is a weighting factor formulated as:

$$W = \frac{\hat{w}}{\sum_{j=1}^N \hat{w}(j)} \quad (4)$$

where,

$$\hat{w} = (D^T \otimes D)^{-1} \cdot (D^T \otimes X_{in}) = G^{-1} \cdot A \quad (5)$$

Moreover, similarity score can be evaluated using the following formula [31].

$$s(x_i, x_j) = \left\{ 1 + \sqrt{\sum_{i,j=1}^N (x_i - x_j)^2} \right\}^{-1} \quad (6)$$

The similarity score yields $0 \leq s \leq 1$ such that if $s \approx 1$ means two vectors are very similar but if $s \approx 0$ corresponds to not-similar vectors.

The D matrix is usually composed from historical data of device in normal condition that regarded to as reference data. This matrix of course can represent full dynamic range of device being monitored during operation. The number of sensors installed in the system device determines the number of vectors used for construction the D matrix. Actually, there is no specific rule for this determination, for example if the correlation of the data sources to build the model is highly linear, it would imply that only a few vectors are needed to construct the D matrix. A general rule of thumb that can be referred to in selecting the number of vectors is at least twice the number of data sources.

In relation with SBM, the proposed approach used in this work called variable-similarity based model (VBM) is utilized for generating the model of anomaly detection. VBM tends to be a local learning-based method that uses each new input to generate a state matrix. Therefore, the characteristics of VBM can capture faster changes in the dynamic range that occur in a given operating time. In VBM, the state matrix $D(t)$ is constructed by selecting of subset data from the most relevant to current input vectors in the superset H that includes the full dynamic operating range of modelled system. Then, the estimated vectors are generated from linear combination of only selected currently relevant vectors. The formulation of VBM is as follow:

$$D(t) = \{H|F(H, X_{in})\} \quad (7)$$

and the weighting factors in Eq. (5) is modified

$$\hat{W}' = (D(t)^T \otimes D(t))^{-1} \cdot (D(t)^T \otimes X_{in}) \quad (8)$$

where H , F and X_{in} are reference data matrix, vector selection process and input vectors, respectively.

3.2. The Proposed Method

The PV inverter system model is constructed using historical data from the monitored power plant. Time-series records of solar irradiance, ambient temperature, and active AC power are utilized to train the VBM, enabling the model to capture system dynamics under normal operating conditions. Once the model is developed, it is deployed to assess the operational health of the solar power system by processing new real-time data streams from the plant. The outputs of this analysis encompass both anomaly detection and fault diagnosis. Specifically, the VBM is employed to estimate the active AC power generated by the inverter, with the model representing the expected behavior of the system under nominal conditions. The discrepancy, or residual, between the model output and the measured power plant data is computed, and a warning is triggered when this residual exceeds a predefined operational threshold.

To enhance the robustness of fault detection, an autocorrelation function is incorporated to capture the inherent periodicity within the residual signals, such as daily cycles. The longest detected periodicity serves as an input parameter for calculating the triangular moving average (TMA). The TMA smooths fluctuations in the residuals and accentuates long-term patterns that may indicate system degradation or performance drift. These trends are further analyzed using a derivative function to identify the onset of critical deviations. Predictive maintenance alerts are subsequently issued when the derivative of the residual trend exhibits persistent positive values, signaling progressive system deterioration. The overall workflow for anomaly detection and predictive alert generation is illustrated in [Figure 1](#).

3.3. Solar Power Plant Data

The dataset utilized in this study was obtained from a 3.5 MWp PV power plant comprising 7,778 PV modules. Operational data were collected through a SCADA system, which recorded one-minute resolution measurements from multiple sensors, including solar irradiance (pyranometer), ambient temperature, and wind speed and direction. The PV system configuration is described as follows:

- Inverters #1–#9 (Feeder F1):
 - Zone B: 13 strings
 - Zone C: 149 strings
- Inverters #10–#18 (Feeder F2):
 - Zone A: 111 strings

- Zone B: 46 strings
- Zone C: 4 strings

The dataset used for validation was recorded between December 1, 2022, and September 30, 2023, yielding a total of 437,759 data points. The dataset was divided into two subsets: one portion was employed for training the VBM, while the remaining portion was used for testing and validation. Figure 2 illustrates the representative daily AC power output profile calculated by the inverters during the observation period.

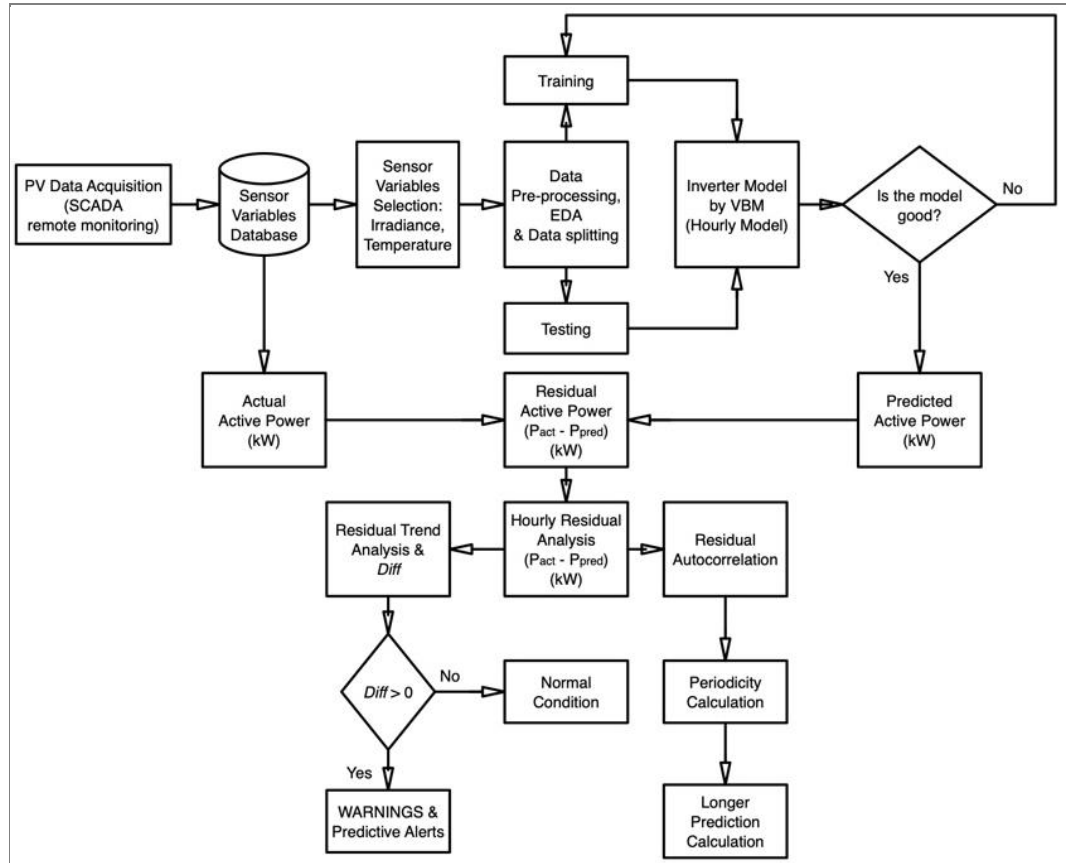


Figure 1.
The proposed method
of anomaly detection
and predictive alerts

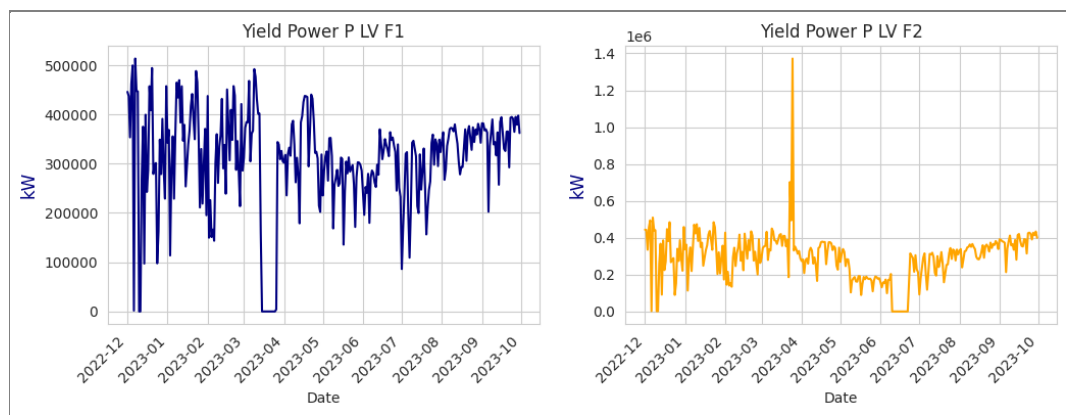


Figure 2.
Data daily yield power
AC calculated from
inverter

3.4. Model Development

3.4.1. Pre-processing data

Prior to analysis, data pre-processing was conducted to eliminate outliers and address irregularities arising from noise and inherent disturbances in the raw measurements. Real-world datasets often exhibit missing values for specific parameters due to occasional sensor malfunctions; however, continuous data acquisition is maintained through redundant SCADA sensors. To address these gaps, imputation techniques were employed to estimate the missing values, thereby ensuring dataset completeness prior to subsequent statistical and modeling processes. Imputation was performed using established approaches, including mean, median,

regression, and nearest-neighbor methods, as widely reported in the literature [32], [33]. Furthermore, to mitigate the influence of extreme values, outlier detection and removal were carried out using the commonly adopted interquartile range (IQR) method [34].

$$Th_u = Q_3 + 1.5 IQR \quad (9)$$

where Th_u represents the upper threshold for outlier removal, and the IQR is defined as $(Q_3 - Q_1)$ with Q_3 and Q_1 corresponding to the 75th and 25th percentiles of the data, respectively [35].

Figure 3 illustrates the irradiance and ambient temperature datasets following outlier removal. Additionally, Figure 4 presents the hourly profiles of active AC power recorded during daylight hours for feeders F1 and F2. The peak system performance is typically observed between 11:00 a.m. and 1:00 p.m. Notably, certain data points within this interval exhibit abnormally low active power output. These instances suggest potential anomalies within the dataset, which must be excluded when constructing the baseline or nominal state model to ensure the accuracy and reliability of subsequent analysis.

Figure 3.
Dataset after outliers
removal

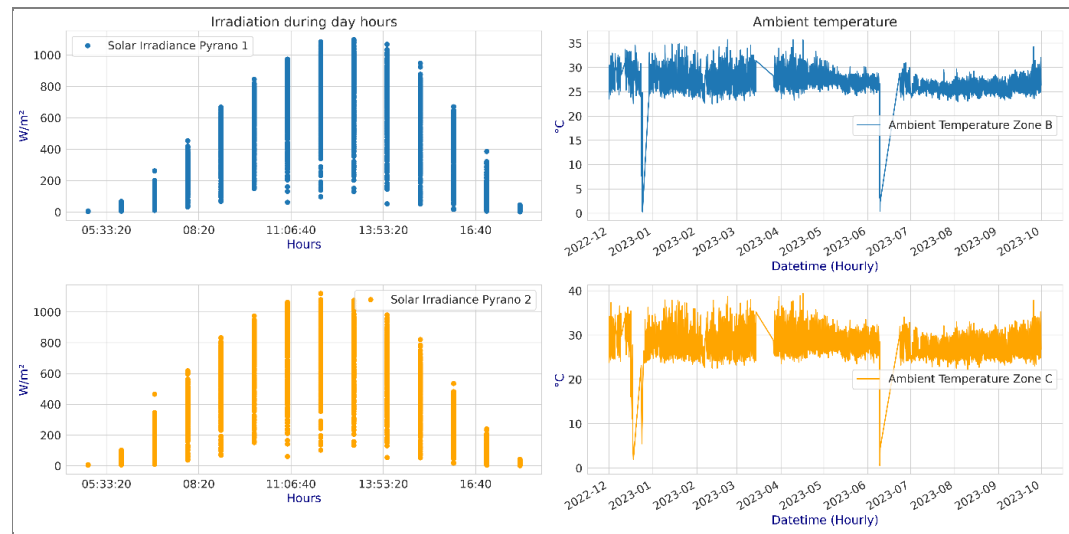
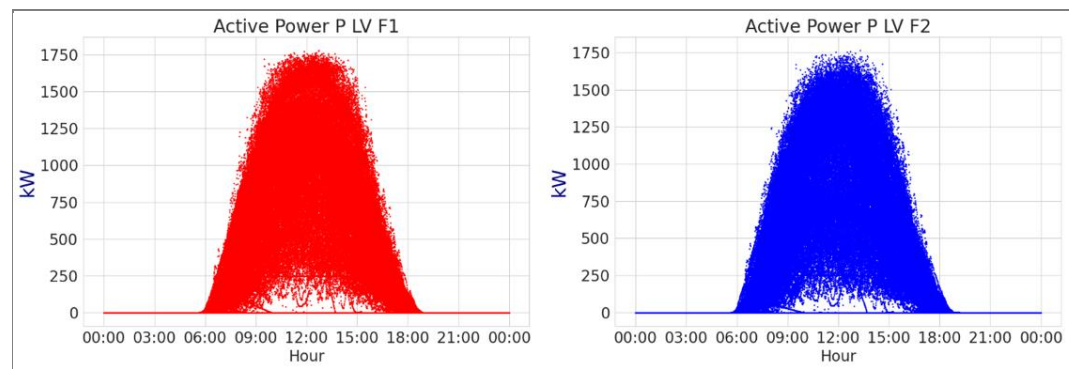


Figure 4.
Hourly active power AC



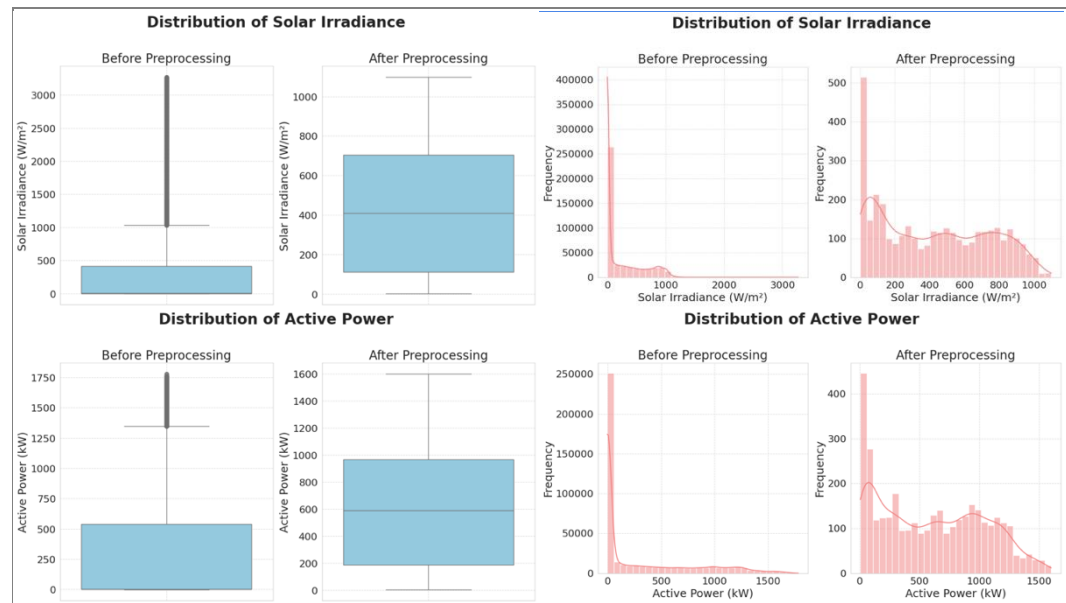
As described in the preceding section, the SCADA system recorded 437,759 data points at a one-minute sampling interval across the monitored parameters. This dataset exhibits substantial variance; however, not all recorded observations were utilized for model development and validation. Many data points correspond to transient or non-representative conditions and, therefore, do not accurately characterize the normal behavior of the monitored physical variables. To refine the dataset, reduce its size, and extract representative information, the following pre-processing steps were implemented:

- Exclusion of data points where the active power is near zero while irradiance remains slightly above zero.
- Removal of data recorded during periods of minimal solar exposure, specifically before 05:30 a.m. and after 06:45 p.m.
- Resampling of the remaining data into hourly intervals using averaged values.

This resampling procedure effectively minimizes data variability and enhances the reliability of the resulting model. Following these steps, the dataset was reduced to 3,520 hourly data points, providing a more robust foundation for subsequent model training and evaluation.

The effectiveness of the preprocessing procedure is illustrated in Figure 5. The boxplot and histogram visualizations demonstrate that the preprocessing steps enhance the quality of the dataset by removing outliers and improving the statistical consistency of the photovoltaic (PV) data distribution.

Figure 5.
The effectiveness of
the preprocessing
method



3.4.2. The VBM training and testing

The PV power generation process is characterized using a generalized mathematical formulation, as expressed in Eq. (10), which defines the relationship between the generated active power and incident solar irradiance.

$$P = f(I, T) \quad (10)$$

where P , I , and T denote the active power output (kW), solar irradiance (W/m^2), and ambient temperature ($^{\circ}\text{C}$), respectively. An examination of the dataset reveals the relationship between active power and solar irradiance, as illustrated in Figure 6. However, establishing an explicit analytical expression to describe this relationship is challenging due to the nonlinear and dynamic characteristics of the system. To address this complexity, a data-driven, black-box modeling approach is adopted. Such models leverage ML techniques to infer the underlying relationships between parameters by learning from historical system data, thereby enhancing the predictive capability for future observations.

In this study, the VBM is employed as the black-box modeling framework. VBM is trained on the available dataset to capture the intrinsic characteristics of system behavior and subsequently utilizes the acquired knowledge to predict the operational status of the solar power system, including scenarios where the system encounters faults or deviations during operation.

The VBM was trained using 2,816 error-free data points, selected with conditioned parameters to represent the system's normal operating state. The remaining data points were subsequently employed as a test dataset to evaluate the performance of the trained model. As an illustrative example, Figure 7 presents a comparison between the actual and VBM-predicted active power recorded by the F1 power meter. The corresponding residuals, representing the deviations between the measured and predicted active power, are also depicted in Figure 7. Overall, the model demonstrates satisfactory predictive performance; however, certain gaps and inconsistencies within the dataset may still be present.

4. Results and Discussion

4.1. Anomaly Detection

By analysing the residual values, the root mean square error (RMSE) loss is computed, which is subsequently used to establish the anomaly detection threshold, as illustrated in Figure 8. Given that RMSE reflects the standard deviation associated with the system's normal operating

conditions, this metric enables the identification of potential abnormalities by determining whether deviations exceed the statistically defined threshold.

Figure 6.
Correlation between
measured irradiance
and corresponding
active power

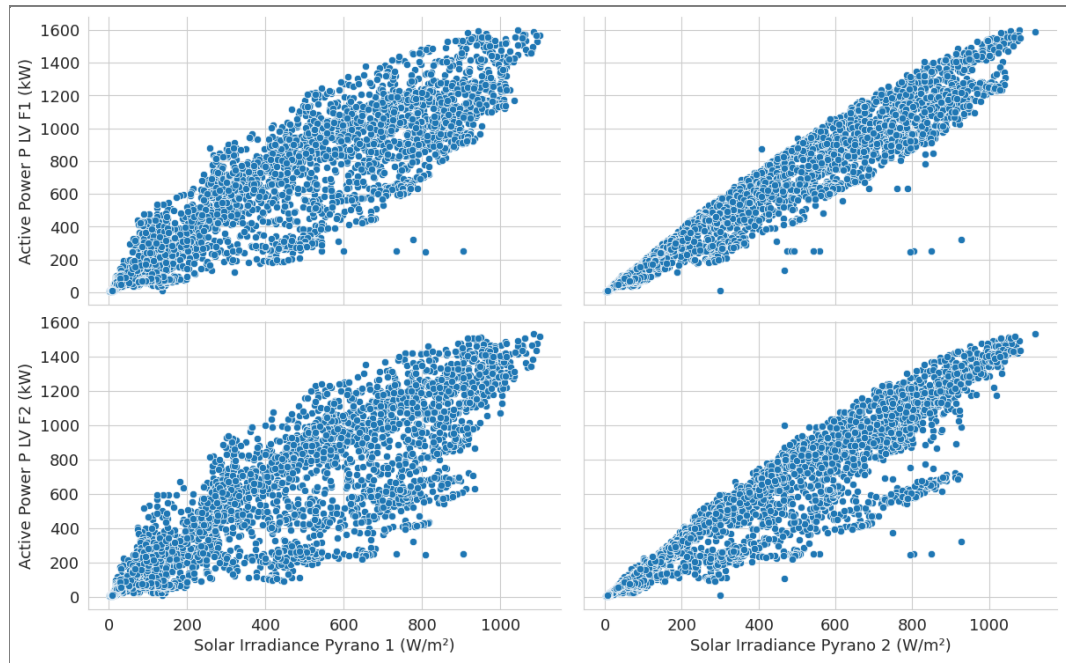
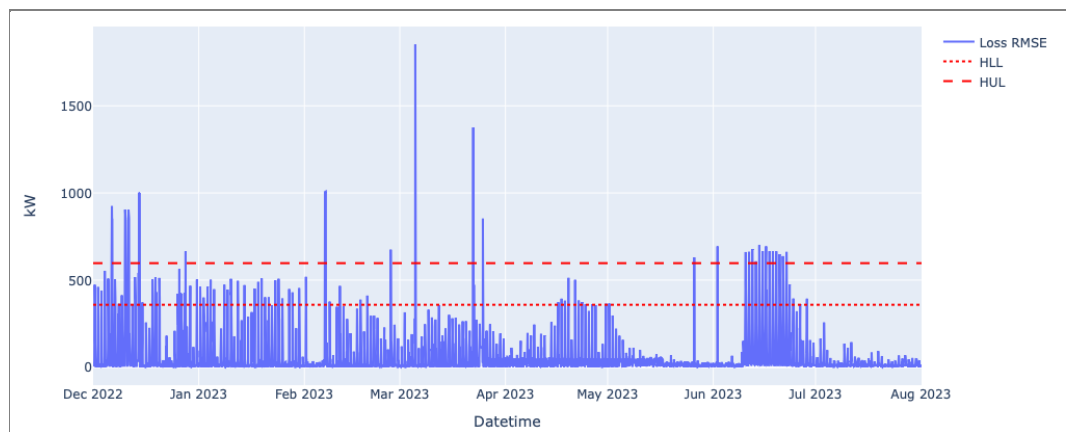


Figure 7.
Training for VBM



Figure 8.
Loss RMSE for a basis
threshold



Under normal operating conditions, PV power plants exhibit system behavior that closely aligns with the predicted parameters, resulting in residuals with relatively low magnitudes. For hourly datasets, anomaly detection thresholds can be defined as the hourly lower limit (HLL) and hourly upper limit (HUL), which are established based on the standard deviation of hourly data σ_h and can be expressed as:

$$HLL = 3\sigma_h \quad (11)$$

$$HUL = 5\sigma_h \quad (12)$$

With respect to these thresholds, when new data points exceed the hourly lower limit (HLL) but remain below the hourly upper limit (HUL), the system is considered to exhibit abnormal behavior. Conversely, values exceeding the HUL indicate a strong anomaly. This thresholding approach, adapted from De Benedetti *et al.* [17] with reference to Montgomery and Runger [36], is justified by statistical properties of normally distributed (Gaussian) data, where approximately 99.7% of observations fall within $3\sigma_h$ and 99.9% within $5\sigma_h$. Even when the dataset deviates from a normal distribution, at least 88.9% of data points typically fall within $3\sigma_h$, thus supporting the selection of $3\sigma_h$ and $5\sigma_h$ as reasonable HLL and HUL values, respectively.

Figure 9 shows a data plot at specific time that clearly exceeds the HLL/HUL threshold and triggers an alert. In a healthy state, the actual active power (top plot) should follow a smooth, high-reaching 'bell curve' between 7 AM and 5 PM. Instead, the plot shows that while the system starts generating power in the morning, it crashes to 0 kW multiple times during peak hours (around 11:00 AM, 1:00 PM, and 3:00 PM).

Looking at the vertical alignment, every time the power hits 0 kW on the top graph, the RMSE (bottom plot) spikes vertically. At 11:00 AM, the power drops, and the RMSE instantly shoots past the HLL (dotted red), triggering the orange circle markers. As the power remains at zero while the sun is still high, the RMSE continues to climb past the HUL (dashed red), resulting in the red 'X' markers. This confirms that the model correctly identifies this as a 'major deviation' because the expected power (based on high irradiance) is completely absent.

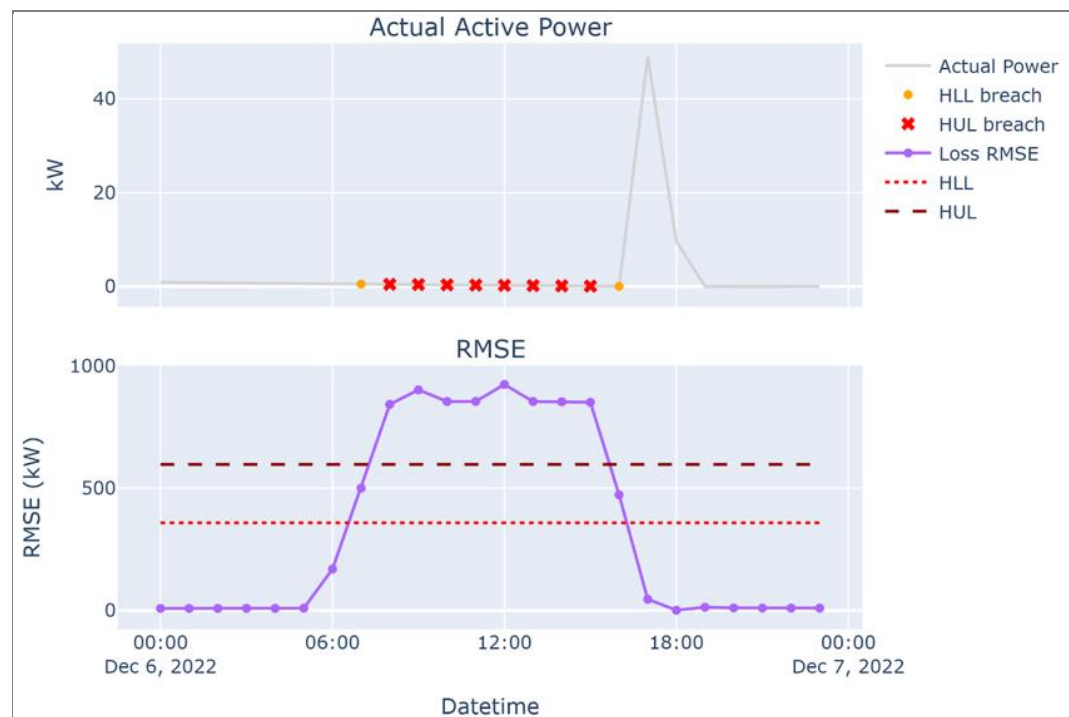


Figure 9.
Actual power data
exceeding the HLL and
HUL thresholds

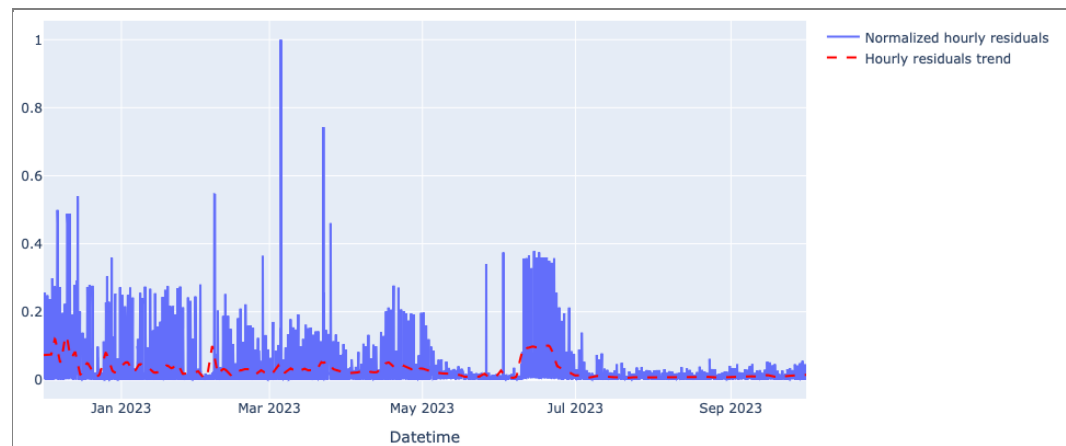
To enhance anomaly detection accuracy and minimize false positive alerts, additional signal processing is applied at the preliminary detection stage. Residual signals often exhibit high-frequency fluctuations, generating numerous transient spikes that do not reflect true anomalies. To mitigate this, residuals are smoothed to emphasize underlying trends while preserving informative anomaly signatures. A triangular moving average (TMA) is employed as the smoothing filter, with its window size determined using the autocorrelation function (ACF), which captures the inherent periodicity of residual fluctuations. The window size selection involves the following steps:

- Identify peak values in the ACF output and designate these as short-term periods.
- Compute the short-term period average (SPA) by averaging the corresponding time values of these peaks.
- Determine the long-term period by selecting all autocorrelation peaks with amplitudes greater than 0.2.

- From these peaks, select time points whose differences exceed the SPA, average these time values to obtain the long-term period average (LPA), and set the TMA window size to $2 \times \text{LPA}$.

Applying these calculations to the dataset yields SPA and LPA values of 24 hours and 48 hours, respectively, resulting in a TMA window size of 48 hours. The normalized residual trend, computed using this triangular moving average, is presented in [Figure 10](#).

Figure 10.
Normalized residual trend



[Table 2](#) shows the comparison of raw counts and TMA-filtered counts showing the reduction of high-frequency noise of warning levels. The effectiveness of the TMA in enhancing anomaly detection is best demonstrated by comparing the raw warning frequencies in [Figure 8](#) with the filtered alerts in [Figure 10](#). The raw residuals are characterized by significant high-frequency noise, manifesting as a "cloud" of transient spikes. As shown in the initial value counts, the system recorded 110 instances of Level 1 warnings (HLL) and 81 instances of Level 2 warnings (HUL). Many of these appear as isolated, single-epoch vertical lines; these represent stochastic fluctuations—likely caused by sensor jitter or momentary operational transients—that do not reflect a genuine shift in machine health.

By applying the TMA filter, this high-frequency is suppressed to reveal the underlying normalized residual trend. The impact of this smoothing is statistically profound: Level 1 warnings (HLL) were reduced from 110 to 33, and Level 2 warnings (HUL) dropped from 81 to 19 (reductions of 70% and 76%, respectively). The insight here is that the TMA acts as a temporal consistency check. Because the TMA assigns a triangular weighting—where the central data point carries the most influence and surrounding points taper off—a single-point noise spike lacks the cumulative "energy" to significantly shift the trend line.

The signal must remain elevated for a sustained period before it can influence the moving average enough to trigger an alert. This process consolidates scattered, unreliable data points into distinct, high-confidence alert clusters. For instance, in the June 2023 period of [Figure 10](#), the TMA trend ignores the "twitchy" raw spikes and instead highlights a sustained plateau of degradation. Consequently, the resulting predictive maintenance alerts become far more actionable, ensuring that maintenance resources are deployed only when the system detects a persistent and statistically significant departure from normal operating conditions, rather than responding to momentary, insignificant fluctuations.

Table 2.

Comparison of raw counts and TMA-filtered counts showing the reduction of high-frequency noise

Warning Level	Raw Count	TMA Count	Reduction
Level 1 (HLL)	110	33	70%
Level 2 (HUL)	81	19	76%
Total Flags	191	52	73%

4.2. Generation of Predictive Alerts

Predictive alerts are generated to notify solar power plant operators of potential anomalies that may precede power losses by several days or hours, thereby enabling proactive maintenance planning to mitigate risks of generation shortfalls. These alerts are derived through differential analysis of the residual trend, which reflects deviations in the system's operational behavior relative to healthy inverter conditions. A positive derivative of the residual trend indicates progressive deterioration of the inverter, as the trend approaches the upper anomaly threshold,

signifying an emerging fault. Conversely, a negative derivative suggests stable operation, with the system remaining within normal conditions.

The predictive capability of the trained VBM was assessed using a test dataset excluded from the training phase. A total of 704 data points, corresponding to the period from August to September 2023, were processed through the model. The prediction outcomes, presented in Figure 11, yielded a root mean square error (RMSE) of 121.5 kW, demonstrating the model's performance in forecasting potential anomalies.

The performance of the proposed VBM demonstrates its capability to accurately represent the operational behavior of the photovoltaic system under normal conditions. The testing results show that the model achieved a root mean square error (RMSE) of 121.5 kW when evaluated on the testing dataset. Considering the dynamic nature of PV power generation, which is influenced by environmental factors such as solar irradiance and ambient temperature, this level of prediction error indicates that the model can reliably capture the nonlinear relationship between the input variables and the generated active power. Furthermore, the residual-based analysis enables effective identification of deviations between predicted and measured system behavior. When the residual values exceed the statistically defined anomaly thresholds (HLL and HUL), the system can be interpreted as entering an abnormal operating condition. This statistical interpretation supports the robustness of the proposed anomaly detection framework, as the threshold selection is grounded in the probabilistic characteristics of the residual distribution. In addition, the integration of autocorrelation analysis and triangular moving average smoothing helps reveal long-term degradation patterns rather than transient fluctuations, thereby improving the reliability of predictive maintenance alerts. These findings indicate that the proposed framework is capable of detecting early signs of performance deterioration in photovoltaic systems, enabling operators to perform proactive maintenance before significant power losses occur.

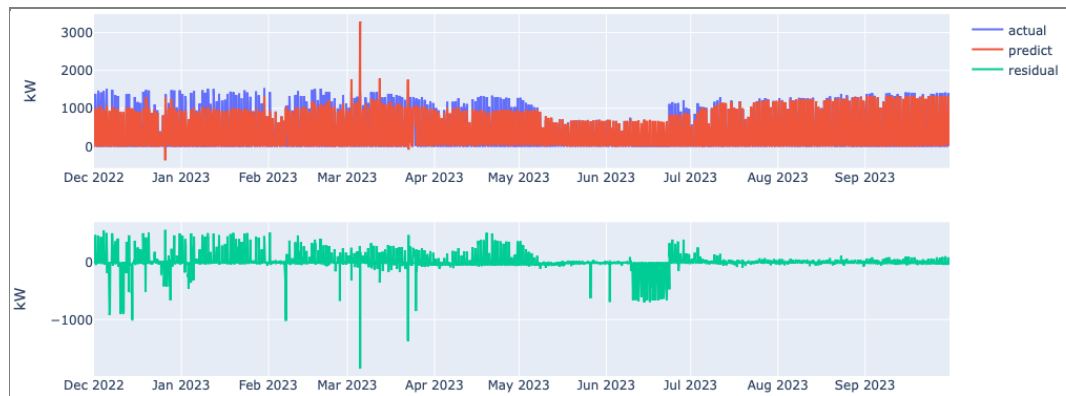


Figure 11.
Testing results of VBM

Following the testing phase, the developed model was applied to the complete dataset spanning December 2022 to September 2023. The prediction outcomes, incorporating residual analysis, generated multiple residual-based warnings, as illustrated in Figure 12. While the model successfully identified numerous potential anomalies, additional verification steps were necessary to reduce false positives. Specifically, residual warnings were cross-validated against the derivative of the residual trend, and predictive maintenance alerts were issued only when residual warnings coincided with positive derivative values, which indicate progressive system deterioration.

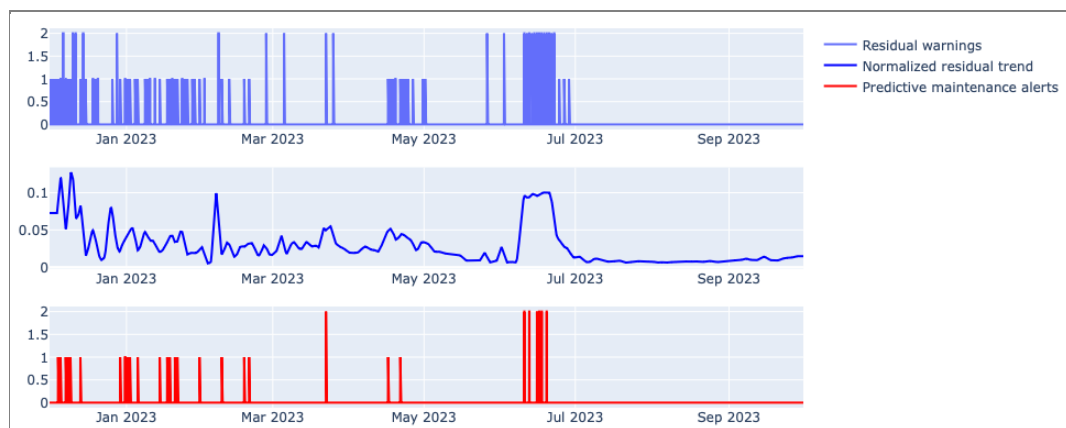


Figure 12.
Predictive maintenance alerts

In comparison with the predictive maintenance alert generation approaches reported in [17], [37], which employed artificial neural networks (ANN), the VBM offers a more interpretable and computationally efficient alternative while maintaining satisfactory performance. Unlike ANN-based methods, which often require substantial computational resources and extensive hyperparameter optimization, VBM relies on simplified matrix-based formulations, facilitating faster implementation and seamless integration into real-time monitoring frameworks.

VBM constructs a state matrix from multivariate input data by exploiting consistent similarity relationships among variables—both linear and nonlinear—without resorting to deep learning architectures. This characteristic renders VBM particularly advantageous for industrial applications, where transparency, interpretability, and computational efficiency are essential. Furthermore, the model's capability to generalize across diverse operational conditions enhances its suitability for practical predictive maintenance applications.

Figure 12 illustrates an example of a predictive maintenance alert generated using the proposed method, designed to notify operators of potential abnormalities within the solar power system. The developed analytical framework is intended to identify performance degradation that may precede imminent or impending system failures. It is important to note, however, that the presence of a warning does not unequivocally confirm the occurrence of a fault. Instead, the alert serves to indicate deviations from typical PV system behavior, prompting operators to assess the system's condition and determine appropriate maintenance actions. Consequently, the proposed method functions as a decision-support tool, augmenting but not replacing the expertise of maintenance personnel.

5. Conclusion

This study presented a predictive maintenance framework for photovoltaic (PV) systems based on a variable similarity-based model (VBM) utilizing real operational data obtained from a SCADA monitoring system. The model was developed using 2,816 data samples representing normal operating conditions and evaluated using a testing dataset of 704 samples. The results demonstrate that the proposed approach can effectively capture the relationship between solar irradiance, ambient temperature, and inverter-generated active AC power. The testing results produced a root mean square error (RMSE) of 121.5 kW, indicating that the model is capable of providing reliable estimations of PV system behavior under varying environmental conditions. By combining residual-based anomaly detection with statistical thresholds and residual trend analysis using triangular moving average filtering, the proposed framework successfully identifies abnormal system behavior and generates predictive maintenance alerts. These alerts provide early indications of potential performance degradation, enabling operators to implement proactive maintenance actions and reduce the risk of power losses in solar power plants.

Acknowledgements

The authors gratefully acknowledge PT PLN Indonesia Power for providing access to operational data from utility-scale photovoltaic (PV) systems, which formed the foundation of this research. The authors also extend their appreciation to the Faculty of Engineering, Universitas Diponegoro, Indonesia, for its institutional support and for fostering an academic environment conducive to the conduct of this study. Special thanks are extended to Yasir Abdur Rahman and Ghiffari Awliya for their valuable assistance in computer programming, data analytics, and data processing activities throughout the research. Their technical contributions were instrumental in supporting the development, validation, and implementation of the proposed methodology.

Authors' Declaration

Authors' contributions and responsibilities - The authors made substantial contributions to the conception and design of the study. The authors took responsibility for data analysis, interpretation and discussion of results. The authors read and approved the final manuscript.

Funding – No funding information from the authors.

Availability of data and materials - All data is available from the authors.

Competing interests - The authors declare no competing interests.

Additional information – No additional information from the authors.

References

- [1] G. E. Durán-Castillo, T. Weis, A. Leach, and B. A. Fleck, "Revenue-optimized photovoltaic orientation in a northern competitive electricity market with carbon offsets," *Energy Reports*, vol. 10, pp. 3133–3145, Nov. 2023, doi: 10.1016/j.egy.2023.09.135.
- [2] M. Islam, M. R. Rashel, M. T. Ahmed, A. K. M. K. Islam, and M. Tlemçani, "Artificial Intelligence in Photovoltaic Fault Identification and Diagnosis: A Systematic Review," *Energies*, vol. 16, no. 21, p. 7417, Nov. 2023, doi: 10.3390/en16217417.
- [3] R. Siddiqui et al., "Comparison of different technologies for solar PV (Photovoltaic) outdoor performance using indoor accelerated aging tests for long term reliability," *Energy*, vol. 107, pp. 550–561, Jul. 2016, doi: 10.1016/j.energy.2016.04.054.
- [4] P. Lauret, E. Lorenz, and M. David, "Solar Forecasting in a Challenging Insular Context," *Atmosphere*, vol. 7, no. 2, p. 18, Jan. 2016, doi: 10.3390/atmos7020018.
- [5] M. Paulescu and E. Paulescu, "Short-term forecasting of solar irradiance," *Renewable Energy*, vol. 143, pp. 985–994, Dec. 2019, doi: 10.1016/j.renene.2019.05.075.
- [6] J. Sánchez-Balseca, J. L. Pineiros, and A. Pérez-Foguet, "Influence of environmental factors on the power produced by photovoltaic panels artificially weathered," *Renewable and Sustainable Energy Reviews*, vol. 188, p. 113831, Dec. 2023, doi: 10.1016/j.rser.2023.113831.
- [7] H. Acikgoz, D. Korkmaz, and U. Budak, "Photovoltaic cell defect classification based on integration of residual-inception network and spatial pyramid pooling in electroluminescence images," *Expert Systems with Applications*, vol. 229, p. 120546, Nov. 2023, doi: 10.1016/j.eswa.2023.120546.
- [8] B. Aboagye, S. Gyamfi, E. A. Ofosu, and S. Djordjevic, "Characterisation of visual defects on installed solar photovoltaic (PV) modules in different climatic zones in Ghana," *Scientific African*, vol. 20, p. e01682, Jul. 2023, doi: 10.1016/j.sciaf.2023.e01682.
- [9] S. R. Madeti and S. N. Singh, "A comprehensive study on different types of faults and detection techniques for solar photovoltaic system," *Solar Energy*, vol. 158, pp. 161–185, Dec. 2017, doi: 10.1016/j.solener.2017.08.069.
- [10] W. Otieno and L. Bosman, "Solar PV selection decision tool: the case of USA Midwest region," *Journal of Civil Engineering and Environmental Technology*, vol. 1, pp. 114–120, 2014.
- [11] F. Ancuta and C. Cepisca, "Fault analysis possibilities for PV panels," in *Proceedings of the 2011 3rd International Youth Conference on Energetics (IYCE)*, 2011, pp. 1–5.
- [12] M. Fernandes, A. Canito, J. M. Corchado, and G. Marreiros, "Fault Detection Mechanism of a Predictive Maintenance System Based on Autoregressive Integrated Moving Average Models," in *Distributed Computing and Artificial Intelligence, 16th International Conference*, F. Herrera, K. Matsui, and S. Rodríguez-González, Eds., Springer International Publishing, 2020, pp. 171–180. doi: 10.1007/978-3-030-23887-2_20.
- [13] G. J.-P. Tevi, M. E. Faye, M. Sene, I. Faye, U. Blieske, and A. Seidou Maiga, "Solar Photovoltaic Panels Failures Causing Power Losses: A Review," in *2018 7th International Energy and Sustainability Conference (IESC)*, IEEE, May 2018, pp. 1–9. doi: 10.1109/IESC.2018.8439986.
- [14] M. L. López, L. E. Olcese, G. G. Palancar, and B. M. Toselli, "Accurate total solar irradiance estimates under irradiance measurements scarcity scenarios," *Environmental Monitoring and Assessment*, vol. 191, no. 9, p. 568, Sep. 2019, doi: 10.1007/s10661-019-7742-3.
- [15] M. Caldas and R. Alonso-Suárez, "Very short-term solar irradiance forecast using all-sky imaging and real-time irradiance measurements," *Renewable Energy*, vol. 143, pp. 1643–1658, Dec. 2019, doi: 10.1016/j.renene.2019.05.069.
- [16] M. Q. Raza, N. Mithulananthan, J. Li, K. Y. Lee, and H. B. Gooi, "An Ensemble Framework for Day-Ahead Forecast of PV Output Power in Smart Grids," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 8, pp. 4624–4634, Aug. 2019, doi: 10.1109/TII.2018.2882598.
- [17] M. De Benedetti, F. Leonardi, F. Messina, C. Santoro, and A. Vasilakos, "Anomaly detection and predictive maintenance for photovoltaic systems," *Neurocomputing*, vol. 310, pp. 59–68, Oct. 2018, doi: 10.1016/j.neucom.2018.05.017.
- [18] S. P. Durrani, S. Balluff, L. Wurzer, and S. Krauter, "Photovoltaic yield prediction using an

- irradiance forecast model based on multiple neural networks," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 2, pp. 255–267, Mar. 2018, doi: 10.1007/s40565-018-0393-5.
- [19] W. Lee, K. Kim, J. Park, J. Kim, and Y. Kim, "Forecasting Solar Power Using Long-Short Term Memory and Convolutional Neural Networks," *IEEE Access*, vol. 6, pp. 73068–73080, 2018, doi: 10.1109/ACCESS.2018.2883330.
 - [20] J. Ramirez-Vergara, L. B. Bosman, E. Wollega, and W. D. Leon-Salas, "Review of forecasting methods to support photovoltaic predictive maintenance," *Cleaner Engineering and Technology*, vol. 8, p. 100460, Jun. 2022, doi: 10.1016/j.clet.2022.100460.
 - [21] D. Rangelov, M. Boerger, N. Tcholtchev, P. Lämmel, and M. Hauswirth, "Design and Development of a Short-Term Photovoltaic Power Output Forecasting Method Based on Random Forest, Deep Neural Network and LSTM Using Readily Available Weather Features," *IEEE Access*, vol. 11, pp. 41578–41595, 2023, doi: 10.1109/ACCESS.2023.3270714.
 - [22] S. Voutsinas, D. Karolidis, I. Voyiatzis, and M. Samarakou, "Development of a multi-output feed-forward neural network for fault detection in Photovoltaic Systems," *Energy Reports*, vol. 8, pp. 33–42, Nov. 2022, doi: 10.1016/j.egy.2022.06.107.
 - [23] Q. Xu, X. Li, X. Yin, and C. Feng, "A Real-Time Fault Detection Technique Based on MPPE in Photovoltaic Systems," in *2023 IEEE 14th International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, IEEE, Jun. 2023, pp. 164–168. doi: 10.1109/PEDG56097.2023.10215228.
 - [24] X. Jiao, X. Li, T. Yang, Y. Yang, and W. Xiao, "A Novel Fault Diagnosis Scheme for PV Plants Based on Real-Time System State Identification," *IEEE Journal of Photovoltaics*, vol. 13, no. 4, pp. 571–579, Jul. 2023, doi: 10.1109/JPHOTOV.2023.3262950.
 - [25] E. H. Sepúlveda-Oviedo, L. Travé-Massuyès, A. Subias, M. Pavlov, and C. Alonso, "Fault diagnosis of photovoltaic systems using artificial intelligence: A bibliometric approach," *Heliyon*, vol. 9, no. 11, p. e21491, Nov. 2023, doi: 10.1016/j.heliyon.2023.e21491.
 - [26] F. A. Tobar, L. Yacher, R. Paredes, and M. E. Orchard, "Anomaly detection in power generation plants using similarity-based modeling and multivariate analysis," in *Proceedings of the 2011 American Control Conference*, IEEE, Jun. 2011, pp. 1940–1945. doi: 10.1109/ACC.2011.5991323.
 - [27] S. Hu, Z. Xiao, Q. Rao, and R. Liao, "An anomaly detection model of user behavior based on similarity clustering," in *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, IEEE, Dec. 2018, pp. 835–838. doi: 10.1109/ITOEC.2018.8740748.
 - [28] Y. Liu, Y. Wang, Y. Chen, Q. Yang, H. Hong, and W. Wang, "Dual Similarity-based Method for Proactive Fault Detection in Power Distribution Networks," in *2022 IEEE 6th Conference on Energy Internet and Energy System Integration (EI2)*, IEEE, Nov. 2022, pp. 1114–1119. doi: 10.1109/EI256261.2022.10116592.
 - [29] M. A. Marins, F. M. L. Ribeiro, S. L. Netto, and E. A. B. da Silva, "Improved similarity-based modeling for the classification of rotating-machine failures," *Journal of the Franklin Institute*, vol. 355, no. 4, pp. 1913–1930, Mar. 2018, doi: 10.1016/j.jfranklin.2017.07.038.
 - [30] S. Wegerich, "Condition based monitoring using nonparametric similarity based modeling," in *Proceedings of the 3rd annual meeting of Japan Society of Maintenance, Japan*, 2006, p. 405. [Online]. Available: http://inis.iaea.org/search/search.aspx?orig_q=RN:42010976
 - [31] S. W. Wegerich, A. Wilks, and R. M. Pipke, "Nonparametric modeling of vibration signal features for equipment health monitoring," in *Proceedings of the IEEE aerospace conference*, 2003, pp. 3113–3121.
 - [32] G. N. Singh, A. K. Pandey, and A. K. Sharma, "Some improved and alternative imputation methods for finite population mean in presence of missing information," *Communications in Statistics - Theory and Methods*, vol. 50, no. 19, pp. 4401–4427, Oct. 2021, doi: 10.1080/03610926.2020.1713375.
 - [33] A. K. Pandey, G. N. Singh, N. Sayed-Ahmed, and H. Abu-Zinadah, "Improved estimators for mean estimation in presence of missing information," *Alexandria Engineering Journal*, vol. 60, no. 6, pp. 5977–5990, Dec. 2021, doi: 10.1016/j.aej.2021.04.053.
 - [34] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, p. 100164,

Mar. 2023, doi: 10.1016/j.dajour.2023.100164.

- [35] M. Seyedan, F. Mafakheri, and C. Wang, "Cluster-based demand forecasting using Bayesian model averaging: An ensemble learning approach," *Decision Analytics Journal*, vol. 3, p. 100033, Jun. 2022, doi: 10.1016/j.dajour.2022.100033.
- [36] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*, 6th ed. John Wiley & sons, 2010.
- [37] 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, p. 103589, Dec. 2024, doi: 10.1016/j.rineng.2024.103589.