International Journal of
Technology and Emerging Research
ISSN: 3068-1995 (Print) | 3068-109X (Online) | https://www.ioro.org/ijter

Hybrid Machine Learning Framework for SCADA-Based Anomaly Detection in Wind Turbine Systems Using Ensemble Learning

Zakir Ahmed Ansari[1*], Dr. Tariq Siddiqui[2]

[1] *M.Tech.Scholar, Department of Computer Science Engineering, Bhabha University Bhopal, India*

[2] *Associate Professor Department of Computer Science Engineering, Bhabha University Bhopal, India*

* **Corresponding author zakiransari952@gmail.com**

Abstract

Supervisory Control and Data Acquisition (SCADA) systems play a crucial role in monitoring the operational health of modern wind turbines by continuously collecting large volumes of sensor data. Efficient analysis of this data is essential for early fault detection, predictive maintenance, and reliable turbine operation. However, anomaly detection in SCADA environments remains challenging due to data imbalance, noise, nonlinear relationships, and dynamic operating conditions. Traditional machine learning approaches often suffer from limited generalization capability and may fail to achieve a balanced trade-off between precision and recall. To address these challenges, this paper proposes a Hybrid Machine Learning Framework for SCADA-based anomaly detection in wind turbine systems using ensemble learning. The proposed framework integrates Isolation Forest, One-Class Support Vector Machine (OCSVM), and Deep Autoencoder models to capture complementary anomaly characteristics from operational data. The outputs of these base models are further combined using an AdaBoost-based stacking architecture to improve classification robustness and anomaly detection performance. Experiments were conducted on a publicly available wind turbine SCADA dataset containing more than 50,000 operational samples and multiple turbine health parameters. The proposed hybrid framework was evaluated using Precision, Recall, F1-score, Area Under Curve (AUC), and confusion matrix analysis. Experimental results demonstrate that the proposed model significantly outperforms standalone approaches, achieving a Recall of 0.8702, F1-score of 0.7118, and AUC of 0.9678. Furthermore, the framework substantially reduces false negative predictions, making it highly suitable for predictive maintenance applications. The findings indicate that integrating machine learning, deep learning, and ensemble learning techniques provides a robust and effective solution for intelligent anomaly detection in industrial SCADA systems.

Keywords: SCADA, Anomaly Detection, Wind Turbine Monitoring, Machine Learning, Deep Learning, Ensemble Learning, Isolation Forest, One-Class SVM, Autoencoder, Predictive Maintenance

1. Introduction

The increasing adoption of intelligent monitoring systems and Industrial Internet of Things (IIoT) technologies has significantly enhanced the capability of industrial assets to generate large volumes of operational data. Supervisory Control and Data Acquisition (SCADA) systems are widely deployed in industrial environments to continuously collect and monitor sensor measurements, enabling real-time condition assessment and predictive maintenance. In wind turbine systems, SCADA platforms record operational parameters such as wind speed, active power generation, rotor behavior, temperature, vibration, and

environmental conditions, providing valuable information for fault diagnosis and performance analysis [1], [2].

Wind turbines operate under highly dynamic environmental conditions and are frequently exposed to varying wind speeds, temperature fluctuations, mechanical stress, and component degradation. These factors often result in abnormal operating conditions that may reduce energy production efficiency and increase maintenance costs. Therefore, early anomaly detection is essential for improving system reliability, minimizing downtime, and preventing catastrophic failures [3]. Traditional monitoring approaches based on threshold analysis and statistical methods are often inadequate for handling the complexity of large-scale SCADA datasets due to nonlinear relationships, noisy measurements, and continuously changing operational patterns [4].

Recent advances in machine learning (ML) and deep learning (DL) have enabled the development of intelligent anomaly detection systems capable of automatically identifying abnormal operational behavior from historical data. Algorithms such as Isolation Forest, Support Vector Machine (SVM), Random Forest, and Autoencoders have demonstrated promising performance in industrial fault diagnosis applications [5], [6]. Among these approaches, unsupervised and semi-supervised methods have gained considerable attention because fault-labeled datasets are often scarce in practical industrial environments. However, standalone anomaly detection models frequently suffer from performance limitations, including high false alarm rates, poor generalization capability, and an imbalance between precision and recall [7].

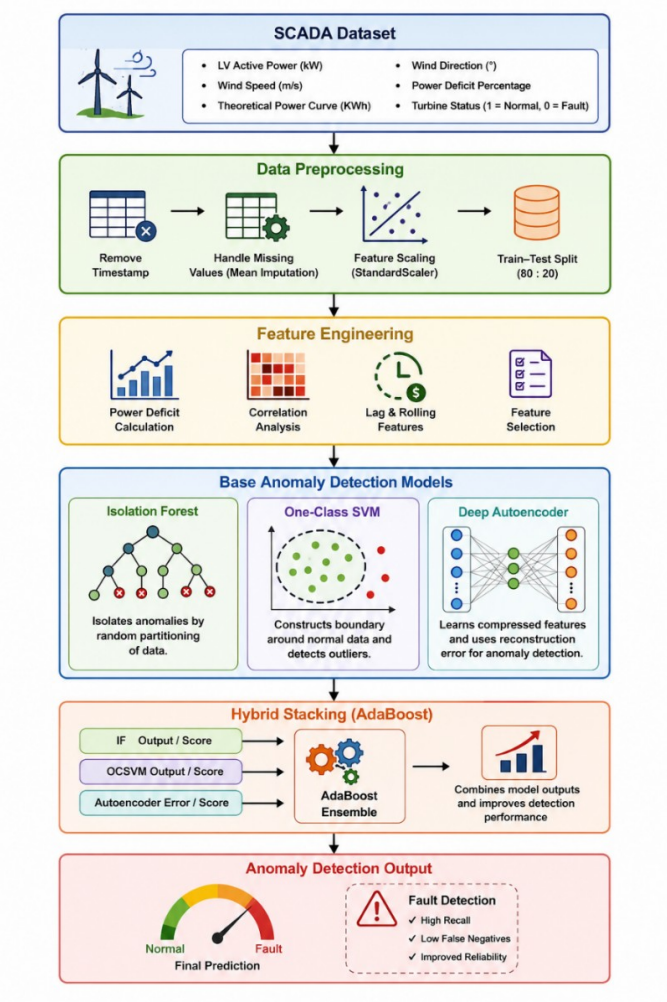

Figure.1 Proposed Hybrid Anomaly Detection Framework

To overcome these challenges, recent studies have explored hybrid and ensemble learning approaches that combine multiple anomaly detection techniques to exploit their complementary strengths. Ensemble frameworks can improve robustness, anomaly sensitivity, and classification accuracy by integrating statistical learning, boundary-based classification, and deep feature extraction mechanisms [8]. Nevertheless, achieving reliable anomaly detection performance in large-scale SCADA environments remains a challenging research problem, particularly when minimizing false negatives is critical for predictive maintenance applications.

Motivated by these challenges, this paper proposes a Hybrid Machine Learning Framework for SCADA-Based Anomaly Detection in Wind Turbine Systems Using Ensemble Learning. The proposed framework combines Isolation Forest, One-Class Support Vector Machine (OCSVM), and Deep Autoencoder models within an AdaBoost-based stacking architecture to improve anomaly detection performance. The framework is evaluated using a publicly available wind turbine SCADA dataset containing more than 50,000 operational samples. Experimental results demonstrate that the proposed hybrid model significantly improves anomaly detection capability, achieving an AUC of 0.9678, Recall of 0.8702, and F1-score of 0.7118 while substantially reducing false negative predictions compared to standalone approaches.

The major contributions of this paper are summarized as follows:

- Development of a hybrid anomaly detection framework integrating Isolation Forest, OCSVM, and Deep Autoencoder models.
- Implementation of an AdaBoost-based stacking architecture for robust anomaly classification.
- Comprehensive evaluation using multiple performance metrics including Precision, Recall, F1-score, AUC, and confusion matrix analysis.
- Demonstration of improved anomaly detection performance and reduced false negative rates for predictive maintenance applications.

2. RELATED WORK

The increasing availability of SCADA data has encouraged extensive research on intelligent anomaly detection and fault diagnosis techniques for wind turbine monitoring systems. Researchers have explored statistical methods, machine learning algorithms, deep learning models, and hybrid frameworks to improve predictive maintenance and operational reliability.

Early studies primarily relied on statistical process monitoring and threshold-based approaches for fault detection. Zaher et al. [9] proposed an automated SCADA data analysis framework for online wind turbine fault detection and demonstrated the effectiveness of statistical monitoring techniques. However, fixed-threshold methods often generate high false alarm rates and exhibit poor adaptability to changing operational conditions.

Despite these advancements, existing approaches still face challenges related to false negative predictions, model robustness, and anomaly sensitivity under dynamic operational conditions. Standalone models often exhibit performance trade-offs between precision and recall, resulting in either excessive false alarms or missed fault detections. Furthermore, limited research has focused on integrating Isolation Forest, One-Class SVM, and Deep Autoencoder models within a unified ensemble framework for SCADA-based wind turbine anomaly detection.

To address these limitations, this work proposes a Hybrid Machine Learning Framework that combines Isolation Forest, OCSVM, and Deep Autoencoder models using AdaBoost-based stacking. The proposed framework aims to improve anomaly detection performance by leveraging complementary anomaly characteristics captured by different learning paradigms.

Table I. Comparative Analysis of Existing Anomaly Detection Techniques

Reference	Method	Strengths	Limitations
Zaher et al. [9]	Statistical Monitoring	Simple implementation	High false alarms
Yang et al. [10]	SCADA Data Analytics	Effective condition monitoring	Limited anomaly sensitivity
Liu et al. [11]	Isolation Forest	Fast unsupervised detection	Moderate recall
SVM-based Methods [12]	One-Class SVM	Good boundary learning	Parameter sensitivity
Sakurada et al. [13]	Autoencoder	Nonlinear feature extraction	Reconstruction threshold dependency
Hybrid Methods [16], [17]	Ensemble Learning	Improved robustness	Increased complexity
Proposed Work	IF + OCSVM + Autoencoder + AdaBoost	High Recall and AUC	Higher computational cost

3. PROPOSED METHODOLOGY

This section presents the proposed Hybrid Machine Learning Framework developed for SCADA-based anomaly detection in wind turbine systems. The framework integrates Isolation Forest (IF), One-Class Support Vector Machine (OCSVM), and Deep Autoencoder models using an AdaBoost-based stacking architecture. The objective is to improve anomaly detection capability by combining statistical outlier detection, boundary-based classification, and deep feature learning within a unified ensemble framework.

Dataset Description

The experiments were conducted using a publicly available wind turbine SCADA dataset containing more than 50,000 operational samples collected from turbine monitoring systems. The dataset includes multiple operational parameters that describe turbine behavior under normal and faulty conditions.

Table II. Dataset Features

Feature	Description
LV Active Power (kW)	Actual electrical power generated
Wind Speed (m/s)	Wind velocity
Theoretical Power Curve (kWh)	Expected power generation
Wind Direction (°)	Wind orientation
Power Deficit Percentage	Difference between actual and theoretical power
Turbine Status	Target label (Normal/Fault)

The turbine status parameter is used as the output class where normal operating conditions are represented by class “1” and anomalous conditions are represented by class “0”.

Data Preprocessing

Raw SCADA data often contains missing values, redundant information, and features with different numerical scales. Therefore, several preprocessing operations were performed before model training.

The preprocessing steps include:

- Removal of irrelevant timestamp attributes.
- Mean imputation for handling missing values.
- Feature normalization using StandardScaler.
- Dataset partitioning using an 80:20 train-test split.

Standardization was applied according to:

Standardization Equation:

$$z = (x - \mu) / \sigma$$

where:
x = original feature value
μ = mean of the feature
σ = standard deviation of the feature

Proposed Hybrid Framework

The proposed framework consists of three base anomaly detection models:

1. Isolation Forest (IF)
2. One-Class Support Vector Machine (OCSVM)
3. Deep Autoencoder

Each model independently analyzes SCADA operational patterns and generates anomaly predictions. The outputs of these models are then combined through an AdaBoost-based stacking architecture to generate the final anomaly classification.

The hybrid architecture leverages the complementary strengths of multiple anomaly detection paradigms, improving classification robustness and reducing false negative predictions.

Isolation Forest

Isolation Forest is an unsupervised anomaly detection algorithm that isolates anomalous observations using recursive random partitioning. Anomalies generally require fewer partitions to become isolated compared to normal observations.

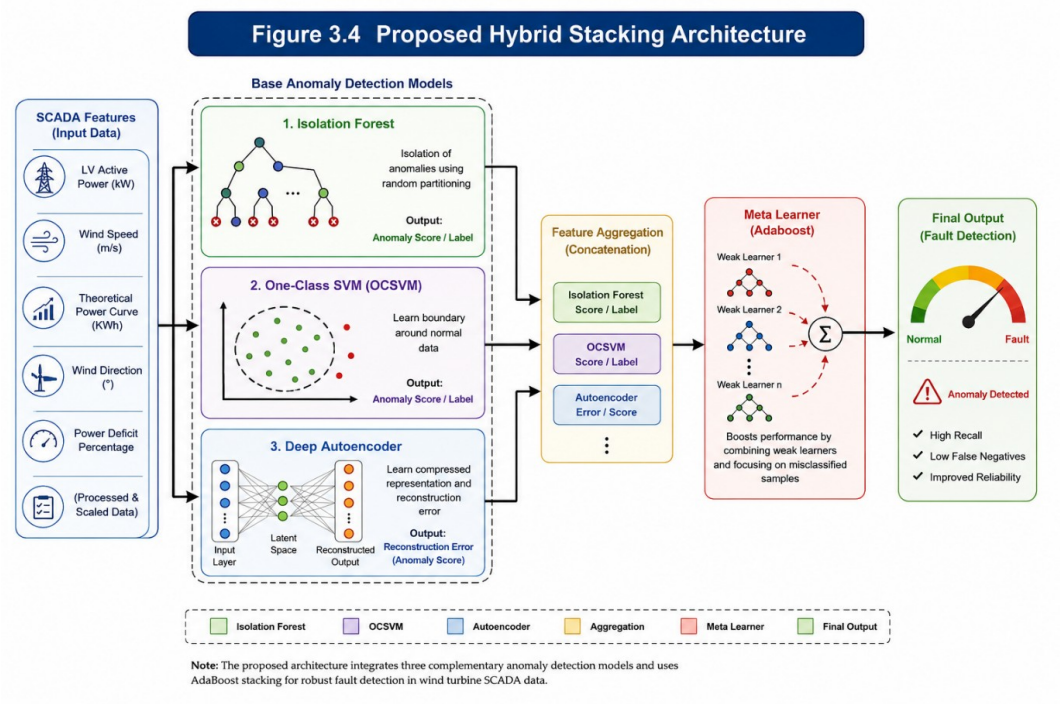

Figure 2. Proposed Hybrid Machine Learning Framework SCADA Dataset

The anomaly score is calculated as:

where:

where (E(h(x))) represents the average path length and (c(n)) denotes the normalization factor.

Observations with shorter path lengths receive higher anomaly scores and are classified as anomalous samples.

One-Class Support Vector Machine

One-Class Support Vector Machine constructs a decision boundary around normal operational data and classifies observations located outside the learned boundary as anomalies.

The optimization objective is defined as:

The One-Class Support Vector Machine (OCSVM) optimization problem can be written as:

subject to

where:

- = weight vector,
- = offset (decision threshold),
- = slack variables,
- = parameter controlling the fraction of outliers,
- = feature mapping function,
- = number of training samples.

OCSVM effectively identifies anomalies in high-dimensional feature spaces and is particularly useful when fault labels are limited.

Deep Autoencoder

The Deep Autoencoder is a neural network architecture that learns compressed latent representations of normal operational behavior through reconstruction learning. During training, the network minimizes reconstruction error between the input and reconstructed output.

The reconstruction loss using Mean Squared Error (MSE) is written as:

where:

- is the original input sample,
- is the reconstructed output sample,
- is the total number of samples,
- MSE quantifies the average squared reconstruction error between the original and reconstructed data.

Anomalous observations typically produce higher reconstruction errors and can therefore be identified using threshold-based classification.

G. AdaBoost-Based Stacking

The outputs generated by Isolation Forest, OCSVM, and Deep Autoencoder are aggregated and provided as input features to an AdaBoost classifier.

AdaBoost improves classification performance by iteratively assigning higher weights to previously misclassified samples. The final prediction is obtained by combining weak learners according to:

The equation can be written as:

where:

- is the final ensemble prediction,
- denotes the weak learner,
- represents the weight assigned to the weak learner,
- is the total number of weak learners in the ensemble.

The stacking mechanism enables the framework to exploit complementary anomaly detection information from all base models, resulting in improved classification accuracy and robustness.

H. Evaluation Metrics

The proposed framework is evaluated using commonly adopted classification metrics:

- Precision
- Recall
- F1-Score
- Area Under Curve (AUC)
- Confusion Matrix Analysis

These metrics provide a comprehensive assessment of anomaly detection performance and facilitate comparison with standalone baseline models.

4. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed Hybrid Machine Learning Framework for SCADA-based anomaly detection in wind turbine systems. The proposed model was compared with three baseline anomaly detection approaches, namely Isolation Forest, One-Class Support Vector Machine (OCSVM), and Deep Autoencoder. Performance was evaluated using Precision, Recall, F1-Score, Area Under Curve (AUC), confusion matrix analysis, and Receiver Operating Characteristic (ROC) curves.

Experimental Configuration

The experiments were performed using Python-based machine learning libraries including Scikit-learn, TensorFlow, NumPy, and Pandas. The SCADA dataset was preprocessed using mean imputation and feature standardization. The dataset was divided into training and testing subsets using an 80:20 split ratio. All models were trained using identical datasets to ensure a fair comparison.

The evaluation metrics used in this study include Precision, Recall, F1-Score, and AUC. Among these metrics, Recall is considered particularly important because it measures the ability of the model to correctly identify actual fault conditions and minimize missed anomalies.

Performance Comparison

The anomaly detection performance of all models was evaluated using Precision, Recall, F1-score, and Area Under Curve (AUC). The comparative performance results are presented in Table III and Figure 4.

Model	Precision	Recall	F1-Score	AUC
Isolation Forest	0.6274	0.5236	0.5708	0.8581
One-Class SVM	0.5220	0.4512	0.4840	0.8734
Deep Autoencoder	0.5752	0.4294	0.4917	0.8578
Hybrid Stacking (AdaBoost)	0.6022	0.8702	0.7118	0.9678

Table III. Performance Comparison of Anomaly Detection Models

The results demonstrate that the proposed Hybrid Stacking framework achieved the best overall performance among all evaluated models. Although Isolation Forest obtained slightly higher precision, the proposed framework significantly improved recall performance. The Hybrid model achieved a Recall value of 0.8702, substantially higher than Isolation Forest (0.5236), OCSVM (0.4512), and Deep Autoencoder (0.4294).

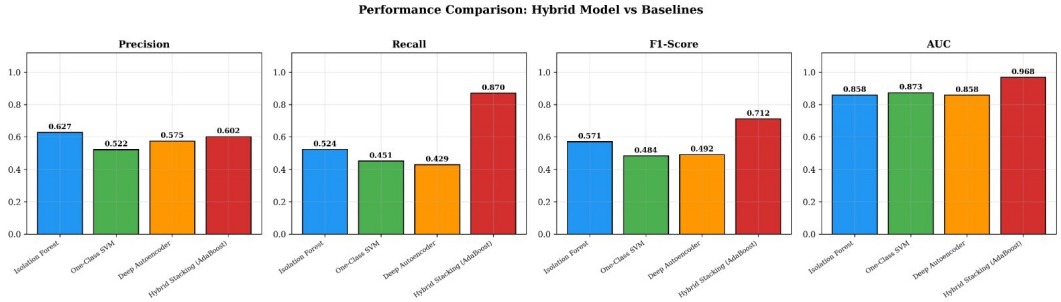

Figure 4 Performance Comparison of Baseline and Hybrid Models

The proposed framework also achieved the highest F1-Score of 0.7118 and AUC of 0.9678, indicating excellent anomaly discrimination capability and balanced classification performance.

Confusion Matrix Analysis

Confusion matrix analysis was performed to evaluate the classification behavior of each anomaly detection model. Figure 5 illustrates the confusion matrices obtained from all evaluated approaches.

The proposed Hybrid Stacking framework correctly identified 3666 anomalous observations while reducing false negatives to only 547 samples. In comparison, Isolation Forest, OCSVM, and Deep Autoencoder produced 2007, 2312, and 2404 false negatives, respectively.

The substantial reduction in false negatives demonstrates the ability of the proposed framework to detect abnormal operational conditions more effectively. This characteristic is particularly important in predictive maintenance systems because undetected faults may lead to severe operational failures and increased maintenance costs.

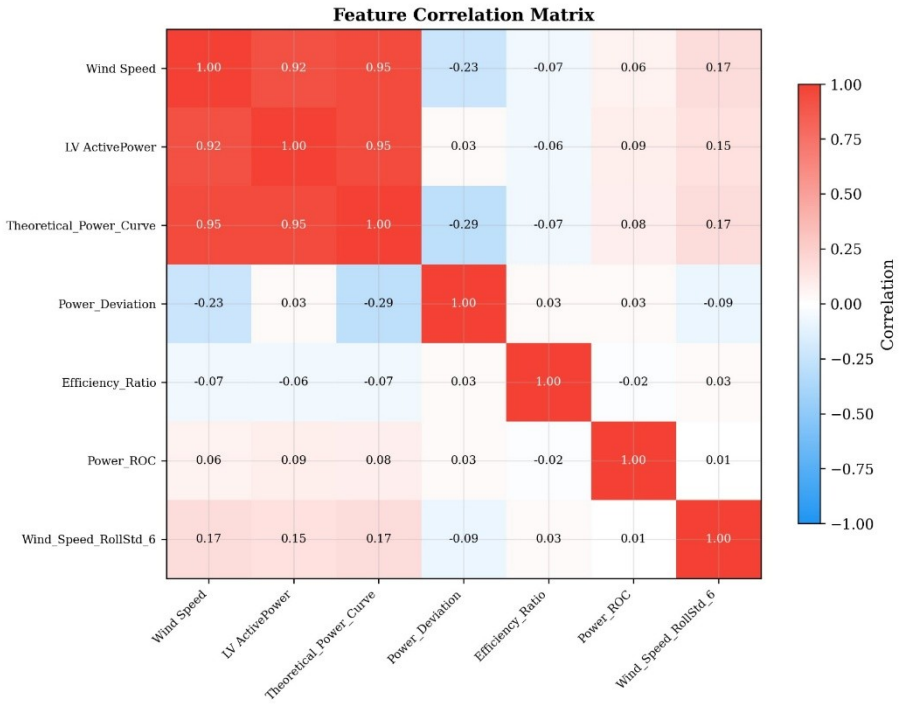

Figure 5 Feature Correlation Matrix

ROC Curve Analysis

Receiver Operating Characteristic (ROC) curves were used to evaluate the discrimination capability of all anomaly detection models under varying classification thresholds. Figure 4 presents the ROC curves obtained from the experiments.

The proposed Hybrid Stacking framework achieved the highest AUC value of 0.9678, significantly outperforming Isolation Forest (0.8581), OCSVM (0.8734), and Deep Autoencoder (0.8578). The ROC curve of the proposed model remains consistently closer to the upper-left corner, indicating improved sensitivity and specificity.

These results confirm that integrating multiple anomaly detection paradigms through ensemble learning improves overall classification reliability and anomaly discrimination performance.

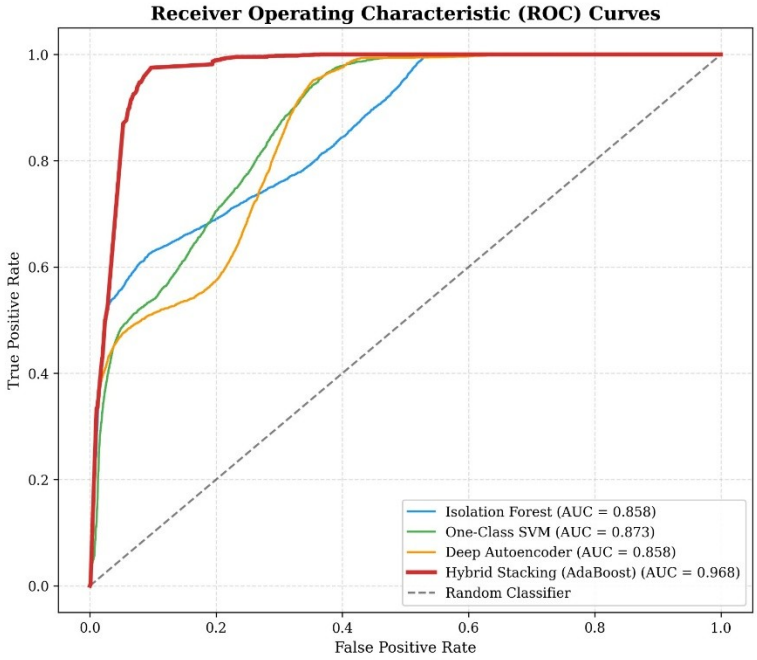

Figure 4 ROC Curve Comparison of Anomaly Detection Models

Autoencoder Reconstruction Error Analysis

The Deep Autoencoder model was evaluated using reconstruction error analysis. Figure 5 shows the distribution of reconstruction errors generated during testing.

Normal operational samples were concentrated around lower reconstruction error values, while anomalous observations exhibited significantly higher reconstruction errors. A reconstruction threshold was applied to distinguish normal and abnormal operational behavior.

The results indicate that the Autoencoder successfully learned latent representations of normal turbine behavior and effectively separated anomalous observations through reconstruction-based anomaly scoring.

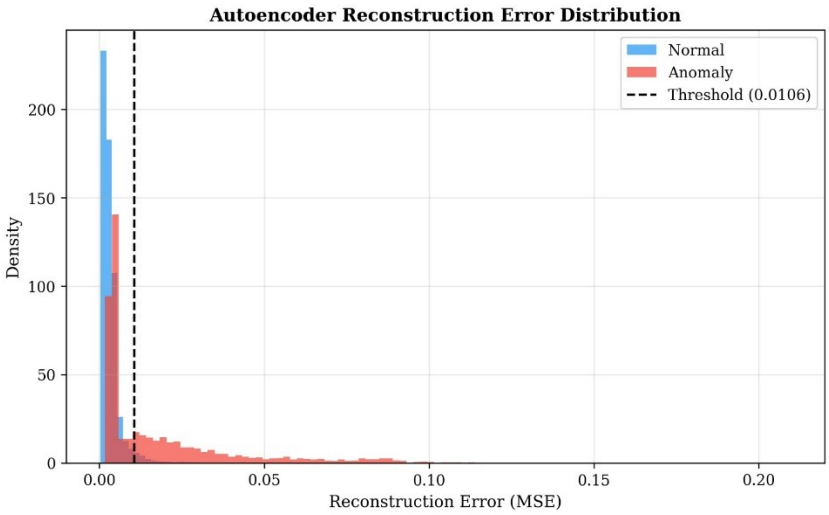

Figure 5 Autoencoder Reconstruction Error Distribution

a. Time-Series and Power Curve-Based Anomaly Visualization

To visually analyze anomaly behavior under real operational conditions, time-series anomaly detection analysis and power curve visualization were performed using the proposed hybrid framework.

Figure 6 illustrates anomaly detection behavior over time using Active Power, Wind Speed, and Hybrid Anomaly Score distributions. The anomaly score exceeds the decision threshold during abnormal operational intervals, indicating successful identification of anomalous turbine behavior. The hybrid framework effectively captures temporal variations and operational disturbances occurring under changing environmental conditions.

Similarly, Figure 7 presents the wind turbine power curve along with anomalies detected by the proposed hybrid framework. Under normal operating conditions, turbine power generation follows the theoretical power curve closely. However, anomalous operational samples deviate significantly from the expected power curve behavior. The proposed hybrid framework successfully identifies these abnormal operational deviations, demonstrating strong practical applicability in wind turbine monitoring systems.

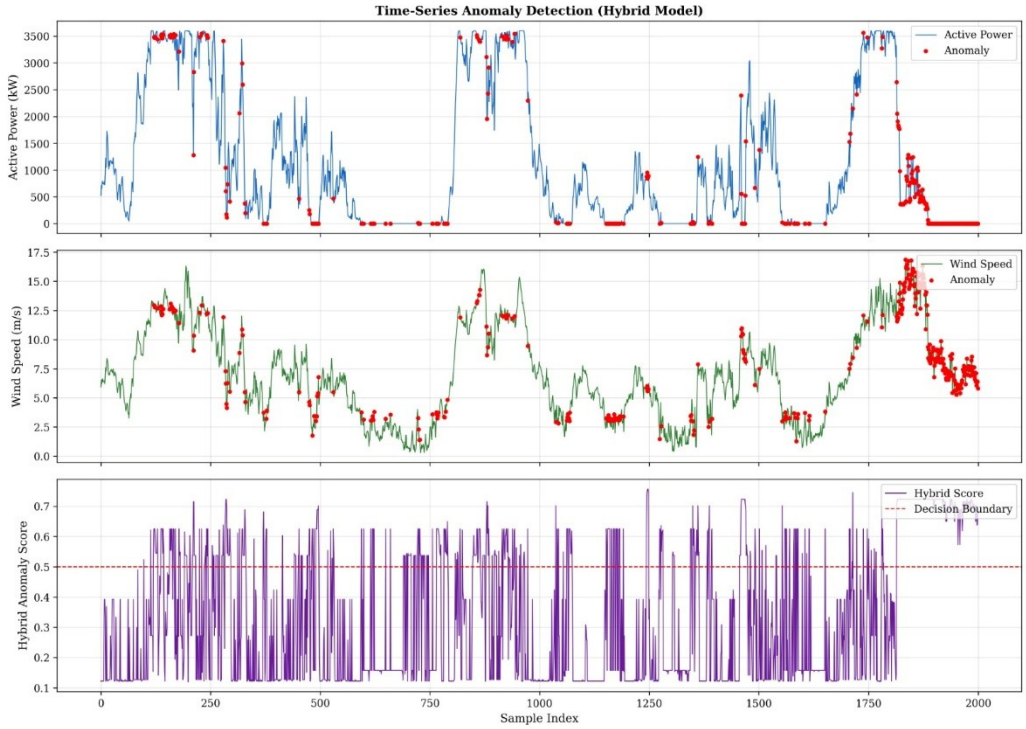

Figure 6 Time-Series Anomaly Detection using Hybrid Framework

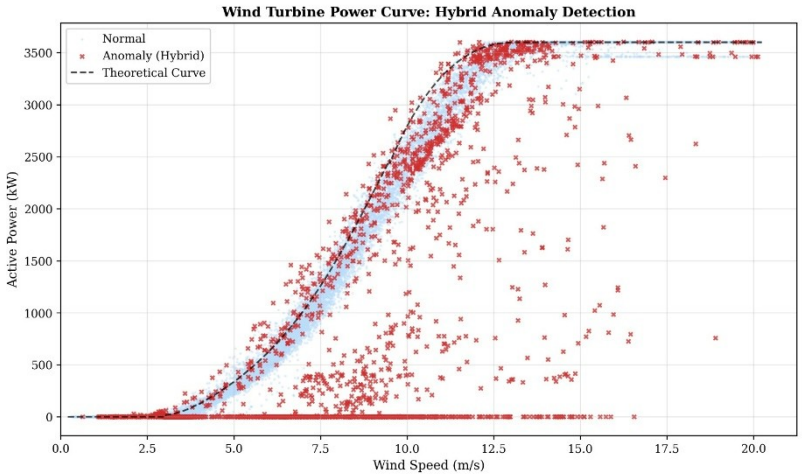

Figure 7 Wind Turbine Power Curve with Hybrid Anomaly Detection

5. Discussion

The experimental results demonstrate that standalone anomaly detection algorithms exhibit moderate performance but suffer from limitations related to anomaly sensitivity and fault detection capability. Isolation Forest achieved balanced performance, whereas OCSVM and Autoencoder demonstrated relatively lower recall values.

The proposed Hybrid Stacking framework successfully addressed these limitations by combining statistical outlier detection, boundary-based learning, and deep feature extraction techniques. The integration of complementary anomaly information enabled the model to achieve superior Recall, F1-Score, and AUC performance.

The most significant improvement achieved by the proposed framework is the reduction of false negative predictions. Since missed anomalies can lead to unexpected turbine failures and increased maintenance costs, the proposed framework provides a practical and reliable solution for predictive maintenance applications.

Overall, the results confirm that ensemble learning combined with machine learning and deep learning techniques provides an effective strategy for intelligent anomaly detection in industrial SCADA systems.

6. Conclusion

This paper presented a Hybrid Machine Learning Framework for SCADA-based anomaly detection in wind turbine systems using ensemble learning. The proposed framework integrates Isolation Forest, One-Class Support Vector Machine (OCSVM), and Deep Autoencoder models within an AdaBoost-based stacking architecture to improve anomaly detection capability and predictive maintenance performance. The framework was designed to address the limitations of standalone anomaly detection approaches, particularly in handling complex nonlinear operational patterns and reducing false negative predictions in large-scale SCADA datasets. Experiments were conducted using a publicly available wind turbine SCADA dataset containing more than 50,000 operational samples. The proposed model was evaluated using Precision, Recall, F1-Score, AUC, confusion matrix analysis, and ROC curve analysis. Comparative evaluation against Isolation Forest, OCSVM, and Deep Autoencoder demonstrated the effectiveness of the proposed hybrid framework. The experimental results showed that the proposed Hybrid Stacking model achieved the highest overall performance, with a Recall of 0.8702, F1-Score of 0.7118, and AUC of 0.9678. Furthermore, the framework significantly reduced false negative predictions compared to standalone models, improving anomaly sensitivity and fault detection reliability. The integration of statistical outlier detection, boundary-based classification, and deep feature learning enabled the framework to capture complementary anomaly characteristics and achieve robust classification performance. The findings indicate that ensemble learning combined with machine learning and deep learning techniques provides an effective solution for intelligent anomaly detection in industrial SCADA environments. The proposed framework can support predictive maintenance strategies by enabling early fault detection, reducing operational downtime, and improving the overall reliability of wind turbine systems.

Future work may focus on real-time deployment of the framework in industrial monitoring environments, integration with Industrial Internet of Things (IIoT) platforms, and the incorporation of Explainable Artificial Intelligence (XAI) techniques to improve model interpretability and decision transparency.

Funding

This research received no external funding

Conflict of Interest

The authors declare no conflict of interest

References

[1] W. Yang, R. Court, and J. Jiang, “Wind turbine condition monitoring by the approach of SCADA data analysis,” Renewable Energy, vol. 53, pp. 365–376, 2013.

[2] Ansari, A.A.; Dyanamina, G. Fault Ride-Through Operation Analysis of Doubly Fed Induction Generator-Based Wind Energy Conversion Systems: A Comparative Review. Energies 2022, 15, 8026. https://doi.org/10.3390/en15218026

[3] A. A. Ansari, G. Dyanamina and A. A. Ansari, "Fuzzy Logic-Driven Protection Coordination for Robust DFIG System Operation Under Symmetrical Fault Conditions," 2024 IEEE 1st International Conference on Green Industrial Electronics and Sustainable Technologies (GIEST), Imphal, India, 2024, pp. 1-7, doi: 10.1109/GIEST62955.2024.10960182.

[4] W. Qiao and D. Lu, “A survey on wind turbine condition monitoring and fault diagnosis,” *IEEE Transactions on Industrial Electronics*, vol. 62, no. 10, pp. 6536–6545, 2015.

[5] A. Zaher, S. McArthur, D. Infield, and Y. Patel, “Online wind turbine fault detection through automated SCADA data analysis,” *Wind Energy*, vol. 12, no. 6, pp. 574–593, 2009.

[6] Ansari, Aftab Ahmed, and Giribabu Dyanmina. "Real-Time Implementation of The Fuzzy Logic Controlled Parallel Protection Technique to Enhance the DFIG System's FRT Capability." Scientia Iranica (2023).

[7] H. Sakurada and T. Yairi, “Anomaly detection using autoencoders with nonlinear dimensionality reduction,” in *Proc. MLSDA Workshop*, 2014, pp. 4–11.

[8] Z. Zhao, L. Bin, and X. Wang, “Machine learning methods for wind turbine condition monitoring and fault diagnosis: A review,” *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 2659–2673, 2018.

[9] T. G. Dietterich, “Ensemble methods in machine learning,” in *Multiple Classifier Systems*, Springer, 2000, pp. 1–15.

[10] A. Zaher, S. McArthur, D. Infield, and Y. Patel, “Online wind turbine fault detection through automated SCADA data analysis,” *Wind Energy*, vol. 12, no. 6, pp. 574–593, 2009.

[11] W. Yang, P. Tavner, C. Crabtree, Y. Feng, and Y. Qiu, “Wind turbine condition monitoring: Technical and commercial challenges,” *Wind Energy*, vol. 17, no. 5, pp. 673–693, 2014.

[12] F. T. Liu, K. M. Ting, and Z. H. Zhou, “Isolation Forest,” in *Proc. IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.

[13] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[14] M. Sakurada and T. Yairi, “Anomaly detection using autoencoders with nonlinear dimensionality reduction,” in *Proc. MLSDA Workshop*, 2014, pp. 4–11.

[15] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] T. G. Dietterich, “Ensemble methods in machine learning,” in *Multiple Classifier Systems*, Springer, 2000, pp. 1–15.

[17] S. Simani and P. Castaldi, “Data-driven fault diagnosis and sustainable control of wind turbines,” *Sustainable Energy, Grids and Networks*, vol. 9, pp. 85–101, 2017.

[18] Z. Zhao, X. Wang, and L. Bin, “Hybrid machine learning approaches for wind turbine fault diagnosis: A review,” *Renewable Energy*, vol. 145, pp. 1355–1372, 2020.