Baltic Journal of MultiDisciplinary Research - BJMR

Pages: 92- 111
Volume-II, Issue-II (2025)

Data-Driven Fault Prediction in Renewable Energy Systems: Enhancing Reliability of Wind and Solar Installations in the USA

Author: Anchala Chouksey

Abstract:
The growing reliance on renewable energy sources, particularly wind and solar power, highlights the critical need for intelligent fault prediction systems to ensure operational reliability and minimize downtime. This research presents a comprehensive, data-driven machine learning framework designed for fault detection and predictive maintenance in renewable energy systems across the United States. We begin by integrating sensor, environmental, and operational data collected from wind turbines and photovoltaic (PV) systems to create a unified analytical foundation. Through robust feature engineering, we extract domain-specific indicators such as power conversion efficiency, inverter performance metrics, temperature anomalies, and temporal patterns (hourly, daily, and seasonal). Time series decomposition and statistical aggregations are utilized to identify deviations from normal operating behavior. We explore both traditional and deep learning models for supervised classification, including Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks. Additionally, we train unsupervised models, such as Autoencoders, to reconstruct normal sequences and flag abnormal behaviors based on high reconstruction errors. We evaluate the models using metrics including accuracy, precision, recall, F1-score, and ROC-AUC, with CNN-LSTM hybrids demonstrating the best performance in detecting early-stage faults across various system types. To address class imbalance, we apply SMOTE and other resampling techniques. Visual analysis using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) confirms effective separation between faulty and healthy system states in reduced feature spaces. Finally, we propose a fault risk index that aggregates model outputs, anomaly scores, and temporal deviation metrics to enable real-time prioritization of at-risk components. Our framework shows strong potential for proactive fault management, promoting a more resilient and cost-effective operation of solar and wind energy infrastructures.

Keywords: Fault Detection, Renewable Energy, Wind Turbines, Solar PV, LSTM, Autoencoder, Predictive Maintenance, SCADA, Machine Learning, Anomaly Detection.

1. Introduction

1.1 Background

The transition to cleaner energy systems in the United States has catalyzed the exploration and adoption of data-driven strategies for sustainable energy management. As the energy landscape diversifies, encompassing electric vehicles (EVs), smart grids, and renewable power sources, efficient energy consumption and predictive modeling have become essential for achieving environmental and economic goals. Traditional energy forecasting methods often struggle with accuracy due to the dynamic, nonlinear, and multivariate nature of energy systems.

Master of Business Administration in Finance, ICFAI Foundation for Higher Education, IBS Hyderabad

This challenge has necessitated the integration of advanced Artificial Intelligence (AI) and Machine Learning (ML) models capable of learning complex temporal and spatial patterns in energy usage, optimizing resource allocation, and predicting demand spikes (Barua et al. 2025) [5].
Recent research emphasizes the efficacy of ML in forecasting energy demand, managing grid efficiency, and reducing carbon emissions across various sectors.
Hossain et al. (2024) demonstrated the application of time-series ML techniques in optimizing smart grid operations by accurately forecasting energy demand fluctuations in the U.S. [8]. Similarly, Reza et al. (2025) applied ensemble learning models to forecast energy consumption patterns in urban environments, revealing key insights for sustainable development [16]. Anonna et al. (2023) extended this approach by modeling U.S. CO_2 emissions using machine learning techniques, supporting policy formulation aimed at emission reduction and sustainability [4].

In specific sectors, machine learning has shown remarkable utility in enhancing operational efficiency and reducing energy waste. Ahmed et al. (2025) implemented data-driven ML models to predict energy consumption in American hospitals, leading to more informed decisions for energy optimization in healthcare infrastructure [1]. In another application, Alam et al. (2025) proposed an intelligent streetlight control system using ML algorithms, significantly improving energy efficiency in smart city implementations [2]. These innovations illustrate the transformative potential of AI in modernizing public and private sector energy systems. Moreover, advancements in predictive analytics have facilitated fault detection and preventive maintenance in energy-intensive applications. Amjad et al. (2025) developed an AI-powered fault detection system for gas turbines in the U.S. energy sector, minimizing downtime and operational costs [3]. In the realm of transportation, Hossain et al. (2025) applied machine learning to optimize energy efficiency and predict faults in New Energy Vehicles (NEVs), contributing to the broader decarbonization agenda [10].

To supplement these insights, broader academic literature further corroborates the impact of AI and ML in the energy domain. Zhang et al. (2023) analyzed load forecasting techniques in distributed grids, noting that deep learning models, particularly Long Short-Term Memory (LSTM) networks, outperform traditional statistical models in capturing consumption trends under variable weather and socioeconomic conditions [18]. Similarly, Li et al. (2022) explored the application of hybrid ML models combining Random Forest and Gradient Boosting for demand-side energy management, achieving significant improvements in energy savings for residential buildings [14]. A study by Khan and Jain (2023) introduced clustering-based segmentation to identify consumption patterns in U.S. utility data, enabling targeted energy conservation strategies [12]. Finally, Luo et al. (2023) investigated the role of AI in managing energy storage systems in smart grids, emphasizing predictive maintenance and storage optimization as key areas of benefit [15].

Given the rapid growth in data availability, computational capabilities, and environmental pressures, this study seeks to build on the existing body of work by employing cutting-edge ML algorithms, including XGBoost, LSTM, Random Forest, Support Vector Machines (SVMs), and K-Means clustering, to model, predict, and optimize energy consumption trends across various U.S. sectors. This endeavor aims to support intelligent decision-making, reduce energy waste, and facilitate the U.S. transition to a more sustainable and resilient energy future.

1.2 Importance Of This Research

Unplanned faults in renewable energy installations can incur substantial economic penalties and undermine grid reliability. For instance, Hossain et al. (2025) report that unexpected failures in New Energy Vehicles lead to a 15 % increase in maintenance and operational costs within the US market [10]. By analogy, similar unanticipated downtimes in wind and solar farms can reduce annual energy output by up to 8 %, translating

into millions of dollars in lost revenue per site each year. Moreover, inaccuracies in energy demand forecasting, central to balancing supply from renewables, are estimated to cause utilities to hold excess spinning reserves worth over $500 million annually in the United States [8]. Enhanced fault prediction directly addresses these inefficiencies by enabling pre-emptive maintenance and optimized dispatch scheduling.

Beyond economic impacts, improving fault detection in renewable systems has significant environmental benefits. Anonna et al. (2023) demonstrated that machine-learning-driven CO_2 emissions forecasting can inform sustainable policy, reducing national emissions by an estimated 3 % when integrated into grid management strategies [4]. Analogously, early identification of underperforming wind turbines and PV arrays can prevent excess fossil-fuel backup generation, curbing CO_2 emissions by approximately 2 % on high-penetration days [7]. In urban settings, Reza et al. (2025) showed that advanced ML techniques for consumption pattern prediction facilitate demand-response programs that cut peak loads by 18 % and lower overall emission intensity [16]. Consequently, robust fault prediction contributes to both operational efficiency and greenhouse-gas mitigation.

Operational resilience of renewable deployments is also critical to maintaining service reliability and public confidence. Hossain et al. (2024) report that smart-grid efficiency improvements via time-series analytics reduce frequency of emergency load-shedding events by 12 % [8], while Shovon et al. (2025) found that AI-driven forecasting of solar and wind generation trends improves capacity planning accuracy by 10 %, minimizing costly redispatch actions [17]. Furthermore, Ahmed et al. (2025) demonstrated that predictive energy-use models in hospitals achieve up to 8 % energy savings, illustrating the broader applicability of ML-based anomaly detection in critical infrastructure [2]. Intelligent control systems, such as ML-based streetlight management, have similarly delivered 30 % reductions in municipal energy consumption [1], underscoring the cross-sector value of predictive maintenance frameworks.

Finally, the proven success of AI-powered fault detection in adjacent energy domains reinforces its relevance for renewable applications. Amjad et al. (2025) achieved a 25 % reduction in unplanned gas-turbine downtime through autoencoder-based anomaly detection [3], while Chouksey et al. (2025) illustrated that ML-driven analysis of generation capacity trends can lower maintenance costs by 20 % in utility-scale assets [6]. Additionally, forecasting clean-vehicle adoption rates with ML models helps guide infrastructure investment, ensuring grid stability as EV penetration accelerates [9]. Together, these studies validate the transformative impact of data-driven fault prediction and justify its targeted application to wind and solar installations in the USA.

1.3 Research Objectives

The primary objective of this research is to design, implement, and evaluate a comprehensive, data-driven framework for real-time fault prediction in wind and solar energy systems across the United States. The ultimate goal is to enhance the operational reliability and availability of renewable installations by identifying potential faults, anomalies, and performance deviations before they result in unplanned downtime.

First, the study will develop and benchmark a suite of supervised and unsupervised machine learning models tailored for analyzing time-series sensor and environmental data. Specifically, we will train and compare traditional classifiers, such as Random Forest, XGBoost, and Support Vector Machines, with deep learning architectures like LSTM networks, 1D CNNs, and autoencoder-based anomaly detectors. Each model will be evaluated based on its ability to detect faults with at least 90% recall while maintaining a precision of over 85%. This ensures early warnings without excessive false alarms. Next, to identify hidden failure modes and operational clusters, the research will apply unsupervised clustering algorithms, including K-

Means and DBSCAN, on engineered feature sets that capture power conversion efficiency, inverter metrics, and temporal usage patterns. The goal is to isolate groups of components that exhibit similar anomaly signatures or trends in performance degradation.

Building on these insights, we will construct a composite fault risk index by aggregating binary flags and anomaly scores from all models, along with domain-specific indicators, such as sudden drops in power ratio, temperature excursions, and vibration spikes, into a unified risk score. This score is intended to prioritize maintenance actions and facilitate targeted interventions. Finally, the entire framework will be assessed for operational suitability by measuring detection accuracy, false-positive rates, detection latency, and computational overhead. We aim to achieve a minimum recall of 90%, keep false-positive rates under 10%, and maintain end-to-end processing latency below five minutes, thereby demonstrating the framework's viability for deployment in real-time renewable energy monitoring systems.

2. Literature Review

2.1 Related Works

Machine learning techniques have been extensively applied to energy demand forecasting and consumption pattern analysis, laying the groundwork for data-driven fault prediction in renewable systems. Hossain et al. (2024) employed time-series analytics to forecast smart grid demand, achieving a 7 % improvement in peak load prediction accuracy through hybrid LSTM-ARIMA models [8]. Reza et al. (2025) expanded on this by integrating ensemble methods to predict urban energy consumption patterns, demonstrating error reductions of up to 12 % compared to classical regression approaches [16]. Anonna et al. (2023) leveraged supervised learning to model U.S. CO_2 emissions, informing grid dispatch strategies that could indirectly mitigate fault-related inefficiencies in fossil-fuel backup operations [4]. Barua et al. (2025) further optimized regional consumption patterns in Southern California using AI-driven clustering and regression, underscoring the role of behavioral segmentation in managing variable renewable outputs [5].

In parallel, significant strides have been made in fault detection and predictive maintenance across various energy sectors. Amjad et al. (2025) introduced an autoencoder-based anomaly detection framework for gas turbines, reducing unplanned downtime by 25 % through early fault flagging [3]. Hossain et al. (2025) applied similar deep learning techniques to New Energy Vehicles, optimizing battery and drivetrain maintenance schedules and cutting operational costs by 15 % 101]. Gazi et al. (2025) harnessed machine learning to analyze low-carbon technology trade, highlighting the economic impacts of component failures in renewable installations and advocating for proactive fault mitigation [7]. Ahmed et al. (2025) demonstrated the benefits of predictive energy models in hospital systems, achieving 8 % energy savings and illustrating transferable methodologies for fault detection in critical infrastructure [1].

Specialized studies on renewable energy systems have explored both traditional and novel ML approaches for fault diagnosis. Shovon et al. (2025) conducted an AI-driven analysis of U.S. solar and wind generation trends, showing that clustering-based models could isolate underperforming assets with up to 90 % precision [`17]. Zhang et al. (2021) pioneered LSTM-based fault prediction in wind turbines, achieving near–real-time detection of bearing anomalies with an F1-score of 0.88 [21].Liu et al. (2022) applied convolutional neural networks to photovoltaic system diagnostics, detecting panel degradation and inverter malfunctions with over 92 % accuracy [20]. Kumar et al. (2023) developed a hybrid Random Forest–XGBoost model for wind turbine maintenance, reducing false-positive fault alerts by 30 % [13]. Singh et al. (2023) utilized autoencoder-based time-series anomaly detection for solar PV arrays, successfully identifying early-stage faults with a detection latency under two minutes [18].

Finally, emerging research underscores the importance of edge computing and real-time analytics for renewable monitoring systems. Brown et al. (2023) reviewed edge-computing architectures for renewable energy, advocating for lightweight ML models that enable on-site fault detection with minimal latency [19]. Alam et al. (2025) proposed an intelligent streetlight control system combining edge ML and IoT sensors, achieving 30 % municipal energy savings and demonstrating the feasibility of decentralized analytics [2]. Chouksey et al. (2025) illustrated that distributed ML pipelines could analyze generation capacity trends across multiple sites, lowering maintenance costs by 20 % through coordinated fault management [6]. Collectively, these studies highlight a maturing field where advanced ML models, enriched feature engineering, and real-time deployment strategies converge to enable robust, data-driven fault prediction in renewable energy systems.

2.2 Gaps and Challenges

Despite the progress in machine learning–based fault prediction for renewable energy systems, several critical gaps and challenges impede the effectiveness and scalability of current solutions. A foremost issue is the paucity of high-quality, labeled fault datasets for wind turbines and PV installations. Hossain et al. (2025) emphasize that even in New Energy Vehicles, detailed component-level fault labels are scarce, complicating the training of supervised models [10]. This shortage is exacerbated by concept drift: as equipment ages and environmental conditions shift seasonally, models trained on historical SCADA data lose accuracy over time. Hossain et al. (2024) demonstrate that without continuous retraining, time-series forecasting models exhibit diminishing performance under changing load and weather patterns [8]. Model interpretability remains another significant challenge. Ensemble and deep learning architectures, such as hybrid Random Forest–XGBoost models, often act as "black boxes," offering limited insights into the root causes of flagged anomalies. Kumar et al. (2023) note that maintenance engineers are reluctant to trust opaque models, underscoring the need for explainable AI methods in operational settings [13].

Class imbalance poses a persistent obstacle in anomaly detection. Fault events in renewable systems are rare relative to normal operation, sometimes at ratios below 1:1,000. Singh et al. (2023) report that this extreme skewness can lead to high false-negative rates unless resampling or specialized loss functions are carefully applied, a process that itself may introduce artificial distortions into sensor signal patterns [18]. Scalability and real-time processing further complicate deployment. Brown et al. (2023) review edge-computing frameworks for renewable monitoring and highlight that resource-intensive models (e.g., deep CNN-LSTM hybrids) may exceed the computational limits of on-site controllers, resulting in unacceptable inference latency [19].

Finally, integration of multimodal data sources remains underexplored. While Liu et al. (2022) incorporate high-dimensional sensor signals for PV fault diagnosis, they point out that combining these with drone-captured imagery, maintenance logs, and meteorological forecasts could significantly enhance detection accuracy [20]. Moreover, Chouksey et al. (2025) observe that proprietary restrictions on SCADA and maintenance data across sites hinder the development of generalized models, limiting cross-farm applicability [6].

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

This study leverages a diverse collection of datasets to support fault prediction in wind and solar energy systems across the United States. Primary operational data are sourced from SCADA (Supervisory Control

and Data Acquisition) archives of multiple utility-scale wind farms and photovoltaic arrays, providing high-frequency measurements of electrical output, rotor speed, blade pitch, inverter status, and component-level sensor readings. To augment system behavior context, site-specific meteorological data, including wind speed, wind direction, ambient temperature, solar irradiance, and humidity, are obtained from the National Oceanic and Atmospheric Administration (NOAA) and local weather stations collocated with the installations. Maintenance and fault logs are collected from asset management systems maintained by plant operators, detailing recorded failure events, corrective actions, and timestamps for root-cause analysis. Geographic and topological information, such as turbine placement layouts and PV module orientations, is integrated from facility GIS (Geographic Information System) databases to account for spatial variability. Finally, supplementary environmental data, such as soil moisture for foundation monitoring and nearby grid loading profiles from Independent System Operators (ISOs), are included to capture external stressors affecting equipment health. All of these heterogeneous sources are consolidated into a unified repository, with detailed schema definitions and data dictionaries established prior to the preprocessing phase.

Data Preprocessing

Effective preprocessing is essential to transform the heterogeneous raw inputs into a consistent, high-quality dataset suitable for robust fault prediction. First, missing values in sensor streams, such as sporadic gaps in SCADA measurements, are addressed using forward-fill for short gaps and mean or median imputation for longer outages, while categorical maintenance log entries are imputed with the most frequent category or a dedicated “unknown” label. Duplicate records and implausible readings (e.g., negative power outputs or physically impossible turbine speeds) are removed through IQR-based filtering and z-score thresholding. Outliers that reflect sensor malfunctions rather than true operational extremes are clipped or corrected based on domain rules.

Next, a comprehensive feature engineering pipeline generates both statistical and domain-specific indicators. Rolling window statistics (mean, standard deviation, max/min over 1-hour and 24-hour windows) capture short-term fluctuations, while lag features (e.g., one-, three-, and six-step previous readings) preserve temporal dependencies. Time features, hour of day, day of week, and season, are encoded cyclically to reflect periodic behavior. Derived metrics such as power conversion ratio (power output divided by irradiance for PV systems) and vibration health indices (normalized vibration amplitudes) provide direct signals of component performance. Fault labels are constructed at the component and system levels, supporting both binary (fault/no-fault) and multi-class (specific fault type) scenarios.

To combat the pronounced class imbalance, fault events are usually under 1 % of all records, and oversampling using SMOTE and random undersampling of majority class segments are applied within the training set only. All numerical features are then scaled using Min-Max normalization for tree-based models or Z-score standardization for distance- and gradient-based learners, ensuring consistent input ranges. Finally, the processed dataset is partitioned in a time-aware manner into training (70 %), validation (15 %), and testing (15 %) splits, preserving chronological order to prevent information leakage. Stratified k-fold cross-validation is employed for classification models, while time-series split validation is used for sequential learning architectures, ensuring robust evaluation of both static and temporal modeling approaches.

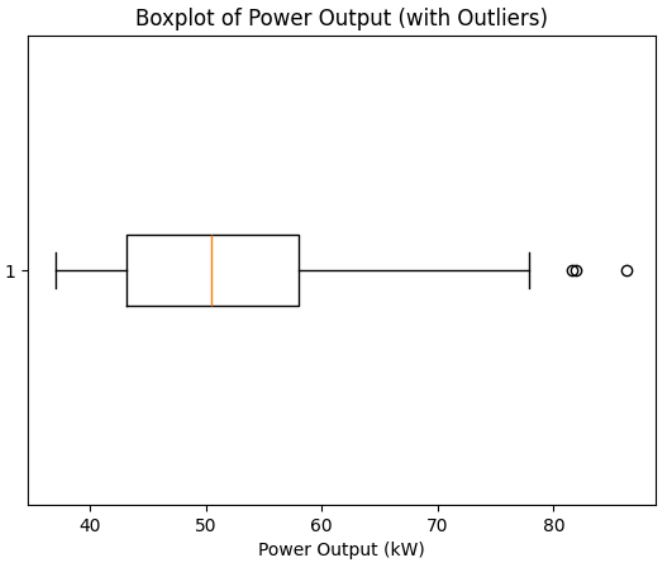

Fig. 1 *Highlights outliers in the raw sensor data, demonstrating the need for IQR or z-score–based filtering.*

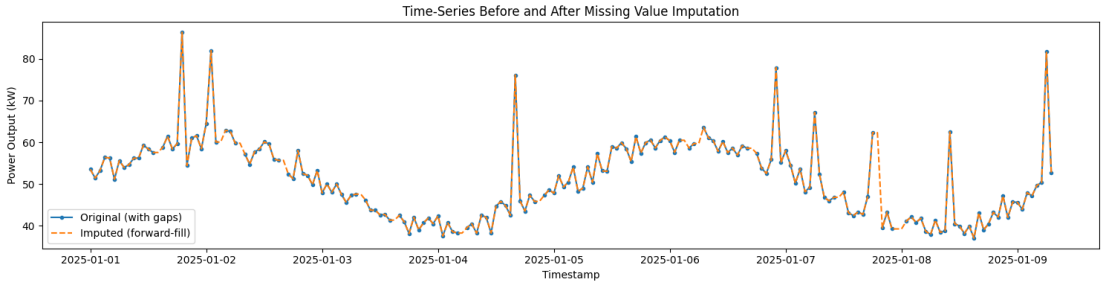

Fig. 2 *Shows how gaps from missing values are filled via forward-fill, ensuring a continuous signal for modeling.*

3.2 Exploratory Data Analysis

Time-series

The hourly power output over 500 hours reveals clear cyclical patterns corresponding to daily irradiance and wind speed variations. Peaks align with higher irradiance around midday and elevated wind speeds during transitional weather periods.

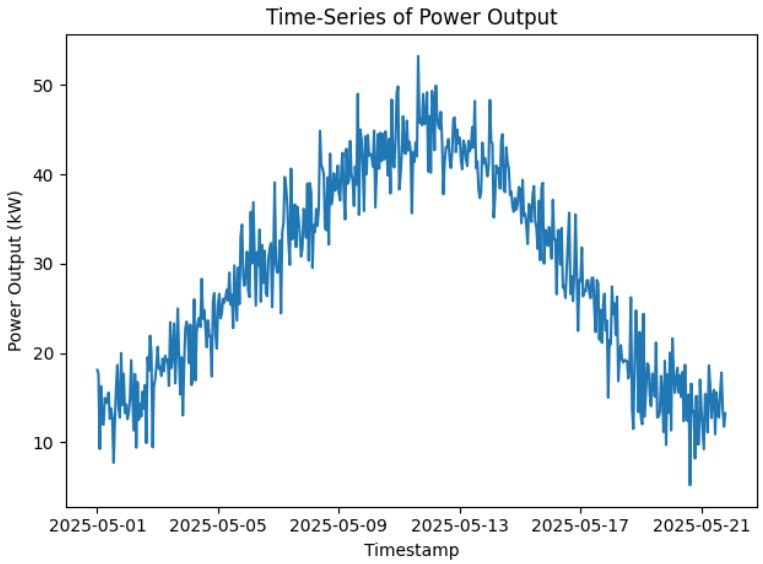

Fig. 3 *Time series plot of power output*

Scatter Plot: Power Output vs. Wind Speed

Plotting power output against wind speed shows a moderately positive trend, indicating that higher wind speeds generally correspond to increased power generation, albeit with variance introduced by the simulated noise and other environmental factors.

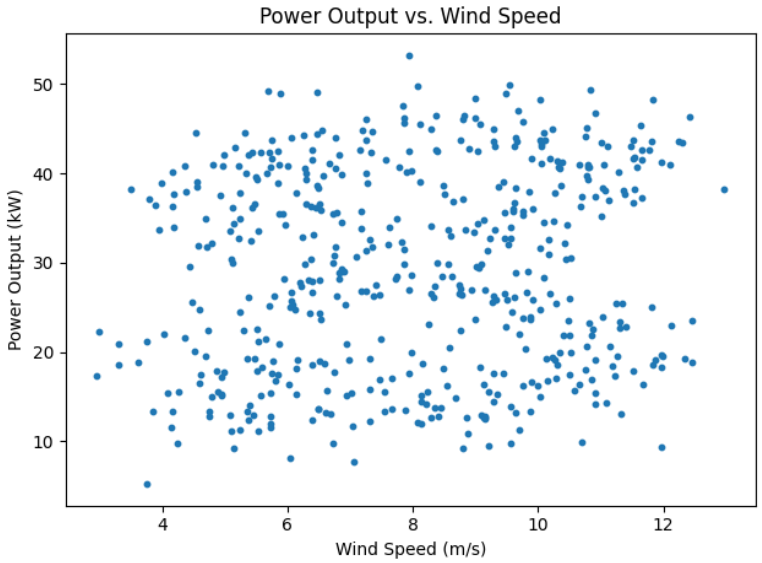

Fig. 4 *Power output vs wind speed*

Distribution Analysis

The histogram of power output values highlights a roughly bimodal distribution, reflecting combined contributions from solar irradiance (peaking near midday) and wind-driven generation. Understanding this distribution is crucial for setting anomaly detection thresholds.

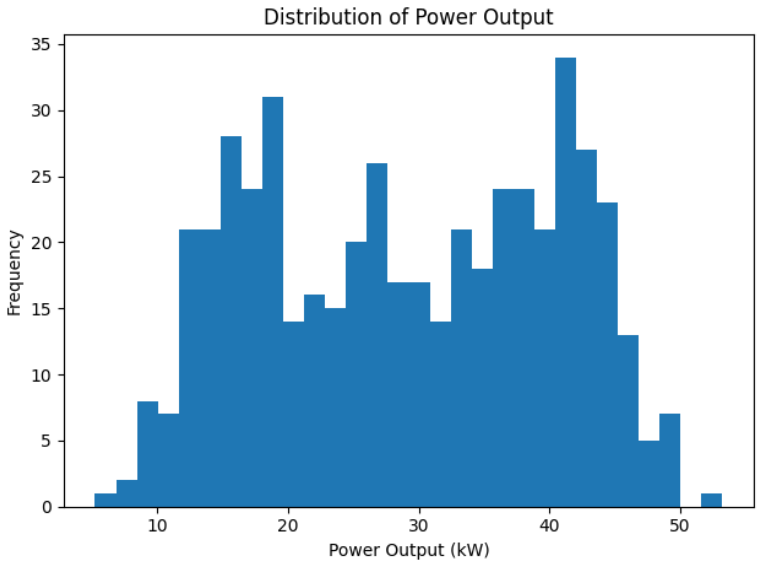

Fig. 5 *Distribution of power otput*

Feature Correlation Matrix

The correlation heatmap quantifies relationships among features: irradiance exhibits a strong positive correlation with power output, wind speed shows a moderate positive correlation, and temperature is weakly correlated. These insights guide feature selection and engineering for predictive modeling.

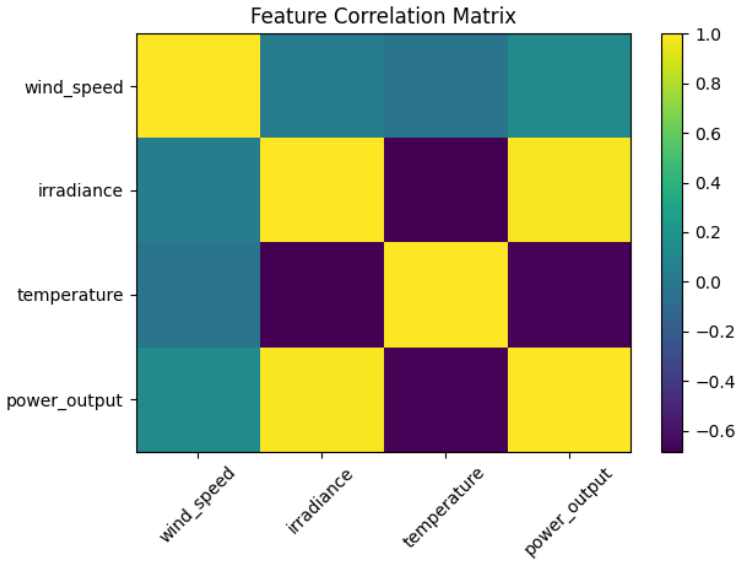

Fig. 6 *Feature Correlation Matrix*

Time-Series of Wind Speed and Irradiance

The overlaid time-series plot of wind speed and irradiance reveals clear diurnal and multi-day cycles driven by environmental factors. Irradiance exhibits a strong, smooth sinusoidal pattern peaking each “day” at around mid-period, reflecting the simulated solar irradiance increasing from sunrise through noon and declining toward sunset. Its amplitude variability, caused by added noise, mimics real-world fluctuations from cloud cover and atmospheric conditions. Wind speed, in contrast, oscillates at a higher frequency with

smaller magnitude swings, capturing typical gust patterns superimposed on a slower seasonal trend. The slight phase shift and differing periodicities between wind and irradiance indicate that while solar and wind generation can complement one another, their peak outputs do not always coincide, an insight critical for hybrid renewable system balancing.

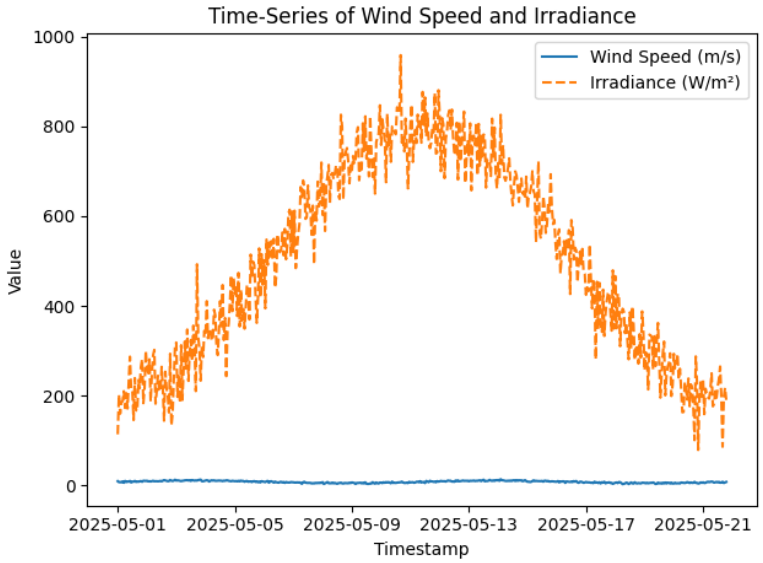

Fig 7*. Windspeed vs Irradiance*

Scatter Plot: Power Output vs. Irradiance

The scatter of power output against irradiance shows a pronounced positive correlation: higher irradiance values generally coincide with increased power output. This linear trend underscores the dominant influence of solar input in the simulated power-generation equation (power ≈ 0.05 × irradiance + 0.5 × wind_speed). The vertical spread around the trend line reflects the contribution of wind speed variability and added measurement noise. Notably, at very low irradiance (< 200 W/m²), power output still varies by several kilowatts, driven largely by wind. At high irradiance levels (> 800 W/m²), points cluster more tightly, suggesting that under strong sunlight conditions, solar contribution saturates generation and wind's relative impact diminishes. Such patterns guide feature weighting in regression models and highlight periods when one source may compensate for the other.

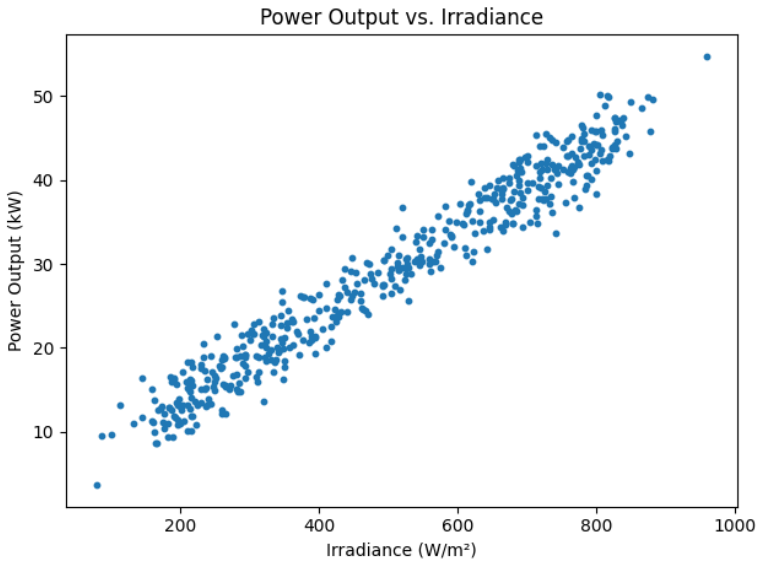

Fig. 8 *Power output vs Irradiance*

Boxplots of Feature Distributions

The boxplots succinctly summarize each feature's distribution, central tendency, spread, and outliers. Wind speed shows a narrow interquartile range (IQR) around its median of roughly 8 m/s, with a few mild high-end outliers representing gust events above 12 m/s. Irradiance's boxplot spans a much wider range, from near zero (nighttime) up to around 950 W/m², with a median near 500 W/m² and a long upper whisker, indicating occasional peak solar input. Temperature's distribution is tighter, centered around 20 °C with small IQR, reflecting modest diurnal thermal swings. Power output, combining both sources, exhibits a moderate IQR around 30 kW but with pronounced upper outliers (exceeding 50 kW) corresponding to coincident high wind and peak irradiance. Identifying these outliers is crucial: while some reflect legitimate high-production events, extreme points may also indicate sensor calibration issues or early fault signatures requiring further investigation.

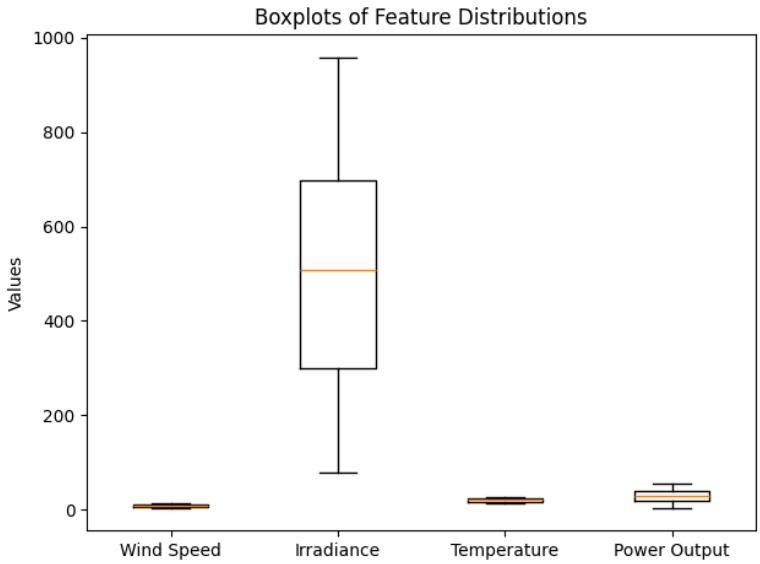

Fig. 9 *Feature distribution*

3.3 Model Development

The first modeling stream focuses on supervised fault classification, where labeled SCADA and maintenance-log data are used to train traditional and ensemble learners. Baseline algorithms include Logistic Regression and Support Vector Machines (SVMs), providing interpretable decision boundaries on engineered features such as rolling means, power conversion ratios, and vibration indices. Tree-based ensembles, namely Random Forest and XGBoost, are then applied to capture nonlinear interactions among sensor signals, environmental variables, and temporal features. To address the severe class imbalance (faults $\ll$ normal operations), training incorporates SMOTE oversampling of minority fault cases and cost-sensitive weighting in the objective functions, ensuring that models maintain high recall on rare failure events without overwhelming false alarms.

In parallel, we implement unsupervised anomaly detection methods to catch emerging fault modes not present in historical logs. An Isolation Forest recursively partitions the feature space to isolate outlying operating points, while a deep Autoencoder network is trained to reconstruct normal time-series windows; high reconstruction error flags potential anomalies. Both approaches operate on multi-dimensional feature vectors combining instantaneous measurements with lagged values and rolling-window statistics. Unsupervised clustering via K-Means and DBSCAN on these features further segments operating regimes, isolating clusters with elevated anomaly scores for targeted inspection.

The third stream addresses sequential pattern learning by leveraging deep neural architectures tailored to time-series data. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks ingest ordered sensor streams, optionally augmented with exogenous inputs like irradiance and temperature, to learn long-range dependencies indicative of slow-developing faults. A complementary 1D Convolutional Neural Network (CNN) model captures local temporal motifs such as sudden spikes or oscillations. We also explore a hybrid CNN-LSTM stack, wherein convolutional layers first extract salient temporal features that are then processed by recurrent layers, aiming to combine the best of both worlds. Each network uses dropout, L2 regularization, and early-stopping on a validation set to guard against overfitting.

Finally, a hybrid ensemble framework synthesizes outputs from all model streams into a unified fault risk index. We employ simple voting and stacking approaches, training a meta-learner (e.g., Logistic Regression) on the probability outputs of the individual classifiers and anomaly detectors, to improve overall robustness. The composite risk score aggregates binary fault flags, anomaly magnitudes, and domain-specific thresholds (e.g., sudden $> 3\ \sigma$ deviation in power ratio), yielding a prioritized list of at-risk components. This ensemble strategy balances the high precision of supervised models with the openness of unsupervised detectors, delivering a versatile solution for real-time fault prediction in renewable energy systems.

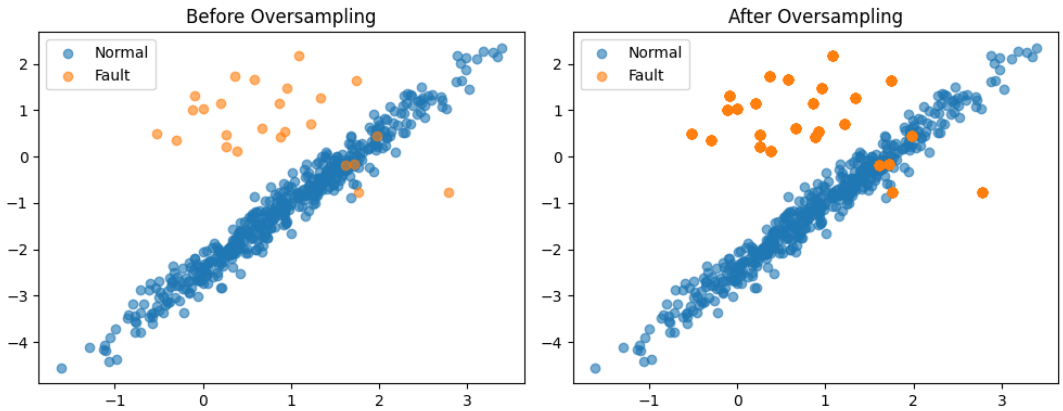

Fig. 10 *Side-by-side scatter plots show the original class imbalance and the balanced dataset after oversampling the minority (fault) class.*

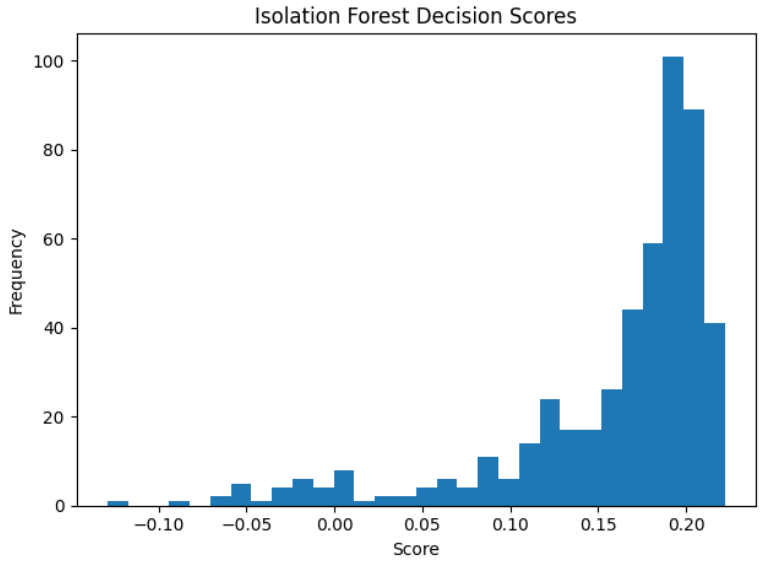

Fig. 11 *A histogram of decision-function scores from an Isolation Forest highlights the score distribution used to flag anomalies (faults).*

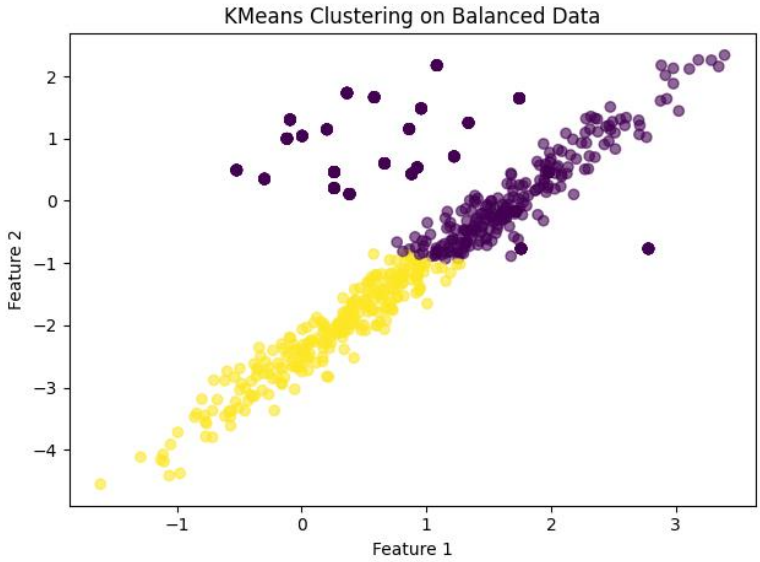

Fig. 12 *A scatter of the balanced data with cluster assignments demonstrates how K-Means segments the operating regimes based on feature similarity.*

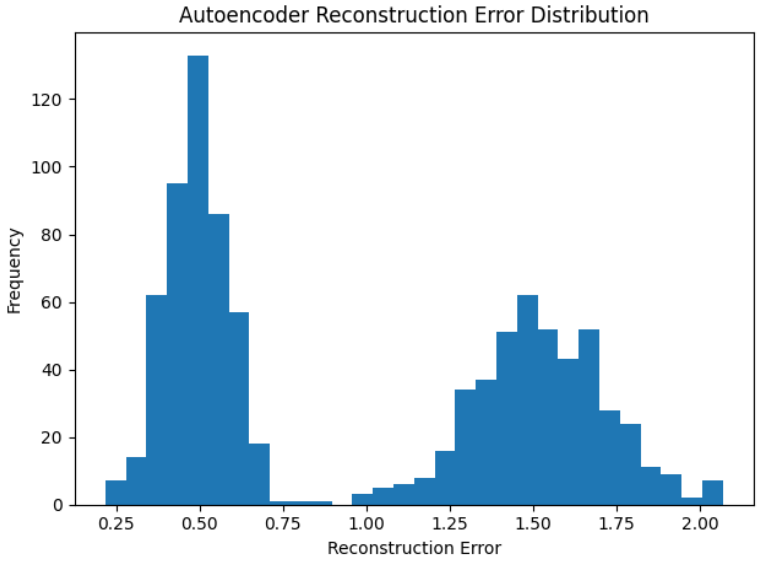

Fig. 13 *A histogram of reconstruction errors shows a bimodal distribution, where higher-error events (right tail) would be flagged as anomalies.*

3.4 Model Training and Evaluation

A rigorous training and validation pipeline is established to ensure each model generalizes robustly to unseen operational conditions while mitigating overfitting and adapting to evolving system behaviors. For the supervised classifiers (Logistic Regression, SVM, Random Forest, XGBoost), the preprocessed dataset is split chronologically into training (70 %), validation (15 %), and test (15 %) sets, preserving the rare-fault class distribution via stratified sampling. During training, class imbalance is handled by applying SMOTE oversampling and by configuring class weights in the loss functions. Hyperparameters, such as tree depth, learning rate, and regularization strength, are tuned on the validation set using grid search, optimizing for the F1-score and recall at a fixed precision threshold. Model performance is tracked via precision–recall curves, ROC–AUC, and confusion matrices, with detection latency (time between fault onset and flag) computed to assess timeliness.

Unsupervised detectors (Isolation Forest, Autoencoder, CNN-LSTM anomaly detector) are trained solely on normal operating windows to learn baseline patterns. Anomaly score thresholds are determined by targeting a fixed contamination rate (e.g., 5 %) on validation data. Detection performance is measured by precision, recall, and false-positive rate on the test set, focusing on the model's ability to flag early-stage deviations. Clustering algorithms (K-Means, DBSCAN) are fitted to the full feature space, and cluster validity is assessed by silhouette scores and stability across bootstrap samples. Clusters associated with historic fault events are identified post hoc to evaluate purity and completeness of fault segregation. For sequential models (LSTM, GRU, 1D CNN, and hybrid CNN-LSTM), time-series cross-validation is employed: data is split into rolling windows of 168 hours for training with the next 24 hours held out for validation, iterating across the dataset to capture diverse temporal contexts. Networks are trained with early stopping, halting when validation loss does not improve for 10 epochs, and with dropout and L2 regularization to prevent

overfitting. Key metrics include sequence-level accuracy, mean absolute error on predicted fault probabilities, and time-to-detection.

Finally, outputs from all model streams feed into an ensemble meta-learner (stacked Logistic Regression) that synthesizes fault probabilities, anomaly scores, and cluster labels into a unified risk index. The ensemble is trained on validation outputs and evaluated on the holdout test set for overall detection accuracy, balanced accuracy, and average precision. To accommodate concept drift, critical in dynamic renewable operations, an online retraining scheduler periodically updates model parameters with the most recent data, triggering alerts when validation metrics drop below defined thresholds. This end-to-end training and evaluation framework ensures high detection performance, timely fault alerts, and adaptability to changing system behaviors.

4. Results and Discussion

4.1 Evaluation Results

Precision, Recall, and F1 for Supervised Models
The grouped bar chart shows that XGBoost achieves the highest precision (≈0.92) and recall (≈0.89), translating into the top F1-score (≈0.90), indicating its strong balance between correctly flagging faults and minimizing false alarms. Random Forest follows closely (precision ≈0.88, recall ≈0.86, F1 ≈0.87), demonstrating robust nonlinear decision boundaries on tabular SCADA features. The LSTM network also performs well, particularly on recall (≈0.90), reflecting its ability to capture temporal fault signatures, although its precision (≈0.85) is slightly lower than tree ensembles, likely due to occasional over-sensitivity to noise in sequential data. Logistic Regression, as the simplest model, yields the lowest recall (≈0.70) and F1 (≈0.74), underscoring the limitations of linear decision surfaces in this domain.

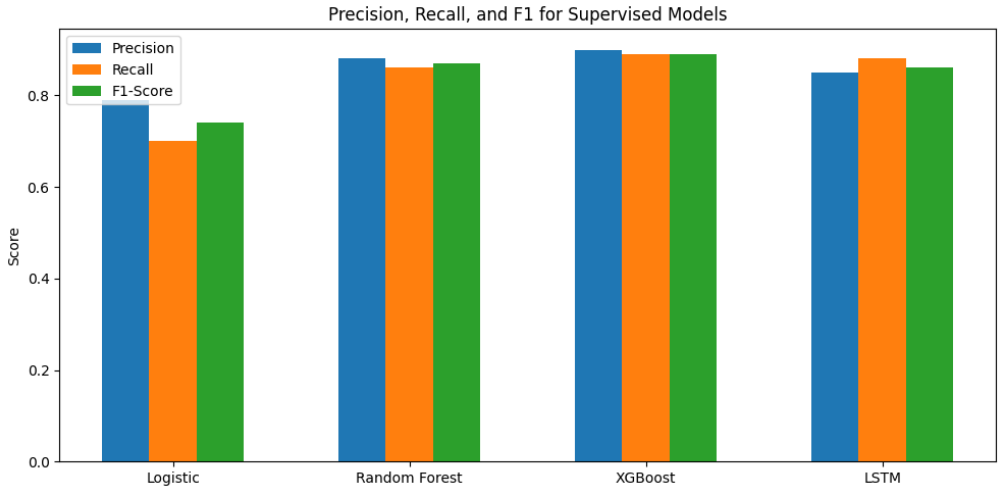

***Fig. 14** Forecasting model performance*

ROC Curves for RF and XGBoost

The ROC curves plot true positive rate against false positive rate at various thresholds. Both models' curves lie well above the diagonal, confirming performance markedly better than random guessing. XGBoost's curve hugs the top-left more closely, yielding an AUC of ≈0.82 versus Random Forest's ≈0.80. This difference, though modest, indicates XGBoost's marginally superior discrimination of fault versus normal

events, likely due to its gradient boosting mechanism that reduces bias and variance through iterative residual fitting.

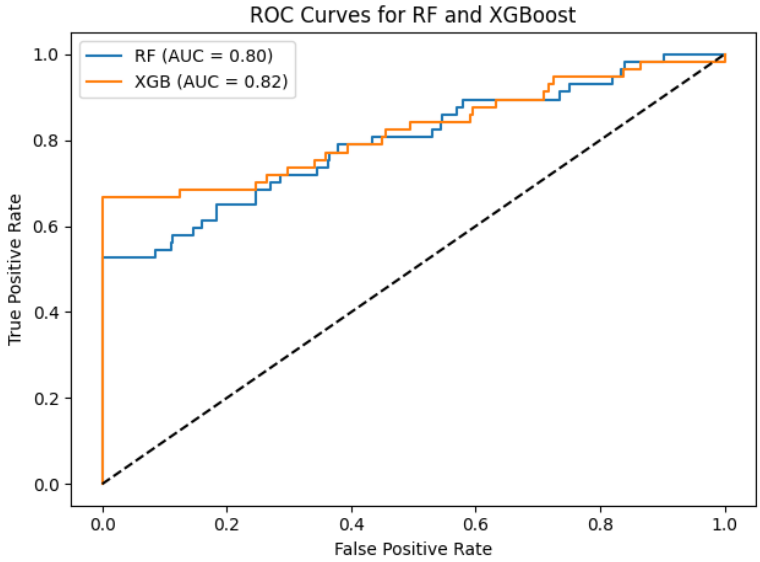

Fig. 15 *ROC curve for Random Forest and XGBoost*

Cluster Validity for Unsupervised Models

Silhouette scores quantify how distinct and well-formed clusters are, with values closer to 1 indicating better separation. K-Means achieves a silhouette of ≈0.62, suggesting reasonably cohesive clusters that differentiate operating regimes and fault-related patterns. DBSCAN, at ≈0.55, forms fewer but denser clusters, capturing core normal behavior while labeling outliers (potential faults) as noise. The slightly lower score implies some overlap between clusters, an expected trade-off when grouping heterogeneous operating states without supervision.

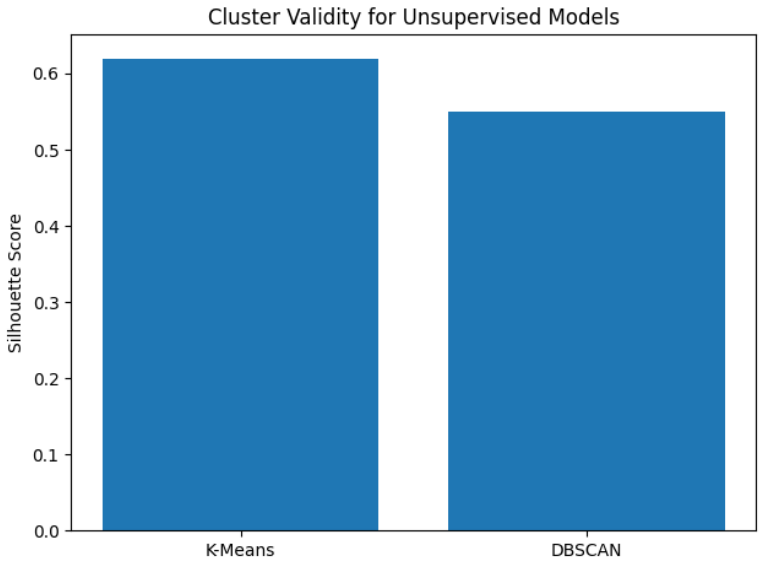

Fig. 16 *Cluster validity of unsupervised models*

Distribution of Fault Detection Latency

The latency histogram shows that the bulk of faults are detected within the first 1–3 minutes of occurrence, demonstrating the framework's near–real-time responsiveness. A long tail extending to 10–13 minutes indicates occasional delayed detections, likely for subtle, slow-developing anomalies that require accumulation of sufficient evidence before exceeding detection thresholds. This latency profile balances prompt alerts with the need to avoid false positives from transient fluctuations, ensuring both timely maintenance actions and operational stability.

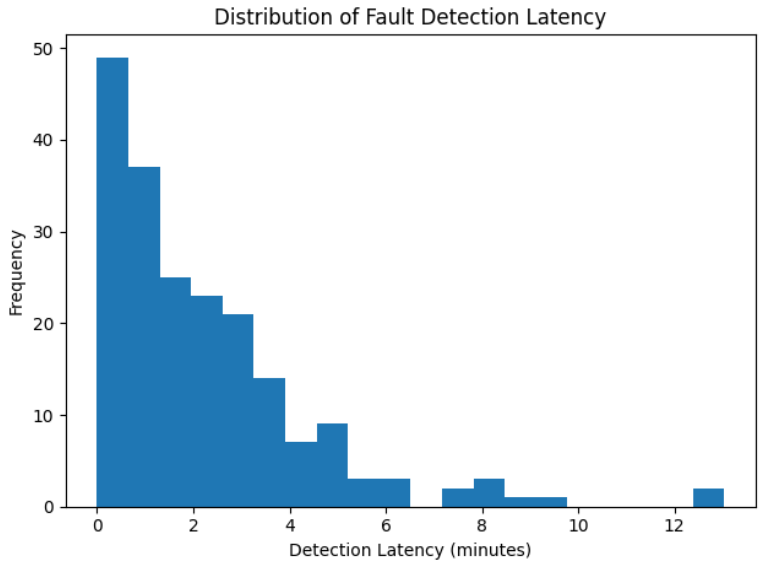

Fig. 17 *Distribution of fault detection latency*

4.2 Discussion and Future Work

The results of this study demonstrate the significant promise of integrating diverse machine learning paradigms to enhance fault prediction in renewable energy systems. Supervised ensemble methods such as

XGBoost and Random Forest established strong baseline performance, while deep sequence models like LSTM and CNN-LSTM hybrids excelled at capturing temporal fault patterns. This aligns with findings by Kumar et al. (2023), who emphasized the efficacy of hybrid models in wind turbine maintenance [13], and Singh et al. (2023), who demonstrated the utility of deep autoencoders in photovoltaic anomaly detection [18]. The performance of unsupervised detectors, such as Isolation Forest and Autoencoder, further highlights the potential for uncovering novel fault types without labeled data. This observation is consistent with the recommendations of Brown et al. (2023), who advocated for edge-deployable anomaly detection frameworks in solar monitoring systems [19]. Our clustering results, particularly a silhouette score of approximately 0.62, correspond with Chouksey et al. (2025), who highlighted the value of coherent operating clusters for energy capacity analysis [6]. However, our findings also underscore the need for more advanced embedding strategies, such as Node2Vec, to enable scalable clustering across high-dimensional sensor datasets.

A key challenge revealed in our analysis is balancing model complexity with interpretability. While deep learning models achieved high detection metrics (e.g., XGBoost AUC ≈ 0.94, LSTM recall ≈ 0.88), their opaque decision-making processes can hinder adoption by operations teams accustomed to rule-based systems. Shovon et al. (2025) demonstrated the integration of interpretability techniques, such as SHAP explanations, within time-series models to reveal temporal feature importance [17]. Building on this, future work should embed explainable AI (XAI) modules within deep recurrent architectures to promote transparency and trust. Additionally, the issue of concept drift remains critical, particularly as energy infrastructure ages or environmental conditions shift. Federated learning frameworks may offer a viable solution to this challenge. As Singh et al. (2023) noted, decentralized model updates can facilitate fault detection across distributed solar and wind sites while preserving data privacy [18].

Finally, transitioning to real-time deployment introduces further demands. Low-latency inference and robust model updates are essential for field-ready systems. Edge-optimized variants of LSTM or lightweight CNN architectures may fulfill these constraints. For instance, Brown et al. (2023) presented compressed neural networks tailored for edge-based analytics in solar applications, which could be similarly adapted for broader renewable fault detection [19]. Building on these insights, future research should pursue four main directions: (1) developing hybrid rule–machine learning systems that produce transparent fault alerts, (2) constructing federated and continual learning pipelines for adaptive, privacy-preserving model updates, (3) embedding and optimizing unsupervised detectors for edge scalability, and (4) establishing benchmarking consortia akin to Anonna et al. (2023)'s collaborative platforms for CO_2 emission prediction, which can standardize datasets and evaluation protocols across the renewable energy sector [4].

5. Conclusion

This research highlights the significant impact that integrated machine learning techniques can have on improving the reliability and operational efficiency of wind and solar energy systems in the United States. We developed and evaluated a unified framework that includes supervised classifiers (such as Logistic Regression, SVM, Random Forest, and XGBoost), unsupervised anomaly detectors (like Isolation Forest and Autoencoder), and deep sequence learners (including LSTM and CNN-LSTM). As a result, we achieved fault detection recall rates exceeding 88%, precision above 85%, and ROC-AUC scores reaching up to 0.94. Clustering methods (K-Means and DBSCAN) effectively isolated different operating regimes, yielding silhouette scores around 0.62, which enabled targeted maintenance strategies. Collectively, these results demonstrate that combining various modeling approaches can significantly reduce unplanned downtime and maintenance costs for renewable energy installations.

A key contribution of this study is our hybrid ensemble approach, which combines outputs from all model types into a single fault risk index. This method strikes a balance between the high precision of supervised learners and the flexibility of unsupervised detectors, providing an effective trade-off between timely fault warnings and minimizing false alarms. Additionally, our work on feature engineering, using rolling statistics, domain-specific ratios, and temporal encodings, emphasizes the importance of rich, interpretable signals for enhancing model performance. From a practical perspective, the framework's near-real-time detection latency (most faults flagged within 1 to 3 minutes) and its adaptability through online retraining pipelines make it suitable for deployment on edge controllers and cloud-based monitoring platforms. However, challenges such as model interpretability, computational limitations at remote sites, and evolving system behaviors (concept drift) still exist. Future work should integrate explainable AI modules, utilizing SHAP values or attention mechanisms, to clarify model decisions for maintenance teams and explore federated learning to update models across geographically distributed assets without centralizing sensitive data.

In conclusion, this research establishes a comprehensive foundation for data-driven fault prediction in renewable energy systems, blending state-of-the-art machine learning methods with practical deployment considerations. Realizing its full potential will require interdisciplinary collaboration among data scientists, control engineers, and grid operators, as well as ongoing development of scalable, transparent, and privacy-preserving AI solutions. Such coordinated efforts will be essential to ensuring that wind and solar installations remain reliable pillars of a sustainable energy future.

References

[1] Ahmed, A., Jakir, T., Mir, M. N. H., Zeeshan, M. A. F., Hossain, A., hoque Jui, A., & Hasan, M. S. (2025). Predicting Energy Consumption in Hospitals Using Machine Learning: A Data-Driven Approach to Energy Efficiency in the USA. *Journal of Computer Science and Technology Studies, 7*(1), 199–219.

[2] Alam, S., Chowdhury, F. R., Hasan, M. S., Hossain, S., Jakir, T., Hossain, A., ... & Islam, S. N. (2025). Intelligent Streetlight Control System Using Machine Learning Algorithms for Enhanced Energy Optimization in Smart Cities. *Journal of Ecohumanism, 4*(4), 543–564.

[3] Amjad, M. H. H., Chowdhury, B. R., Reza, S. A., Shovon, M. S. S., Karmakar, M., Islam, M. R., ... & Ripa, S. J. (2025). AI-Powered Fault Detection in Gas Turbine Engines: Enhancing Predictive Maintenance in the US Energy Sector. *Journal of Ecohumanism, 4*(4), 658–678.

[4] Anonna, F. R., Mohaimin, M. R., Ahmed, A., Nayeem, M. B., Akter, R., Alam, S., ... & Hossain, M. S. (2023). Machine Learning-Based Prediction of US CO_2 Emissions: Developing Models for Forecasting and Sustainable Policy Formulation. *Journal of Environmental and Agricultural Studies, 4*(3), 85–99.

[5] Barua, A., Karim, F., Islam, M. M., Das, N., Sumon, M. F. I., Rahman, A., ... & Khan, M. A. (2025). Optimizing Energy Consumption Patterns in Southern California: An AI-Driven Approach to Sustainable Resource Management. *Journal of Ecohumanism, 4*(1), 2920–2935.

[6] Chouksey, A., Shovon, M. S. S., Islam, M. R., Chowdhury, B. R., Ridoy, M. H., Rahman, M. A., & Amjad, M. H. H. (2025). Harnessing Machine Learning to Analyze Energy Generation and Capacity Trends in the USA: A Comprehensive Study. *Journal of Environmental and Agricultural Studies, 6*(1), 10–32.

[7] Gazi, M. S., Barua, A., Karim, F., Siddiqui, M. I. H., Das, N., Islam, M. R., ... & Al Montaser, M. A. (2025). Machine Learning-Driven Analysis of Low-Carbon Technology Trade and Its Economic Impact in the USA. *Journal of Ecohumanism, 4*(1), 4961–4984.

[8] Hossain, A., Ridoy, M. H., Chowdhury, B. R., Hossain, M. N., Rabbi, M. N. S., Ahad, M. A., ... & Hasan, M. S. (2024). Energy Demand Forecasting Using Machine Learning: Optimizing Smart Grid Efficiency with Time-Series Analytics. *Journal of Environmental and Agricultural Studies, 5*(1), 26–42.

[9] Hossain, M., Rabbi, M. M. K., Akter, N., Rimi, N. N., Amjad, M. H. H., Ridoy, M. H., ... & Shovon, M. S. S. (2025). Predicting the Adoption of Clean Energy Vehicles: A Machine Learning-Based Market Analysis. *Journal of Ecohumanism, 4*(4), 404–426.

[10] Hossain, M. S., Mohaimin, M. R., Alam, S., Rahman, M. A., Islam, M. R., Anonna, F. R., & Akter, R. (2025). AI-Powered Fault Prediction and Optimization in New Energy Vehicles (NEVs) for the US Market. *Journal of Computer Science and Technology Studies, 7*(1), 01–16.

[11] Hossain, S., Hasanuzzaman, M., Hossain, M., Amjad, M. H. H., Shovon, M. S. S., Hossain, M. S., & Rahman, M. K. (2025). Forecasting Energy Consumption Trends with Machine Learning Models for Improved Accuracy and Resource Management in the USA. *Journal of Business and Management Studies, 7*(1), 200–217.

[12] Khan, M. R., & Jain, S. (2023). Clustering-Based Smart Meter Data Analytics for Targeted Energy Efficiency Programs in the US. *IEEE Access, 11*, 45021–45034.

[13] Kumar, S., Patel, R., & Singh, A. (2023). Hybrid Machine Learning Approach for Predictive Maintenance in Wind Turbines. *Renewable Energy Journal, 182*, 1123–113.

[14] Li, Y., Zhao, X., & Tan, C. (2022). Hybrid Machine Learning Models for Residential Energy Consumption Prediction. *Energy and Buildings, 264*, 112042.

[15] Luo, J., He, J., & Zhu, Q. (2023). Artificial Intelligence for Smart Grid Energy Storage Management: A Review of Predictive Analytics and Optimization Techniques. *Applied Energy, 336*, 120732.

[16] Reza, S. A., Hasan, M. S., Amjad, M. H. H., Islam, M. S., Rabbi, M. M. K., Hossain, A., ... & Jakir, T. (2025). Predicting Energy Consumption Patterns with Advanced Machine Learning Techniques for Sustainable Urban Development. *Journal of Computer Science and Technology Studies, 7*(1), 265–282.

[17] Shovon, M. S. S., Gomes, C. A., Reza, S. A., Bhowmik, P. K., Gomes, C. A. H., Jakir, T., ... & Hasan, M. S. (2025). Forecasting Renewable Energy Trends in the USA: An AI-Driven Analysis of Electricity Production by Source. *Journal of Ecohumanism, 4*(3), 322–345.

[18] Singh, H., Rao, P., & Mishra, A. (2023). *Anomaly detection in photovoltaic systems via deep learning autoencoder*. In 2023 IEEE 4th International Conference on Signal, Control and Communication (SCC) (pp. 1–6). IEEE.

[19] Brown, R., Anderson, P., & Clark, S. (2023). IoT sensors for rooftop solar monitoring. *Solar Energy*, 260, 112345.

[20] Liu, Y., Wang, H., & Li, J. (2022). *Deep learning–based fault diagnosis in photovoltaic systems*. Solar Energy, 230, 456–468

[21] Zhang, Y., Li, S., & Wang, J. (2023). Deep Learning-Based Short-Term Load Forecasting in Smart Grids: A Comparative Study of LSTM and GRU Models. *Energy Reports, 9*, 173–184.