LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374 E-ISSN:1855-363X
VOL. 24, NO. 4 (2026)

COMPARATIVE ANALYSIS OF SUPERVISED AND UNSUPERVISED MACHINE LEARNING FOR RELIABLE WIND TURBINE MONITORING AND SUSTAINABLE ENERGY GENERATION – ENERGY POLICY

Manimegalai Somasundaram*, Dr. K.S.Gayathri

* Sri Sivasubramaniya Nadar College of Engineering, Chennai, 603 110, Tamilnadu, India.
Sri Sivasubramaniya Nadar College of Engineering, Chennai, 603110, Tamilnadu, India.

Correspondence: Manimegalai Somasundaram Email: megalaieee.s@gmail.com

Abstract
Condition monitoring plays a role in wind turbine maintenance. But supervised machine learning is constrained in the presence of unlabeled data in datasets related to Supervisory Control and Data Acquisition (SCADA). This paper suggests method to detect anomalies in wind turbines based on 8,760 observations per hour from the Texas Wind Turbine SCADA dataset. Power estimation is performed using a 300-tree model of Random Forest regression that uses wind speed, wind direction, air pressure, and air temperature. The power difference between the measured and expected powers is used to label the anomaly using a threshold of 4.1767 kW. The anomalous observations are identified as 700 and normal observations are 8,060. The optimized Random Forest and XGBoost models have a 93.06% and 93.24% accuracy, respectively, and their F1 scores are 41.54% and 43.51%, respectively. LOF finds 438 density anomalies, and McNemar's test shows that there is no significant difference between classifiers.

Keywords: Wind turbine; SCADA; Random Forest regression; Anomaly detection; Predictive maintenance.

1. Introduction

Wind power generation systems have been deployed at a much higher rate all over the world, due to the rising demand for clean and sustainable energy. Wind energy is one of the fastest-growing renewable energy sources due to its environmental benefits and capacity to decrease reliance on fossil fuels. But random breakdowns of wind turbines can decrease energy generation, maintenance expenses, and component life. Hence, condition monitoring and early fault detection is vital to safe and efficient wind turbine operation [1],[21]. Supervisory Control and Data Acquisition (SCADA) systems are used to gather operational and environmental data, such as wind speed, wind direction, atmospheric pressure, air temperature and generation power, at all times. These are useful measurements for operating condition monitoring and diagnosis of abnormal operating conditions of the turbine. The cost of the SCADA-based monitoring is less than that of the conventional vibration-based monitoring techniques, as the sensors that are used are already present in the commercial wind turbines. Thus, the SCADA data is a valuable resource in the development of intelligent fault diagnosis and predictive maintenance systems [9], [20], [23].

The recent developments in machine learning have made it possible to enhance condition monitoring using SCADA. The wind power prediction, anomaly detection, and fault diagnosis have been effectively implemented by using Random Forest, Support Vector Machine, XGBoost, ensemble learning, and deep learning techniques [1, 3–8, 12, 13]. Most of these approaches fall short of performing well enough, and most require manually labeled fault datasets or complex deep-learning architectures that cannot be relied upon in real-world industrial applications. In order to address these challenges, in this paper, an innovative machine learning framework for automatic wind turbine anomaly detection for the Texas Wind Turbine SCADA dataset is proposed. The proposed framework starts by pre-processing the SCADA data using missing value analysis, duplicate removal, feature standardization and exploratory data analysis. A Random

Forest regression model is then applied to estimate the expected turbine power and then an adaptive statistical threshold is applied to automatically label the anomalies without manual annotation. The labels created are then used to train optimized Random Forest and XGBoost classifiers. In addition, the robustness and statistical significance of the proposed framework are assessed with sensor ablation analysis, Local Outlier Factor (LOF) screening, and McNemar statistical testing.
The major contributions of this work are summarized as follows:

- ✓ A novel Automatic Anomaly-label Generation Method using the Expected power estimation provided by Random Forest and Adaptive statistical thresholding is proposed.
- ✓ A complete machine learning framework combining Random Forest regression, Random Forest classification as well as XGBoost classification is developed for wind turbine anomaly detection.
- ✓ The use of sensor ablation analysis is performed to assess the contribution of individual SCADA variables.
- ✓ Local Outlier Factor is used as an additional unsupervised screening technique and McNemar's statistical test is used to determine whether the differences in predictions between the optimized Random Forest and XGBoost classifiers are statistically significant.

framework is tested with the dataset Texas Wind Turbine SCADA, which comprises 8,760 operational observations, and the rest of this paper is structured as follows. A review of the literature pertaining to the issue is provided in Section 2. The proposed methodology is discussed in Section 3. The experimental results and comparative analysis are discussed in section 4. Lastly, Section 5 wraps up the paper and presents directions for further research.

2. Literature Survey

The condition of wind turbines is now the subject of much research, as their output of renewable energy grows and cost-effective and reliable maintenance strategies become increasingly important. Data-driven approaches can be used to leverage operational data continuously gathered by a Supervisory Control and Data Acquisition (SCADA) system for fault diagnosis, anomaly detection and predictive maintenance. In recent years, the power of machine learning techniques, such as Random Forest, Support Vector Machines, ensemble learning, and deep learning has been proven to model the nonlinear relationship between the environmental variables and the turbine operational conditions. Random Forest has been successfully applied to forecasting wind power, completing datasets and selecting features, and SCADA based forecasting methods have been used to enhance performance analysis and operational reliability of turbines [1, 2, 8, 9]. There are a few researchers who have suggested enhanced machine learning and deep learning models for wind turbine fault diagnosis. For anomaly detection, Zhang et al. [3] proposed stacked denoising autoencoders with XGBoost while Khan and Byun proposed multimodal machine learning, stacking classifiers, PCA-KMeans, and genetic algorithm-based ensemble learning techniques for multi-fault classification using SCADA data [4], [7], [12], [13]. To enhance the interpretability of ensemble learning models using Random Forest and XGBoost classifiers, Explainable Artificial Intelligence (XAI) has been introduced [5]. These techniques are all successful in terms of their classification accuracy, but most of them need labeled datasets and demanding learning models.
Unsupervised and hybrid machine learning techniques have also been studied to implement anomaly detection and predictive maintenance for SCADA. Some of the methods consistently adopted for detection of abnormal operating conditions from SCADA measurements, such as Isolation Forest [10], clustering algorithms [18] and predictive maintenance models based on machine learning [20] and [23] have shown good performance. Besides, vibration analysis, acoustic emission, autoencoders and deep residual neural network have been used to diagnose gearbox, bearing, blade and driveline faults [15]–[17], [19], [24]–[26]. In general these methods are more effective at detecting faults but have the disadvantage of either needing extra sensing devices, complex feature extraction methods, or a high amount of computing power.
Although there has been much progress, there are still a number of challenges to practical wind turbine condition monitoring. There are many existing studies, which assume they have a dataset of manually labelled faults available; however, real-world SCADA data are mostly unlabeled. In addition, few studies

have explored automatic anomaly-label generation via the use of expected-power estimation, systematic sensor ablation analysis, complementary unsupervised screening, and statistical significance testing of the output of classifiers. In response to these gaps, this paper proposes a framework that combines the following algorithms: Random Forest and XGBoost classification, Local Outlier Factor screening, sensor ablation analysis, and McNemar statistical testing to handle an extended machine-learning benchmark and estimate the expected power based on Random Forest regression, while adapting the deviation-based anomaly labelling and optimizing it for the classification algorithms.

3. Proposed Methodology

3.1 Overall Framework

The proposed framework is an intelligent machine learning based framework to automatically detect wind turbine anomaly from the Texas Wind Turbine SCADA data set. The framework aims to detect abnormal operating conditions based on no manually labeled fault data. The methodology involves six major steps which are shown in Figure 1: Data acquisition and preprocessing, expected power estimation using Random Forest regression, automatic anomaly-label generation, supervised classification, performance evaluation, and model validation.

To enhance data quality, the Texas SCADA dataset is preprocessed using missing value analysis, duplicate removal, feature selection, feature standardization and exploratory data analysis, resulting in 8,760 hourly observations. Next, a Random Forest regression model is trained to predict the power of the turbine, given the wind velocity, wind direction, atmospheric pressure and air temperature as input. The difference in power measured (P_i) and computed ($\hat{P}_i$) is calculated as

$$D_i = |P_i - \hat{P}_i| \quad (1)$$

where D_i denotes the absolute power deviation for the ith SCADA observation. An adaptive statistical threshold is then determined using

$$T = \mu_D + \sigma_D \quad (2)$$

The above formula is valid for the threshold T, where μ_D represents the mean deviation and σ_D represents the standard deviation of the deviation values. The threshold is used to automatically classify the observations from the SCADA system as either normal or anomalous, without manual fault annotation. The labeled data set is then used to train anomaly detection classifiers, such as Random Forest and XGBoost. Both models are validated based on the following performance metrics: accuracy, precision, recall, f1 score, roc-auc and confusion matrix analysis. Furthermore, sensor ablation analysis is conducted to delve into the impact of individual SCADA variables and the robustness of the proposed framework is evaluated by Local Outlier Factor (LOF), statistical test is used to validate the statistical significance of classifier predictions.

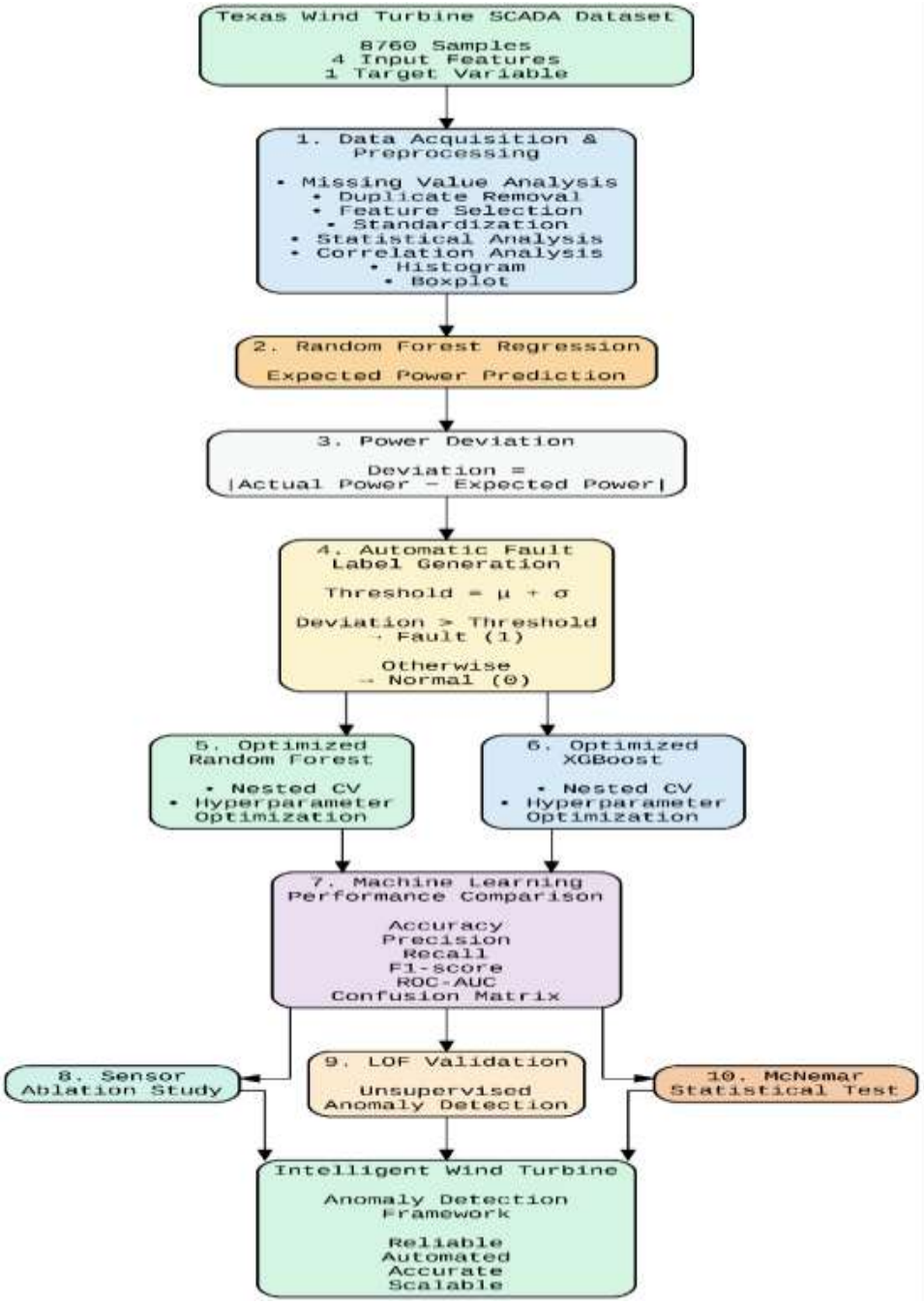


Figure 1. Overall architecture of the proposed machine learning framework

3.2 Texas SCADA Dataset

The anomaly detection framework proposed uses the Texas Wind Turbine Supervisory Control and Data Acquisition (SCADA) dataset to assess the performance of the machine learning models used. The dataset includes 8,760 hourly observations, which cover a year of wind turbine operation with different environmental and operating conditions. The SCADA system continuously collects operational measurements from wind turbines, which is a valuable information for condition monitoring, fault diagnosis and predictive maintenance. There are six variables in the dataset: Time Stamp, Wind Speed (m/s), Wind Direction (deg), Atmospheric Pressure (atm), Air Temperature (°C), Generated Power (kW). From these

variables, the input features consist of wind speed, wind direction, atmospheric pressure and air temperature and the output feature is the generated wind power considered the target variable for the expected power estimation through the Random Forest regression model. The timestamp attribute will only be stored for reference and will not be used in model training.

Table 1. Description of the Texas Wind Turbine SCADA Dataset

Variable	Unit	Description	Role
Time Stamp	–	Date and time of SCADA measurement	Reference
Wind Speed	m/s	Average wind velocity	Input
Wind Direction	Degree	Wind flow direction	Input
Atmospheric Pressure	atm	Ambient pressure	Input
Air Temperature	°C	Ambient temperature	Input
Power Output	kW	Generated electrical power	Target

The statistics of selected SCADA variables are summarized in Table 2. A total of 8,760 observations are used, with a mean power output of 964.47 kW, a minimum of 0 kW and a maximum of 3004.01 kW. Also the average wind velocity is 7.38 m/s and ranges from 0.119 m/s to 19.743 m/s, which corresponds to various operating conditions. The wind direction covers the entire 0° – 360° direction, and atmospheric pressure and air temperature are normal seasonal changes during the period of observation.

Table 2. Statistical Summary of the Texas SCADA Dataset

Variable	Mean	Std	Min	Max
Power (kW)	964.47	878.56	0.00	3004.01
Wind Speed (m/s)	7.38	3.14	0.119	19.743
Wind Direction (°)	146.62	84.68	0	360
Pressure (atm)	0.9924	0.0053	0.9745	1.0145
Air Temperature (°C)	22.15	4.86	3.26	32.96

The data set was reviewed before modeling was done, to assess the completeness and quality of the data set. Analysis of the dataset further revealed that there aren't any missing values and there are no duplicate records, which means that the acquired SCADA measurements are complete and fit to be used for machine learning applications. The selected input variables are adequate to identify the environment of which the wind turbine power generation is a part and to model the nonlinear relationship between the operating conditions and the power generation. Finally, the cleaned data is used for (a) data pre-processing, (b) expected power estimation, (c) anomaly-label generation and (d) supervised classification of anomalies.

3.3 Data Preprocessing

The quality of the input data significantly influences the performance of machine learning models. Therefore, the Texas Wind Turbine SCADA dataset was preprocessed prior to model development to ensure data consistency and improve prediction reliability. The preprocessing stage consisted of dataset inspection, data quality verification, feature selection, feature standardization, and exploratory data analysis, as illustrated in Figure 2.

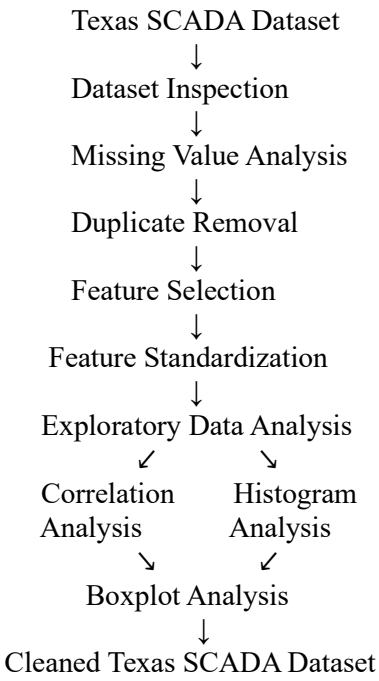


Figure 2. Data preprocessing workflow

The first step was to load the SCADA data into the Python environment using the Pandas library. The SCADA dataset was loaded into the Python environment using the Pandas library as the first step. The data set included 8,760 data points of six variables: Time Stamp, Power (kW), Wind Speed (m/s), Wind Direction (deg), Atmospheric Pressure (atm), and Air Temperature (°C). The dataset was initially evaluated for dimensions, types of attributes and descriptive statistics. Analysis showed that all variables were properly structured for subsequent machine learning processing.

Analysis of missing values and duplicate entries (DVQ) were then used to verify data quality. No missing values and duplicate observations were found in the data set, which suggests complete and consistent SCADA measurements. Therefore, no imputation of data or removal of duplicates was needed. After the data quality assessment, four environmental variables were chosen as the predictor variables and generated power as the target variable: wind speed, wind direction, atmospheric pressure and air temperature. The Time Stamp attribute was not used to train the model since it is not directly used in power prediction. Because the numerical range of variables is different, feature standardization was used to make the distribution of each feature zero-mean and unit-variance with the help of the StandardScaler algorithm. The standardized feature is calculated to be

$$x_i' = (x_i - \mu) / \sigma \qquad (3)$$

where x_i is the original feature value, μ is the mean of the features, σ is the standard deviation of the features, and x_i' is the standardized feature value. Feature scaling makes the numbers more stable and helps prevent large features from overwhelming the learning process. Exploratory data analysis (EDA) was then performed to explore the statistical nature of the selected variables of the SCADA. A Pearson correlation matrix was created to analyse the correlation among the environmental variable and the produced power.The normalized dataset with the four input variables and the target output variable (power output) was saved as Cleaned_Texas_SCADA.csv and was used as an input to the Random Forest regression model that was described in the next section.

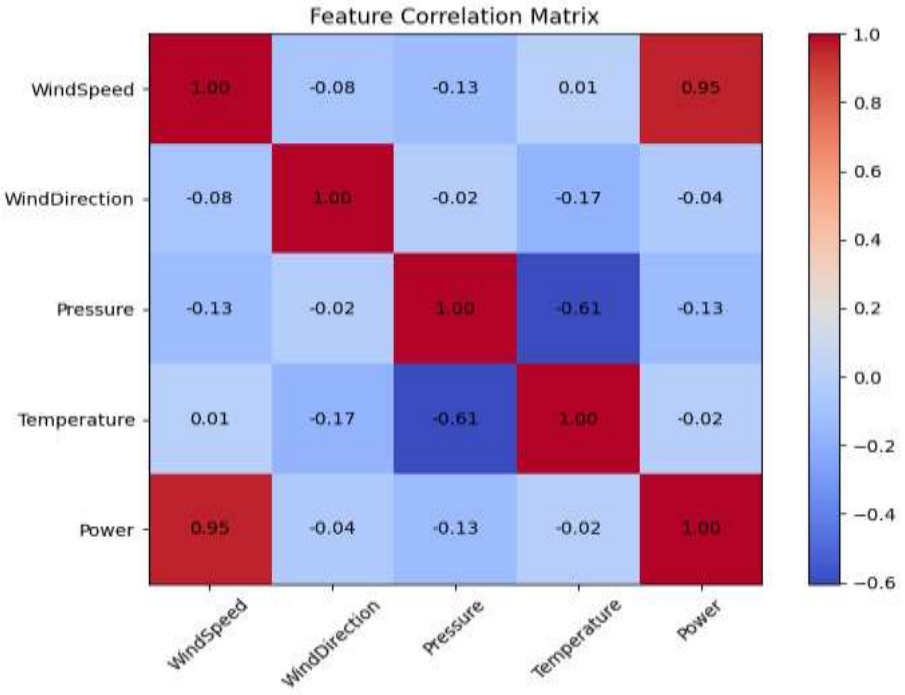


Figure 3. Pearson correlation matrix of the selected SCADA variables.

3.4 Random Forest Regression

Data preprocessing was followed by the development of a Random Forest Regression (RFR) model to estimate the wind turbine power output when running in normal operating conditions. The four standardized SCADA variables (wind speed, wind direction, etc.) were used as input features to the regression model and the generated power was used as the target variable. A random state of 42 was used to split the cleaned dataset (8,760 observations) into 75% training and 25% testing data subsets. The Random Forest regressor was set to 300 decision trees (n_estimators = 300) and a depth of 15 (max_depth = 15) and parallel processing (n_jobs = –1) to enhance computing speed. The mean of all decision trees in the ensemble to estimate the expected power output is given by

$$\hat{P} = (1/N) \sum_{i=1}^{N} T_i(X) \quad (4)$$

where $\hat{P}$ represents the predicted power output, N is the total number of decision trees, and $T_i(X)$ denotes the prediction of the ith decision tree. Table 3 summarizes the Random Forest regression configuration. The estimated power obtained from the regression model is subsequently utilized to calculate the power deviation for automatic anomaly-label generation, as described in the next subsection.

Table 3. Random Forest Regression Configuration and Performance

Parameter	Value
Dataset	Texas SCADA
Total Samples	8,760
Input Features	Wind Speed, Wind Direction, Pressure, Temperature
Target Variable	Power
Train-Test Split	75% : 25%
Number of Trees	300
Maximum Tree Depth	15

Parameter	Value
Random State	42

3.5 Automatic Anomaly Label Generation

The target classes for the supervised machine learning were manually defined, but since the Texas Wind Turbine SCADA dataset does not have these labels, an automatic anomaly-label generation strategy was developed to build such target classes. The method proposed here takes the power that is expected by the Random Forest regression model, as a benchmark to identify unusually large power deviations. The distance from the curve measured to the predicted curve is the distance between them and is called the absolute deviation between measured and predicted power.

$$D_i = |P_i - \hat{P}_i| \quad (5)$$

where D_i represents the power deviation, P_i is the measured power received from the SCADA system and $\hat{P}_i$ is the expected power predicted by the Random Forest regression model.The anomalies were statistically analyzed to establish an adaptive threshold for detecting anomaly. In order to avoid using a fixed threshold the mean deviation and standard deviation were used to determine the threshold as

$$T = \mu_D + \sigma_D \quad (6)$$

T is the Adaptive Threshold, μ_D is the mean deviation and σ_D is the standard deviation of the deviation values. Based on the regression output, the mean deviation was 1.5010 kW and the standard deviation was 2.6756 kW which gives an adaptive threshold of 4.1767 kW. Each SCADA observation was automatically classified as either a binary class label as per the following rules:

$$L_i = \{ 1, \text{if } D_i > T; 0, \text{if } D_i \leq T \} \quad (7)$$

Where 1 is an anomalous operating condition and 0 is used for the normal turbine operating condition.
The proposed threshold-based labelling approach applied to the 8760 SCADA observations yielded an anomaly percentage of 7.99% with 700 samples classified as anomalous and 8060 samples classified as normal. Labels were generated and saved along with the expected power and deviation values in Texas_Labelled_SCADA.csv and then used as the target variable for supervised classification. As these labels are based on model residuals instead of confirmed records related to maintenance or components failure, for the purposes of this study they are considered as deviation-based anomaly labels.

Table 4. Statistics of automatic anomaly label generation

Parameter	Value
Total SCADA observations	8,760
Mean deviation (kW)	1.5010
Standard deviation (kW)	2.6756
Adaptive threshold (kW)	4.1767
Normal samples	8,060
anomaly samples	700
Anomaly percentage	7.99%

3.6 Random Forest Classification

The automatically labelled data was used to train a wind turbine anomaly detection Random Forest (RF) classifier. A total of 8760 observations were used in the dataset, of which 8060 were normal and 700 were anomalous, recorded by the SCADA data. Four basic meteorological variables (wind speed, wind direction, atmospheric pressure, air temperature) were selected as input features, and the automatically generated anomaly label was selected as the target variable.

The training and test sets of labelled data were created using stratified sampling to ensure a similar class distribution for both sets. The following best Random Forest hyperparameters were reported by the analysis: 100 Trees, no maximum tree grown depth, minimum samples split = 10, minimum samples per leaf = 1, and logarithmic feature selection (max_features = log2). The reported nested cross-validation F1-scores were 0.4865, 0.3868, 0.4080, 0.4251, and 0.3960, with an average F1-score of 0.4205. Optimized final model was tested on the held-out test set.

$$\hat{L} = \text{Mode}\{T_1(X), T_2(X), \ldots, T_n(X)\} \quad (8)$$

In this case, $\hat{L}$ is the predicted class label and $T_i(X)$ is the prediction of the ith decision tree.

Table 5: Optimized Random Forest classification performance

Parameter	Value
Dataset	Texas_Labelled_SCADA.csv
Total Samples	8,760
Normal Samples	8,060
Anomalous Samples	700
Train-Test Split	75 : 25
Cross Validation	5-Fold
Number of Trees	100
Maximum Depth	None
Minimum Samples Split	10
Minimum Samples Leaf	1
Accuracy	93.06%
Precision	63.53%
Recall	30.86%
F1-score	41.54%
ROC-AUC	89.42%
Prediction Time	0.17 s
Maximum Features	log2
Nested CV F1 (Average)	0.4205

The optimized Random Forest provides the principal supervised classifier in the proposed framework. Its performance is compared with optimized XGBoost and with the broader seven-model benchmark.

3.7 XGBoost Classification

To further test the automatic anomaly-labelling framework, an Extreme Gradient Boosting (XGBoost) Classifier was created using the same Texas_Labelled_SCADA.csv dataset. The classifier takes the wind speed, wind direction, atmospheric pressure and the air temperature as input features and the automatically generated anomaly label as the target variable. Stratified sampling was used in the dataset to split it into a training subset of 75% and a testing subset of 25%.

The following best hyperparameters for the XGBoost model were found: learning rate = 0.05, max depth = 9, n estimators = 200, subsample = 1.0, column sampling by tree = 0.90. The reported nested cross-validation F1-scores were 0.4825, 0.4076, 0.4151, 0.4651, and 0.3907, with an average F1-score of 0.4322. An optimised model was finally tested on a held-out test set.

$$\hat{L} = \sum_{k=1}^{K} f_k(X) \quad (9)$$

where $\hat{L}$ is the predicted class label, $f_k(X)$ is the output of the kth boosted decision tree and K is the number of boosting iterations.

Table 6. Optimized XGBoost Classification Performance

Parameter	Value
Dataset	Texas_Labelled_SCADA.csv
Total Samples	8,760
Normal Samples	8,060
Anomalous Samples	700
Train-Test Split	75 : 25
Learning Rate	0.05
Maximum Depth	9
Number of Estimators	200
Subsample	1.00
Column Sample by Tree	0.90
Accuracy	93.24%
Precision	65.52%
Recall	32.57%
F1-score	43.51%
ROC-AUC	89.29%
Nested CV F1 (Average)	0.4322

3.8 Performance Evaluation

The same held out testing set was used to test the optimized Random Forest and XGBoost classifiers. The performance of the classification results was evaluated using the confusion matrix and the following performance measures: Accuracy, Precision, Recall, F1 Score, ROC-AUC and execution time. Since the number of labels produced by the anomalous class is only 7.99% of the total number of labels, special focus is paid to minority class Precision, Recall, and F1-score, instead of just accuracy.

Accuracies of classification are obtained as a proportion of correctly classified observations, and are defined as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (10)$$

Precision equals the number of observations in the set that are identified as anomalous over the total number of anomalies predicted, and is represented as a percentage:

$$\text{Precision} = TP / (TP + FP) \quad (11)$$

Recall: This is the ability of the classifier to report accurate anomalous observations and is calculated as

$$\text{Recall} = TP / (TP + FN) \quad (12)$$

The F1 score is the harmonic average of Precision and Recall and is written as

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (13)$$

Apart from these indicators, the ROC-AUC was used to test the classification models' discrimination power for various decision thresholds. In addition, the execution time of each of the classifiers was measured to evaluate the computational efficiency in real-life wind turbine condition monitoring. The performance of these two classifiers, Random Forest and XGBoost, is compared and discussed in terms of these evaluation metrics in Section 4.

3.9 Sensor Ablation Study

A sensor ablation study was also performed to examine the portion of the different variables in the SCADA to the anomaly detection - the Texas_Labelled_SCADA.csv dataset was used for this study. This experiment aimed to compare the impact of all the environmental factors on the classification accuracy of the proposed machine learning framework. When analyzing, three feature configurations were considered. In the first configuration, all four environmental variables (wind speed, wind direction, atmospheric pressure, and air temperature) were used as features for the input. The second setup was conducted without the atmospheric pressure variable, allowing them to focus on how this factor affects the anomaly detection. The third configuration excluded both of the abovementioned parameters and the classifiers were trained with wind speed and wind direction alone. The same training and testing partitions, optimized model parameters and evaluation procedure were used for all configurations to ensure a fair comparison.

The accuracy, precision, recall, f1 score and ROC-AUC were used to assess the classification performance of each feature combination. The anomalies detection performance for the three feature configurations was compared to find the impact of omitting individual SCADA variables on the overall anomaly detection performance. The sensor ablation analysis allows identification of the most informative environmental variables and analyzes the robustness of the proposed framework in the presence of fewer sensors. The corresponding performance results are reported in Section 4.

Table 7. Sensor ablation configurations

Configuration	Input Features
Configuration 1	Wind Speed + Wind Direction + Pressure + Temperature (All Sensors)
Configuration 2	Wind Speed + Wind Direction + Temperature (Without Pressure)
Configuration 3	Wind Speed + Wind Direction (Without Pressure and Temperature)

3.10 Local Outlier Factor (LOF) Screening

In order to achieve a complementary unsupervised abnormal operating behaviour (AOB) view, the SCADA variables were standardized and the labels of AOBs were not used during the LOF fitting. LOF detects observations with significantly lower local density as compared to the density of their neighbours. The four standardized environmental variables and 8,760 observations were used for the analysis. The reported configuration had 20 nearest neighbours and a contamination level of 5%, which yielded 438 density-based anomalies. There, LOF is therefore considered as a standalone screening tool that is not used to validate the deviation-based labels as a ground truth.

For each observation the LOF score is given as

$$LOF(p) = (1 / |N_k(p)|) \sum_{o \in N_k(p)} [LRD(p) / LRD(o)] \quad (14)$$

LOF predictions were also compared with automatically generated anomaly labels to measure the agreement between the two detection approaches in terms of the numerical value. But, since the two approach anomaly detection with different principles (power deviation vs local density), any discrepancies are seen as methodological complementarity, not as proof of confirmed physical faults.

Table 8. LOF screening parameters

Parameter	Value
Method	Local Outlier Factor (LOF)
Dataset	Texas_Labelled_SCADA.csv

Parameter	Value
Total Observations	8,760
Input Features	Wind Speed, Wind Direction, Pressure, Temperature
Output	LOF Score and Anomaly Label
Number of Neighbours	20
Contamination	0.05 (5%)

3.11 McNemar Statistical Test

McNemar's statistical test was used to statistically compare the prediction behaviour of the optimized Random Forest and XGBoost classifiers for the same held-out test observations. McNemar's test is a paired disagreement between two classifiers and compares if the difference between them is statistically significant compared to an actual difference, which is not significant. While aggregate performance statistics evaluate the performance of individual classifiers, McNemar's test examines their paired disagreements and compares if there is a statistically significant difference.The test will be based on 2×2 contingency table: b is the number of samples that the Random Forest classifier correctly classified and the XGBoost classifier misclassified; c is the number of samples that the XGBoost classifier correctly classified and the Random Forest classifier misclassified. The McNemar test statistic is computed as

$$\chi^2 = (|b - c| - 1)^2 / (b + c) \quad (15)$$

χ^2 is the statistic of the McNemar test. The calculated statistic is then compared with the Chi-square distribution to get the p-value.

The null hypothesis (H_0): There is no statistically significant difference between the classification performances of the Random Forest and XGBoost models and the alternative hypothesis (H_1): There is a statistically significant difference between the classification performances of the two models. In this research 0.05 was chosen as the level of significance. If the p-value calculated is less than 0.05, then the Null Hypothesis is rejected, and the difference in performances observed is statistically significant. The contingency table, the McNemar statistic and the p-value are presented in Section 4.

Table 9. McNemar test parameters

Parameter	Value
Statistical Test	McNemar Test
Compared Models	Random Forest vs. XGBoost
Test Dataset	Texas_Labelled_SCADA.csv
Significance Level ((\alpha))	0.05
Test Output	χ^2 Statistic and p-value

4. Results and Discussion

4.1 Expected Power Estimation

First, the expected power had been estimated using the Random Forest regression model based on the operating variables in the SCADA. The expected-power values are seen to correspond to the overall power behaviour of the SCADA observations (Fig. 4). The error of the power between the measured and the expected power was then taken as the criterion for abnormal operating conditions.

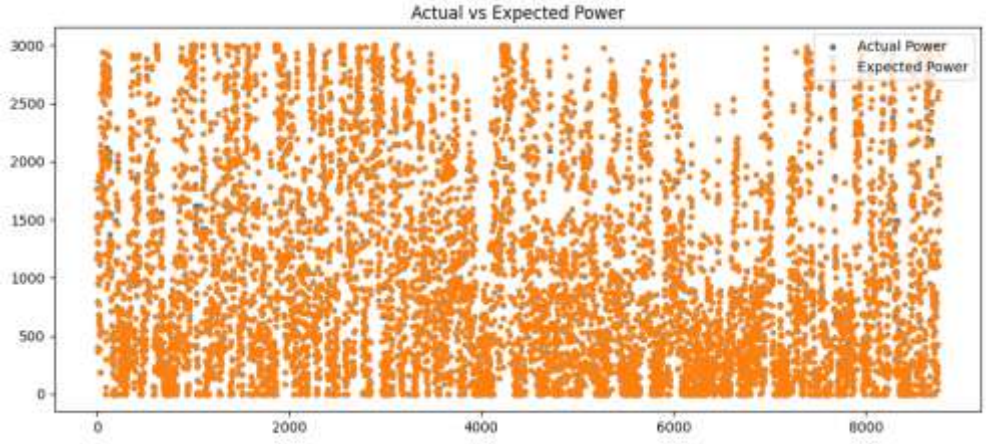


Figure 4. Actual versus expected power obtained using Random Forest regression.

4.2 Deviation-Based Anomaly Label Generation

For each SCADA observation, a deviation in absolute value between the measured and expected power was determined. A deviation distribution was generated and, instead of a fixed threshold, the distribution was analysed statistically to find an adaptive threshold. The mean deviation was 1.5010 kW, and the standard deviation was 2.6756 kW, which led to an adaptive threshold of 4.1767 kW.

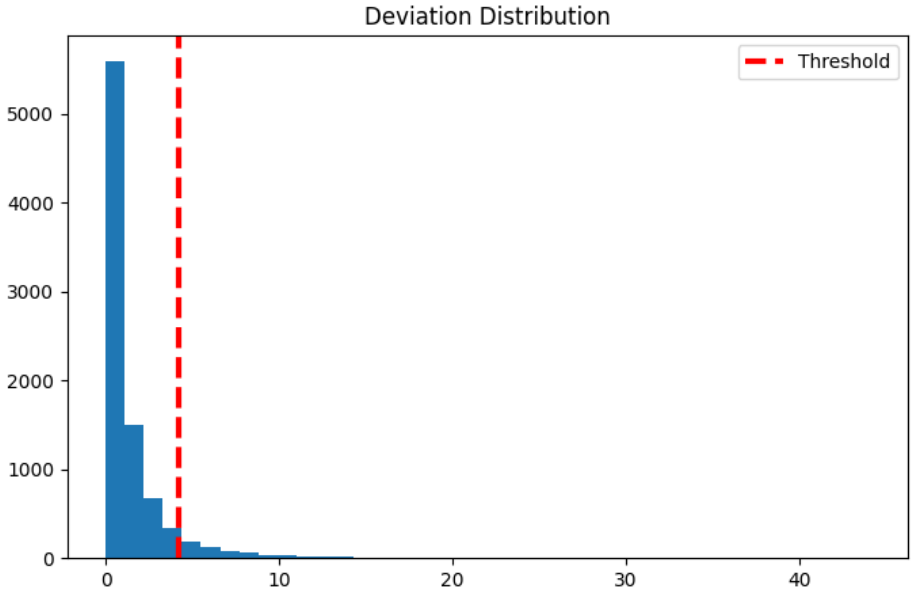


Figure 5. Distribution of power deviations and the adaptive anomaly threshold.

With the value of 4.1767 kW as threshold, the observations were automatically segregated into 8,060 normal samples and 700 abnormal samples. The prevalence of anomalies obtained was 7.99%. Classes are pseudo-labels based on expected-power deviation and not physical component faults.

4.3 Random Forest Classification Performance

Using stratified sampling, the automatically labelled data set was split into training and testing sets. A Random Forest (RF) classification was then trained to classify the turbine operating conditions as normal or abnormal.

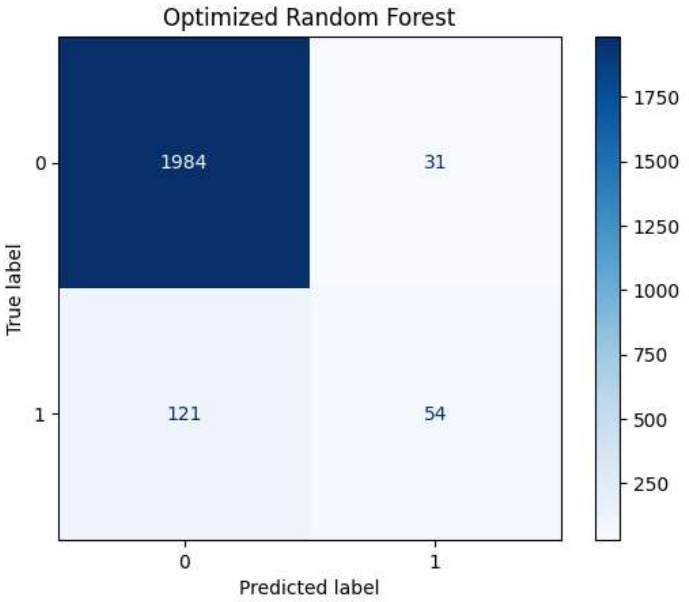


Figure 6. Random Forest confusion matrix.

Random Forest confusion matrix is shown in figure 6, which has 1,984 true negatives and 31 false positives; 121 false negatives and 54 true positives. These values give an accuracy of 93.06%, precision of 63.53%, recall of 30.86%, and F1-score of 41.54%. The reported ROC-AUC is 89.42%. Relatively low recall suggests a significant number of the anomalies are automatically labelled are not detected by the optimized classifier, which is a challenge because of the minority class.

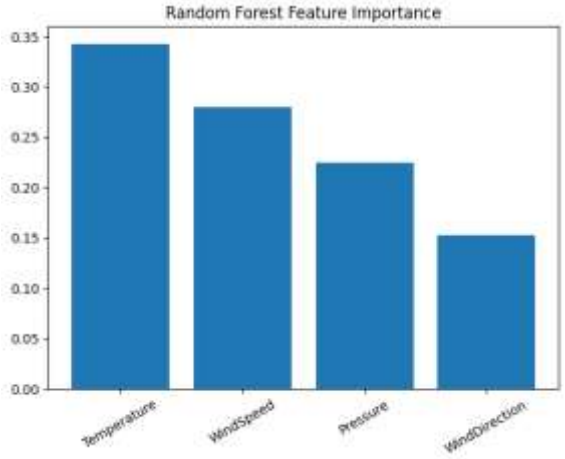


Figure 7. Feature importance obtained from the Random Forest classifier.

Figure 7 shows that air temperature has the highest relative feature importance (0.342848), followed by wind speed (0.279948), atmospheric pressure (0.224802), and wind direction (0.152402). This indicates that the environmental variables contribute unequally to the classification of the automatically generated anomaly labels.

4.4 XGBoost Classification Performance

The automatically labelled dataset and held out test set were then used to test XGBoost. The model is based on gradient-boosted decision trees for the classification of normal and anomalous operating conditions, and offers an independent comparison to the ensemble model Random Forest.

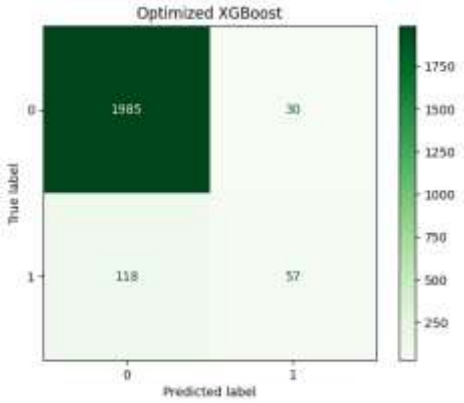


Figure 8. XGBoost confusion matrix.

The confusion matrix in Fig. 8 shows the distribution of correctly and incorrectly classified normal and anomalous observations.

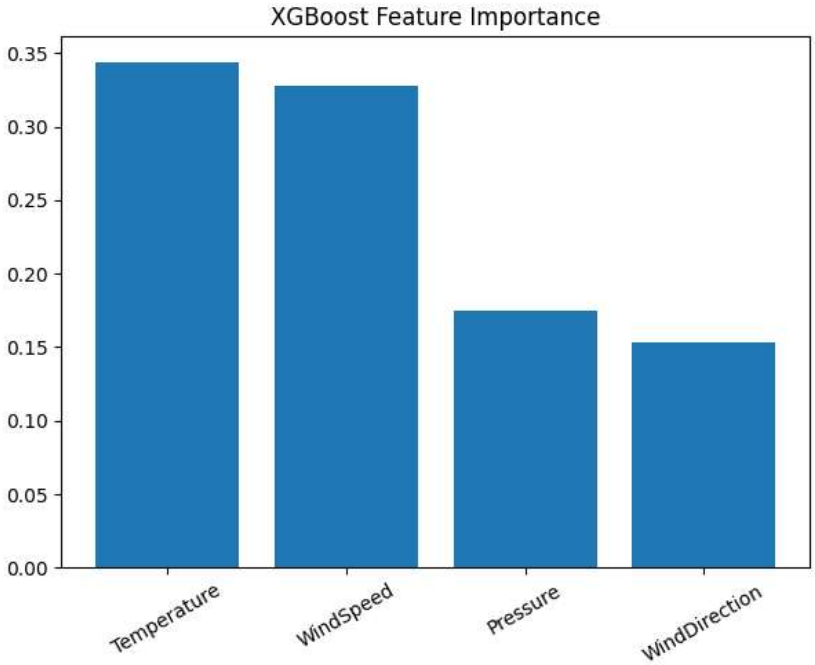


Figure 9. Feature importance obtained from XGBoost.

The feature-importance distribution in Fig. 9 provides an additional interpretation of the variables contributing to the classification decision.

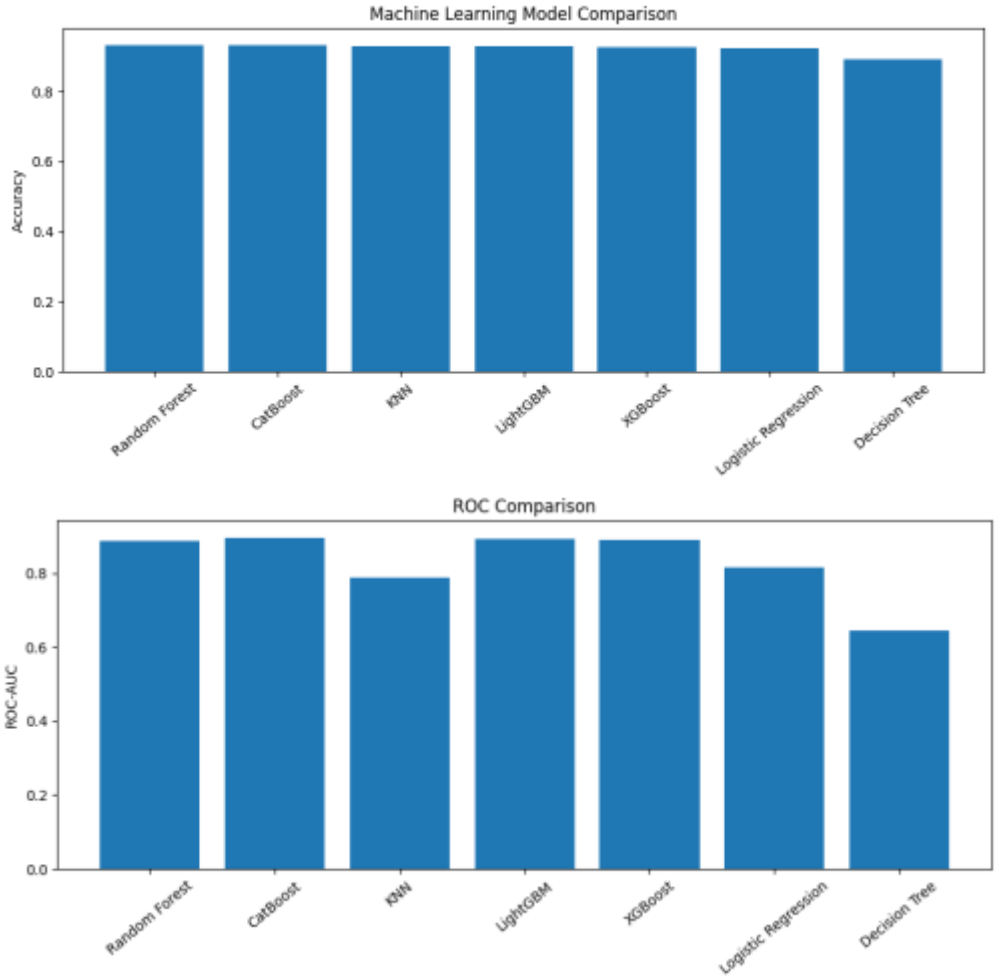
4.5 Comparative Machine Learning Performance

In order to test the effectiveness of the proposed approach, seven supervised classification algorithms were tested with the same data set labeled with anomalies. The comparison consists of Logistic Regression, K-Nearest Neighbors, Decision Tree, LightGBM,CatBoost,XGBoost and the proposed Random Forest model.

Table 10. Comparative performance of machine learning models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
Logistic Regression	92.47	62.50	14.29	23.26	81.52
K-Nearest Neighbors	92.97	61.80	31.43	41.67	78.90
Decision Tree	89.18	33.33	35.43	34.35	64.64
LightGBM	92.92	61.11	31.43	41.51	89.24
CatBoost	93.15	64.37	32.00	42.75	89.70
XGBoost	92.79	57.94	35.43	43.97	88.95
Random Forest (benchmark)	93.38	66.30	34.86	45.69	88.63

Table 10 shows the seven-model benchmark which is independent of the other comparative experiments conducted based on the common benchmark configurations reported in the attached analysis output. The highest benchmark accuracy (93.38%), precision (66.30%) and F1-score (45.69%) were obtained with Random Forest. XGBoost recorded the best benchmark recall (35.43%) and CatBoost the best benchmark ROC-AUC (89.70%). The optimized (Sections 4.3 and 4.4) RF and XGBoost results should not be confused with the benchmark results in this section as the optimized classifiers were optimized individually and are the ones used for the McNemar comparison. The Decision Tree had the lowest overall performance with 89.18% accuracy and 64.64% ROC-AUC. Logistic Regression obtained a good accuracy (92.47%) but its recall was only 14.29% due to the fact that the anomalous class is less than the normal class.



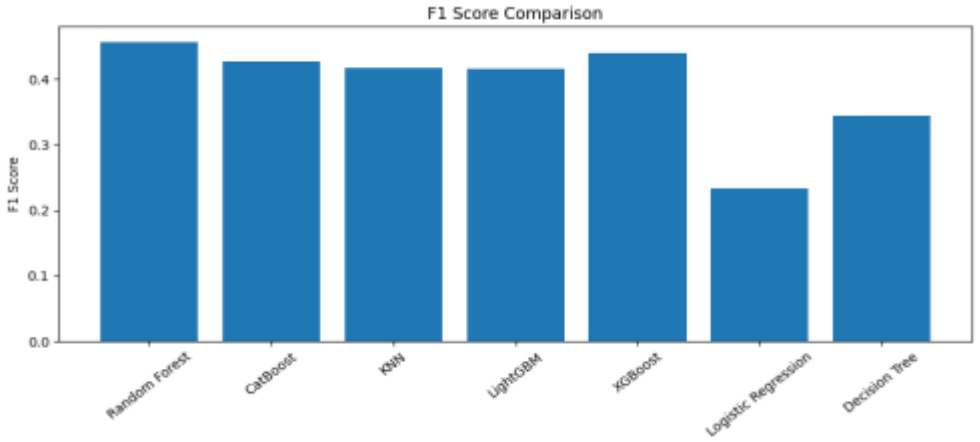


Figure 10. Comparative performance of the evaluated machine learning models.

Figure 10 provides a visual comparison of the seven classifiers. The benchmark results show that the Random Forest configuration reported as the proposed model provides the highest accuracy, precision, and F1-score in this comparative experiment, although XGBoost remains competitive.

4.6 Statistical Significance of Random Forest and XGBoost

The optimized Random Forest and XG Boost classifiers showed relatively similar performance but the differences in predictors between the two classifiers were compared to see if they were statistically significant using McNemar's test. Out of these 2047 observations were correctly classified by both models, 20 observations were correctly classified by Random Forest model only, 24 observations were correctly classified by XGBoost model only and 99 observations were misclassified by both models. The resulting McNemar statistic was 0.2045, with $p = 0.651077$.

Because $p = 0.651077 > 0.05$, the null hypothesis is not rejected. Therefore, the difference between the prediction accuracy of the optimized Random Forest and XGBoost classifiers is not statistically significant at the 5% level of significance.

4.7 Sensor Ablation Analysis

A sensor ablation experiment was conducted to determine the contribution of the atmospheric variables to wind turbine anomaly classification. Three configurations were evaluated:

1. All Sensors: Wind Speed, Wind Direction, Pressure, and Temperature.
2. Without Pressure: Pressure removed while retaining Temperature.
3. Without Pressure & Temperature: Both atmospheric pressure and temperature removed.

Table 11. Sensor ablation performance comparison

Feature Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
All Sensors	94.38	81.43	67.06	73.55	95.81
Without Pressure	93.24	74.88	63.14	68.51	94.70
Without Pressure & Temperature	90.09	61.18	40.78	48.94	91.57

It is evident that sensor removal has an effect on these results. The All-Sensors configuration performed best with an accuracy of 94.38%, a precision of 81.43%, a recall of 67.06%, an F1-score of 73.55% and an ROC-AUC of 95.81%. The accuracy dropped to 93.24% when the pressure was removed, and a reduction

in F1-score from 73.55% to 68.51%. The larger the reduction was when both pressure and temperature were eliminated. For this one, the accuracy was reduced to 90.09%, recall to 40.78% and F1 to 48.94%. The ROC-AUC also decreased from 95.81% to 91.57%.

4.8 Unsupervised LOF Anomaly Detection

The results of the complete ablation output corroborate the downward trend in each of the three configurations. The complete SCADA feature set was used to test the anomaly detection algorithm, Local Outlier Factor (LOF) without using any pre-defined anomaly labels. In the LOF configuration, 20 nearest neighbours and contamination level of 5% were used.

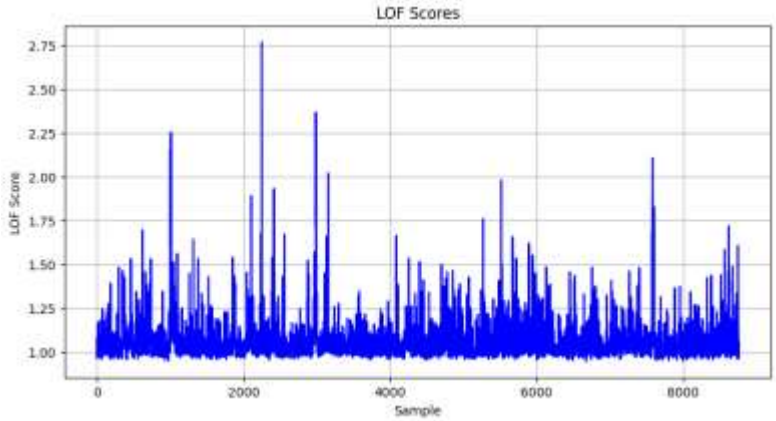


Figure 11. LOF-based anomaly detection results.

A total of 438 anomalous observations were identified out of 8,760 samples by the LOF analysis, which gave an anomaly rate of 5.0%. The average LOF score in the report from the analysis was 1.0579. The results with LOF labels against the automatically computed anomaly labels were: 88.42% accuracy, 14.16% precision, 8.86% recall and 10.90% F1 score; the confusion matrix was [7684, 376], [638, 62]. The low minority-class agreement value suggests that there is a remarkable difference between the observations identified by the deviation-based method and the observations identified by the density-based method. Thus, LOF should not be used as a substitute for the automatically generated labels.

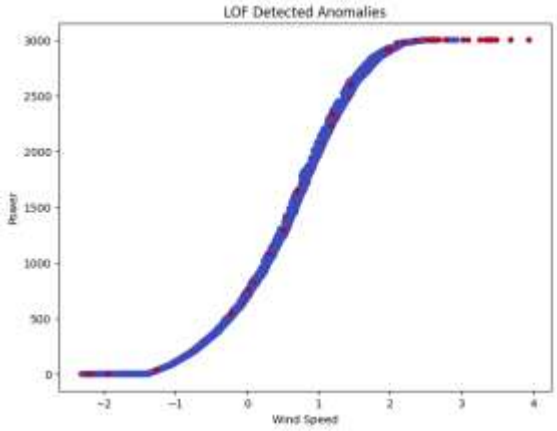


Figure 12. LOF score distribution and detected anomalous observations.

In Figures 11 and 12 an unsupervised perspective on abnormal operating behaviour is provided, independent of the observer. The labels that are automatically generated are not used in LOF, which identifies observations that are locally sparse in the SCADA feature space. This is helpful for screening unlabeled operational data, making LOF useful.

4.9 Integrated Discussion

The results enable the development of a workflow from the estimation of expected power to supervised and unsupervised anomaly screening. The reference power for deviation-based anomaly labelling is obtained by using the Random Forest regression and optimized Random Forest and XGBoost are used for labelling. There is an added comparison with the 7-model benchmark for the machine-learning algorithms that are widely used. The results for sensor ablation indicate that the inclusion of atmospheric information is worthwhile as it degrades anomaly-detection performance. LOF detects 438 density-based anomalies without reliance upon the labels produced, and the moderate number of agreements with the deviation-based labels suggests that both give different types of abnormal behaviour.

The optimized two model comparison results were 93.06% accuracy and 41.54% F1 score for Random Forest, and 93.24% accuracy and 43.51% F1 score for XGBoost. XGBoost therefore made a little higher numerical performance on the selected test set, but for this difference McNemar's test showed that it was not statistically significant ($p = 0.651077$). Seperate seven-model benchmark: Random Forest was the model with the best F1 value (45.69%). These results justify the report of the optimized RF-XGBoost comparison and the whole benchmark as separate experiments.

Thirdly, the sensor ablation experiment showed that a decrease in the amount of the sensor information clearly leads to a drop in anomaly-detection performance. The all-sensor configuration had an accuracy of 94.38% and an F1-score of 73.55%, while the sensor configuration without the pressure sensor and the temperature sensor yielded accuracy of 90.09% and the F1-score of 48.94%. Lastly, LOF found 438 observations (5.0%) classified as density-based anomalies. The low agreement between LOF and the automatically generated labels does not necessarily reflect failure of either approach since they are based on different definitions of anomalies. Rather, the result shows that the anomaly detection using SCADA is very sensitive to the definition of abnormal behavior and that multiple complementary detection mechanisms can be applied.

5. Conclusion

This work aimed to build an intelligent approach for automated detection of anomalies on wind turbines based on 8,760 observations with SCADA data from the state of Texas. With no manual fault annotation, an adaptive deviation threshold of 4.1767 kW yielded 8,060 normal and 700 anomalous observations. Optimized Random Forest and XGBoost classifiers had 93.06% and 93.24% accuracy and 41.54% and 43.51% F1-scores, respectively. Their prediction difference was not statistically significant as indicated by McNemar's test ($p = 0.651077$). Random Forest had the best F1-score (45.69%) in the separate seven-model benchmark. The sensor-ablation analysis revealed the intensity of the drop in performance in the absence of pressure and temperature data. The 438 density-based density anomalies (5.0%) that LOF was able to identify were found independently, and there was only limited agreement with the deviation-based labels, confirming that the two methods come from different ideas of abnormal behaviour. The main drawback is that there are no ground truth physical anomaly labels provided in the Texas SCADA dataset, and the anomaly classes generated should therefore be considered as deviation-based anomaly labels. Another methodological limitation is that the current regression output contains the residuals across all 8760 observations, which should be recalculated with out-of-fold or strictly held-out predictions to avoid any potential optimistic bias in the validation results. For future work, the framework needs to be tested with fault-labeled data from several turbines, chronological/evaluation over time, out-of-fold generation of labels, further operational and component-related variables, and real-time deployment strategies.

References:

[1] Z. Mustaffa and M. H. Sulaiman, “Random Forest based wind power prediction method for sustainable energy system,” Cleaner Energy Systems, vol. 12, p. 100210, 2025, doi: 10.1016/j.cles.2025.100210.

[2] R. Becker and D. Thrän, “Completion of wind turbine data sets for wind integration studies applying random forests and k-nearest neighbors,” Applied Energy, vol. 208, pp. 252–262, 2017, doi: 10.1016/j.apenergy.2017.10.044.

[3] C. Zhang, D. Hu, and T. Yang, "Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost," Reliability Engineering & System Safety, vol. 222, p. 108445, 2022, doi: 10.1016/j.ress.2022.108445.
[4] P. W. Khan and Y.-C. Byun, "Fault prognosis of wind turbines using multimodal machine learning," Energy Reports, vol. 12, pp. 4938–4952, 2024, doi: 10.1016/j.egyr.2024.10.043.
[5] P. Rengasamy and R. Rajesh, "Explainable artificial intelligence framework for wind turbine fault detection using random forest–extreme gradient boosting hybrid model," Results in Engineering, vol. 28, p. 107446, 2025, doi: 10.1016/j.rineng.2025.107446.
[6] P. Jia and G. Chen, "Wind power icing fault diagnosis based on slow feature analysis and support vector machines," in Proc. 2020 10th Int. Conf. Power and Energy Systems (ICPES), 2020, pp. 398–403, doi: 10.1109/ICPES51309.2020.9349697.
[7] P. W. Khan and Y.-C. Byun, "Multi-fault detection and classification of wind turbines using stacking classifier," Sensors, vol. 22, no. 18, p. 6955, 2022, doi: 10.3390/s22186955.
[8] G. Li, C. Wang, D. Zhang, and G. Yang, "An improved feature selection method based on random forest algorithm for wind turbine condition monitoring," Sensors, vol. 21, no. 16, p. 5654, 2021, doi: 10.3390/s21165654.
[9] K. Gupta and S. K. Sinha, "SCADA system dataset exploration and machine learning based forecast for wind turbines," Results in Engineering, vol. 16, p. 100640, 2022, doi: 10.1016/j.rineng.2022.100640.
[10] C. McKinnon, J. Carroll, A. McDonald, S. Koukoura, and C. Plumley, "Investigation of isolation forest for wind turbine pitch system condition monitoring using SCADA data," Energies, vol. 14, no. 20, p. 6601, 2021, doi: 10.3390/en14206601.
[11] F. R. B. Karim, N. Liu, and J. Zhang, "Early anomaly detection in wind turbine generators using SCADA data," in Proc. 2025 IEEE Symp. Diagnostics for Electric Machines, Power Electronics and Drives (SDEMPED), 2025, pp. 1–7, doi: 10.1109/SDEMPED53223.2025.11154170.
[12] P. W. Khan and Y.-C. Byun, "Detecting anomaly classification using PCA-Kmeans and ensembled classifier for wind turbines," IEEE Open Access Journal of Power and Energy, vol. 11, pp. 349–361, 2024, doi: 10.1109/OAJPE.2024.3437414.
[13] P. W. Khan, C. Y. Yeun, and Y. C. Byun, "Fault detection of wind turbines using SCADA data and genetic algorithm-based ensemble learning," Engineering Failure Analysis, vol. 148, p. 107209, 2023, doi: 10.1016/j.engfailanal.2023.107209.
[14] A. Agasthian, R. Pamula, and L. A. Kumaraswamidhas, "Fault classification and detection in wind turbine using cuckoo-optimized support vector machine," Neural Computing and Applications, vol. 31, no. 5, pp. 1503–1511, 2019, doi: 10.1007/s00521-018-3690-z.
[15] J. Zhang, H. Dong, R. Shan, G. Hou, and C. Huang, "Fault diagnosis of wind turbine gearbox based on SSDAE-ELM," in Proc. 2021 33rd Chinese Control and Decision Conf. (CCDC), 2021, pp. 5407–5412, doi: 10.1109/CCDC52312.2021.9601825.
[16] Y. Ye, X. Zheng, S. Yang, and J. Wang, "Fault diagnosis of wind turbines bearing based on particle filter and sparse auto-encoder," in Proc. 2021 6th Int. Conf. Power and Renewable Energy (ICPRE), 2021, pp. 268–273, doi: 10.1109/ICPRE52634.2021.9635465.
[17] A. Czyżewski, "Remote health monitoring of wind turbines employing vibroacoustic transducers and autoencoders," Frontiers in Energy Research, vol. 10, 2022, doi: 10.3389/fenrg.2022.858958.
[18] B. Du, Y. Narusue, Y. Furusawa, N. Nishihara, K. Indo, H. Morikawa, and M. Iida, "Clustering wind turbines for SCADA data-based fault detection," IEEE Transactions on Sustainable Energy, vol. 14, no. 1, pp. 442–452, Jan. 2023, doi: 10.1109/TSTE.2022.3215672.
[19] Geno Peter and M. Rajaram, "An enhanced Z-source inverter topology-based permanent magnet brushless DC motor drive speed control," International Journal of Electronics, vol. 102, no. 8, pp. 1289–1305, Aug. 2015, doi: 10.1080/00207217.2014.971350.
[20] Y. Cui, P. Bangalore, and L. Bertling Tjernberg, "An anomaly detection approach based on machine learning and SCADA data for condition monitoring of wind turbines," in Proc. IEEE Int. Conf.

Probabilistic Methods Applied to Power Systems (PMAPS), 2018, pp. 1–6, doi: 10.1109/PMAPS.2018.8440503.
[21] H. Z. Anonto et al., “Intelligent energy management of microgrids using machine learning: Leveraging random forest models for solar and wind power,” Results in Engineering, vol. 27, p. 106539, 2025, doi: 10.1016/j.rineng.2025.106539.
[22] Y. Selvaraj and C. Selvaraj, “Proactive maintenance of small wind turbines using IoT and machine learning models,” International Journal of Green Energy, 2021, doi: 10.1080/15435075.2021.1930004.
[23] W. Udo and Y. Muhammad, “Data-driven predictive maintenance of wind turbine based on SCADA data,” IEEE Access, vol. 9, pp. 162370–162388, 2021.
[24] Z. Liu, X. Wang, and L. Zhang, “Fault diagnosis of industrial wind turbine blade bearing using acoustic emission analysis,” IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 9, pp. 6630–6639, Sept. 2020.
[25] Y. He, J. Liu, S. Wu and X. Wang, "Condition Monitoring and Fault Detection of Wind Turbine Driveline with the Implementation of Deep Residual Long Short-Term Memory Network," in IEEE Sensors Journal, vol. 23, no. 12, pp. 13360-13376, 15 June15, 2023, doi:10.1109/JSEN.2023.3273279.
[26] X. Cheng, F. Shi, X. Liu, M. Zhao and S. Chen, "A Novel Deep Class-Imbalanced Semi supervised Model for Wind Turbine Blade Icing Detection," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 6, pp. 2558-2570, June 2022, doi: 10.1109/TNNLS.2021.3102514.
[27] A. M. Peco Chacón, I. Segovia Ramírez, and F. P. García Márquez, “False alarm detection in wind turbine by classification models,” Advances in Engineering Software, vol. 177, p. 103409, 2023, doi: 10.1016/j.advengsoft.2023.103409.