FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme / specialisation: MSc in Computer Science - Reliable and Secure Systems	2022 HØST-2023 VÅR Open
Author: Manisha Pranav, Sakarvadia	
Supervisor at UiS: Antorweep Chakravorty Co-supervisor: External supervisor(s): Amund Haugeseth (Tietoevry Create, Norway)	
Thesis title: Predictive maintenance with industrial sensor data Credits (ECTS): 30	
Keywords: Wind turbine, fault detection, predictive maintenance	Pages: + appendix: Stavanger June 15,2023

Predictive maintenance with industrial sensor data

Manisha Sakarvadia[a]

[a] *University of Stavanger, Norway*

Abstract

The Norwegian Ministry of Petroleum and Energy Commissions report shows that the government is making a large step closer to its ambition of allocating regions for 30,000 MW offshore wind via way of means of 2040. According to a report by IRENA, offshore wind operation and maintenance (O & M) costs make up a significant portion of the overall cost of electricity for offshore wind farms in G20 countries, ranging from 16-25%. To address this issue, it is essential to explore methods for improving operational reliability and reducing the maintenance costs of wind turbines. One promising approach is predictive maintenance, which involves leveraging data collected from sensors already equipped with the turbines to detect and address potential issues before they become more serious. Predictive maintenance is important in wind farms to reduce downtime and optimize the performance of wind turbines. Various rotating components in wind turbines make them complicated machinery, and if any of those parts fails, it can cause the entire turbine to shut down. This can result in lost revenue for the wind farm operator and lead to higher maintenance costs if the problem is not addressed quickly. This can be possible through a Supervisory Control and Data Acquisition (SCADA) system, which collects and analyzes data from various turbine components. We have developed a method for detecting and monitoring failures in critical components such as the gearbox and generator, based on historical SCADA data. Our approach utilizes machine learning models, specifically extreme gradient boosting (XGBoost), and has been tested on two real-world case studies involving eight different turbines. The outcomes show both the effectiveness and usefulness of our technique for boosting wind turbine reliability and minimizing maintenance costs.

Keywords: Wind turbine, fault detection, supervisory control and data acquisition (SCADA), extreme gradient boosting (XGBoost), predictive maintenance

*Corresponding author
Email address: mp.sakarvadia@stud.uis.no (Manisha Sakarvadia)

1. Introduction

There is little doubt that the global energy sector has a large role in greenhouse gas emissions. To reduce carbon emissions, sustainable low-carbon development is necessary. Modern energy requirements have pushed renewable energy forward significantly. Wind energy is generating a lot of interest from renewable energy investors due to its advantageous characteristics. Wind energy is a renewable energy source that is free of pollution and has very high growth potential[1]. The number of wind turbines in use has risen as a result of the rapid development of wind power. However, components such as blades, bearings, gears and generators are susceptible to failure, resulting in higher maintenance and operational costs. To improve the reliability of wind turbines and prevent potential accidents, it is essential to monitor their condition and detect faults. This can help reduce economic losses and support the continued growth of the wind power industry through effective maintenance and planning.

In an effort to eliminate emissions related to climate change by 2040, the Norwegian government's goal is to increase the country's total wind power capacity to 12-14 gigawatts (GW) by 2030 and 30-34 GW by 2040. As stated in the 2021 annual report of the IEA Wind TCP - Global Wind Energy Research Collaboration, wind energy in Norway generates 11.8 TWh of electricity, representing 8.5% of the country's total electricity consumption [2].

According to recent studies, there is a growing trend towards the installation of wind turbines in offshore environments rather than onshore. However, this shift brings with it a higher level of complexity when it comes to maintaining such equipment. The International Renewable Energy Agency (IRENA) has reported that offshore wind operation and maintenance (O & M) costs make up a significant proportion of the total cost of electricity for offshore wind farms in G20 countries, typically ranging from 16% to 25%.

To reduce these costs, there is a need to optimize O & M practices and minimize unscheduled maintenance. This can be achieved through advancements in data collection and analytics, which can enable predictive maintenance and improve production output optimization. Therefore, it is essential to explore innovative solutions that allow for more efficient and cost-effective O & M practices to be developed and implemented in the offshore wind industry[3].

A wind turbine is a device that utilizes rotating blades to harness the energy of the wind and transform it into electrical energy through drive trains. These drive-trains can be classified into two types: direct drive and gear type, which utilizes a gearbox[4]. Both types have a hub as their input, a main shaft as their transfer, and a generator as their output. Other essential components of a wind turbine include main shaft bearings, mechanical brake, shaft bearing, yaw systems, power electronic systems, and hydraulic and cooling systems. The gearbox and generator play crucial roles in the conversion of energy from the wind turbine components mentioned above.

To reduce weight and enhance the transmission ratio, planetary transmission is widely used in wind turbine gearboxes, which operate in high-altitude nacelles[5]. This has resulted in the design of planetary/spur gearbox systems,

where the spur gearbox is the fixed gearbox stage. Wind turbine gearboxes have a fixed gearbox stage that increases the rotational speed of the planetary gear, leading to induced vibration, which manifests as strong noise in the gearbox [6]. However, diagnosing faults in the gearbox system can be difficult due to the random nature of wind and the time-varying rotational speed. When wind hits the turbine's blade, the kinetic energy is converted into rotational energy by the shaft connected to the blade. The moving shaft is then connected to a generator that produces electrical power through electromagnetism [3].

In gearbox-driven wind turbines, the double-fed induction generator is extensively used. Its operation mode is based on the rotational speed of the rotor winding's and stator windings connected to the transformer. The rotor windings are connected to the power grid via an inverter, which controls slip power depending on the rotor's rotational speed. The rotor sends electricity to the grid at ultra-synchronous pace, while the stator transfers all lively electricity to the grid on the synchronous pace of the generator [7]. The generator shaft is supported by bearings and is one of the most critical components in a wind turbine. As the shaft continuously rotates, bearing damage may occur, and effective fault detection is necessary.

Wind turbines are often deployed in harsh and remote locations, such as offshore environments, to maximize wind motion. Maintenance can be risky and costly, requiring using cranes or helicopters to raise upkeep crews. Therefore, monitoring the equipment is necessary to avoid such activities and perform maintenance when needed. Organizations typically adopt reactive, preventive, or predictive maintenance programs to increase operational reliability and decrease costs. In reactive maintenance, repairs are performed when components become defective. Preventive maintenance is achieved at a ordinary price to keep away from failures, however the challenge is figuring out while to carry out maintenance. Organizations use a conservative approach in planning maintenance for safety-critical equipment, which can result in machine life being wasted if maintenance is scheduled too early. Predictive maintenance is an effective approach because it predicts when failure will occur and schedules maintenance just before it [8].

This has resulted within side the layout of planetary/spur gearbox systems, wherein the spur gearbox is the constant gearbox stage. Wind turbine gearboxes have a fixed gearbox stage that increases the rotational speed of the planetary gear,leading to induced vibration, which manifests as strong noise in the gearbox[5]. However, because wind is stochastic and the rotational speed changes over time, diagnosing problems in the gearbox system can be difficult. When wind hits the turbine's blade, the kinetic energy is converted into rotational energy by the shaft connected to the blade. The moving shaft is then connected to a generator that produces electrical power through electromagnetism [3]. In gearbox-driven wind turbines, the double-fed induction generator is extensively used depending on the rotational speed of the rotor windings and stator windings connected to the transformer. The rotor windings are linked to the power grid through an inverter that regulates slip power according to the rotational speed of the rotor. The generator sends power to the grid at a

frequency slightly above or below the grid frequency, at the same time as the stator transfers all energetic strength to the grid on the synchronous pace of the generator. This system allows the wind turbine to generate electrical power efficiently and effectively. However, the generator shaft is a critical component that requires effective fault detection and maintenance to prevent costly and dangerous failures, especially in remote or offshore locations. Maintenance activities can be dangerous and costly, requiring the use of cranes or helicopters to lift maintenance crews. Therefore, monitoring the equipment is necessary to avoid such activities and perform maintenance when needed.

1.1. Maintenance optimization in Wind industry

Organizations normally undertake reactive, preventive, or predictive maintenance programs to increase operational reliability and decrease costs. In reactive maintenance, repairs are performed when components become defective. Preventive maintenance is carried out at a regular rate to avoid failures, but the challenge is determining when to perform maintenance. Organizations use a conservative approach in planning maintenance for safety critical equipment, which can result in machine life being wasted if maintenance is scheduled too early. Predictive maintenance is an effective approach because it predicts when failure will occur and schedules maintenance just before it[8].

Predictive maintenance relies on condition monitoring (CM), which allows maintenance of equipment and components that are likely to fail, and replace them at the appropriate time[9]. By carrying out maintenance just-in-time, predictive maintenance can help asset managers bridge the gap between reactive and scheduled maintenance. Predictive maintenance involves estimating the remaining useful life, detecting anomalies and identifying faulty components that need fixing. The challenge of predictive maintenance can be addressed through first-principles modelling, which is based on a physics-based approach that does not require data from the wind turbines, but requires a substantial amount of expert domain knowledge. It involves deriving equations that describe the system's behavior and using them to determine how the equipment will degrade and eventually fail over time. On the other hand, data-driven modeling does not require expertise in the system evaluation but instead requires significant amounts of data collected from the real-world system. Statistical and machine learning techniques are then used to develop models based on the data to understand the system's behavior and how it fails [10]. Hybrid approaches, where data-driven strategies are used to fill gaps in knowledge about the system's first principles, are also being explored.

Autonomous condition monitoring systems have seen a rapid increase in popularity over the past decade, with wind turbines being one of the equipment types monitored. One strategy for condition monitoring is to retrofit vibration sensors, strain gauges, or oil particle counters to sub-components of the turbine for localized monitoring[11]. However, the cost of retrofitting sensors and collecting and analyzing data for performance insights can be a problem with this strategy[12]. Wind turbines are already equipped with sensors that record data on equipment status. These sensors form part of a Supervisory Control

and Data Acquisition (SCADA) system that was initially installed for monitoring and operating the turbine system. However, engineering data obtained from this system is now being used to identify anomalies and assess the health status of wind turbines, allowing for data-driven predictive maintenance[13]. The SCADA system's sensors are typically located in the turbine's main components, and data is usually sampled at 10-minute intervals, making it easy to transfer and store data in a database for later retrieval[14].

Wind turbines equipped with SCADA systems record various wind and performance parameters such as wind speed, power output, rotor speed, blade pitch angle, tower and drivetrain acceleration, bearing temperature, and gearbox temperature. The data captured by SCADA systems can be utilized for fault detection and prognosis activities[15]. The availability of this data has facilitated the development of a condition monitoring system based on SCADA data analysis, which can be assessed at various granularities. Monitoring at the sub-component level, such as the drivetrain, can help to detect faults more accurately. On the other hand, monitoring at the whole wind turbine level by combining signals of different components can provide a higher-level warning[9]. To prioritize components for monitoring, the decision should be based on their failure rates and downtime per failure. Components that are more prone to failure and have longer lead times for replacement should be given more attention[16]. According to a survey conducted on two wind farms in China, 68% of the total downtime was caused by faults in the generator, converter, and pitch systems[17].

Typically, SCADA systems provide data that represents both normal and faulty operations. However, sometimes we may lack sufficient data for faulty operation, for instance, when some sensors are broken. In such cases, we can build a mathematical model of the equipment and estimate its parameters using available sensor data. To generate failure data, we can then simulate this model with various fault states under different operating conditions. This generated data can then be used alongside sensor data to develop our algorithm. After acquiring the data, the next step is to remove outliers and filter out noise from the data to ensure its accuracy[8]. In our research, we only have sensor data for normal operation and not for faulty operation. Therefore, in order to develop an algorithm for predictive maintenance, it is essential to construct a mathematical model of the wind turbine and generate failure data. This process requires extensive knowledge of the wind turbine's performance. In this paper, we will review various approaches that have used SCADA data for wind turbine fault detection and prediction. Our contribution to knowledge will involve:

1. developing a data-driven approach for predictive maintenance using SCADA data without failure data
2. validating the approach with data from a different wind farm having failure data.

1.2. Outline

In this research, we intend to examine data obtained from a wind farm located in France that is managed by ENGIE. The wind farm comprises four

2MW wind turbines. Section 2 of the paper will review previous research related to the ENGIE dataset and similar solutions proposed by other studies. In Section 3, we will outline our methodology for identifying failures by pre-processing the data, developing a model, and performing post-processing. The findings of our approach will be showcased in Section 4, where we will demonstrate its application through a real-world case study conducted on an operational wind farm situated in Meuse, France. To assess the efficacy of our proposed solution, we will compare its performance against data obtained from a wind farm that already has documented failure data. In Section 5, we will discuss the effectiveness and practicality of our fault detection algorithm, and in Section 6, we will summarize the study and consider possible future steps for this research.

2. Literature Review

In recent years, machine learning techniques have been used to achieve predictive maintenance by developing inductive models that can learn the underlying structures in SCADA data of wind turbines. These models can predict potential faults and anomalies in advance[8]. Many of the existing studies in this field have utilized supervised methods such as regression or classification. These methods offer the benefit of establishing a clear relationship between input and output variables[18]. This section will review the current research on regression-based anomaly detection and its application to the ENGIE dataset.

2.1. Regression-based research works

Wind farms utilize a technique called condition monitoring, which involves constructing a model of the typical behavior of wind turbines and their components. This approach entails using a set of independent input variables, like wind speed, to create a regression model that can predict a numerical dependent output variable, such as power, under the assumption that the component is ideal. One critical aspect of wind turbine power curve modeling is that the power curves provided by manufacturers were tested under specific weather conditions, which may differ from those at the installation site [19]. To address this challenge, a study [20]compared four data-mining techniques, namely, cluster center fuzzy logic, neural network, K-Nearest Neighbour, and Adaptive Neuro-Fuzzy Inference System (ANFIS), to monitor wind turbine power output and detect deviations. The models were initially created using only one input variable, wind speed, and an output variable, power. However, by incorporating wind direction and ambient temperature as input variables, the models fit better with the data. The research concluded that ANFIS, which combines neural networks and fuzzy theory, achieved the highest performance. A study [21]investigated the use of machine learning to model wind turbine components, specifically the generator, comparing the performance of extreme gradient boosting (XGBoost) and long-short term memory (LSTM) based on mean absolute error (MAE). Results showed that XGBoost had a lower MAE and was more computationally efficient, executing 150 times faster than LSTM. The predicted outcomes

were compared with field measurements to detect anomalies. A separate investigation [22] has devised a framework for anomaly detection and parameter identification. This framework utilizes an auto-encoder neural network, which incorporates an LSTM network within its neuronal structure. A support vector regression-based adaptive threshold was applied to decrease the false alarm rate for anomaly detection. To validate the effectiveness of this approach, the researchers employed SCADA data from a wind farm situated close to the southern coast of Ireland. In another investigation [12], SCADA data was utilized, specifically focusing on the generator temperature and gearbox oil temperature aiming to establish a baseline temperature model. The variations between the predicted and actual values was calculated and analyzed using an exponentially weighted moving average (EWMA) control chart proposing a fixed threshold vs dynamic threshold for fault detection.

In study [23], an adaptive elastic network was used for feature selection, and a combination of convolutional neural network (CNN) and long-short term memory (LSTM) was employed to establish logical relationships between observed variables. This method was effective in detecting over-temperature in the high-speed side of the gearbox bearing. Another study [24] proposed a model for detecting abnormal spikes in wind turbine components by adjusting temperature data for the effects of ambient temperature and power output. Regression models were built using input variables (power output and ambient temperature) and an output variable (component temperature). Linear regression was selected as the best model, and the residual between the model's output temperature and raw temperature data was used to detect abnormal component behavior. Predictive analytics of wind turbine gearbox based on support vector regression (SVR) models for accurate prediction of gearbox oil and bearing temperature were carried out in another study [25]. Statistical tests were used to analyze the residuals and establish the robustness of the tested SVR model. The Mahalanobis distance method was applied for feature selection in another study [26], reducing the input variables fed into the LSTM prediction model. Fault detection was measured using the error between the predicted temperature of component wih the actual measurement [27] yielded more efficient and accurate results, lowering root mean square error by 4% compared to traditional backpropagation neural networks. Parameters of SCADA measurements used to build data-driven normal behavior models using SVR with a Gaussian kernel and principal components analysis (PCA) to orthogonalize and reduce feature dimensions.

In the study [28], a comprehensive methodology for fault detection of wind turbines using artificial neural networks and statistical process control was proposed. The methodology was tested on an operational wind turbine in Italy to compare its effectiveness and applicability. The evaluation concerned evaluating the normal behavior model of a healthy wind turbine with that of a target faulty turbine, and the results showed that faults could be detected two weeks prior to occurrence.

2.2. Related works on dataset

A study [18] proposed a new approach to predicting anomalies in wind turbines by combining LSTM and XGBoost models. The model was trained on a labeled dataset (LDT dataset) and then transferred to an unlabeled dataset (Engie dataset) to detect anomalies for wind farm operators who lack access to historical data. Another study [29] developed a system for signal reconstruction from low correlated parameters when SCADA sensors fail to send data. The goal of the model was to predict wind power using other SCADA parameters. The study explored linear and non-linear algorithms and used multiple linear regression, random forest, and Cartesian genetic programming evolved Artificial Neural Network (CGPANN) to inform the generalized model. In a different study[22], the Gaussian mixed model was used to cluster operating conditions of mechanical equipment to detect anomalies without mixing them up with normal operating conditions. The isolation forest method was used to identify critical attributes responsible for equipment degradation.

One another study [30] applied the improved dragonfly algorithm (IDA) to choose optimal parameters of support vector machine (SVM) for short-term wind power forecasting. The hybrid model (IDA-SVM) outperformed the traditional grid search algorithm (Grid-SVM). These studies demonstrate effective methods for condition monitoring and wind power forecasting. The IDA-SVM model utilized adaptive learning factors and differential evolution strategies to enhance the optimization ability of the dragon algorithm (DA) and was employed on the ENGIE dataset during various seasons.

Lastly the study [31], evaluated the k-means-based Smoothing Spline hybrid model using the ENGIE dataset as a validation set and demonstrated that it provides the most accurate power curve based on better goodness of fit statistics compared to other k-medoids++ -based Gaussian hybrid models.

3. Methodology

This study aims to develop a reliable workflow using XGBoost for detecting faults in wind turbines. The main objective is to create a predictive maintenance system that can identify potential issues without prior knowledge of what those faults might look like, in the absence of failure data. Previous research has explored predictive maintenance for wind turbines using machine learning and SPC based on SCADA data, as discussed in section 2. However, these studies have typically relied on available failure data, maintenance logs, alarm logs, or status logs from the same wind farm. In contrast, our study will validate our model's predictive performance using data from a different wind farm with failure data, utilizing transfer learning. Our approach aims to assess the effectiveness of our methodology in predicting failures when there is no historical data available for the wind farm under investigation.

Main steps of our approach are as below:

1. Data acquisition and pre-processing: We gather data from open-source platforms, perform data cleaning, remove outliers, and filter out normal operational data points for further model processing.

2. Model processing: We build models for each turbine in the wind farm to represent normal behavior.
3. Post-processing: We evaluate the deviations between the model predictions and actual measured data using the z-score threshold chart.

Overall, our study aims to develop a robust workflow for fault detection in wind turbines that can be applied even in the absence of failure data.

To develop our fault detection model, we first create a representation of the wind turbines' normal behavior under healthy conditions. We assume that this model will always provide accurate information about the turbine's health status. During the testing phase, we use the model to predict the wind turbine's health status. This healthy reference state will serve as a benchmark for asset managers. When new SCADA data is acquired, we compare the deviations between the healthy wind turbine model and the latest data. We monitor these deviations using a Z-score chart, and any data points outside the allowable fault threshold are considered anomalous. To validate our approach, we train the model on new data from a different wind turbine that has failure data. Only after successful validation, do we deem the model ready for real-time monitoring.

In summary, our approach involves building a reliable fault detection model based on a representation of normal turbine behavior, which can be used to predict the turbine's health status. The Z-score chart is used to monitor deviations between the model and real-time data, and any anomalies are flagged for further investigation. Successful validation using data from a different wind turbine with failure data is necessary before deploying the model for real-time monitoring.

3.1. Data acquisition and data pre-processing

To develop a reliable model for wind turbines, historical monitoring data was collected from the La Haute Borne wind farm in Meuse, France. The data was obtained from the wind farm's SCADA system, which provides real-time monitoring and control information. The wind farm is operated by ENGIE Green and has four wind turbines based on Senvion MM82 technology. The data includes 34 parameters, such as average, maximum, minimum, and standard deviation, sampled at a frequency of 10 minutes. In this study, only the average values were used as they contain the most significant information. The wind turbines at La Haute Borne are characterized by a power of 2050kW and a rotor diameter of 82m and a hub height of 80m. The study considers several key parameters, including active power, wind speed, outdoor temperature, generator bearing temperature, gearbox bearing temperature, generator speed, gearbox oil sump temperature, rotor speed, and nacelle temperature. The cut-in wind speed is 3. 5m/s, the rated wind speed is 14.5m/s, and the cut-out wind speed is 25m/s.

3.2. Feature selection

To determine the healthy behavior of a wind turbine, it's crucial to identify the input and output variables. However, it can be challenging to determine

these variables since the SCADA system's sensors measure numerous parameters. Therefore, this study relied on a literature review to identify the best combinations of variables needed to monitor critical components, such as the gearbox and generator. The bibliographic search involved several methods to arrive at a list of the most influential variables. Table 1 displays the input and output variables defining the behavior of the components of interest based on the scientific literature review.

Table 1: Input and output variables for gearbox and generator components for various models

Component	Gearbox	Generator
Input variables	Nacelle temperature Rotor Speed Active Power Outdoor temperature Gearbox oil sump temperature	Nacelle temperature Active Power Generator stator temperature Generator speed
Output variables	Gearbox Bearing Temperature	Generator Bearing Temperature
Ref	[31,33,38]	[11,20,35]

3.3. Regression Based model

The performance of wind turbines is influenced by a variety of factors, including the stochastic nature of wind. To effectively model these turbines, an algorithm must be able to accurately capture the complex relationship between the variables that define the system's behavior. In this study, we will analyze the use of regression models to build a healthy behavior profile for WT components using input and output variables. To accomplish this, we will divide the dataset instances into training and testing sets, with a 70:30 split for each component model. Model accuracy on the training set will be compared to that of the test set to identify any instances of overfitting.

The input variables used in this study have different dimensions and ranges, making it necessary to standardize their values within a defined range. The sklearn standard scalar function was used to standardize the input variables. This function calculates the z-score of the variables using the mean and standard deviation of the output variables, transforming them into z-scores as shown in equation 1.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the input variable, μ is the mean of the input variable, σ is the standard deviation of the input variable, and z is the transformed variable (i.e. the Z-score).

In this study, the input variables of the training set will be utilized to predict the output variables belonging to the same set, analyzing their interrelationships. The training accuracy is defined by how well the model forecasts the output variable. Subsequently, the model will be tested on the input variables

from an unseen test set to predict the output variable. The model's accuracy will then be evaluated based on how accurately it predicts the output variable. The predicted output variables, which represent the healthy state of the WT, will then be compared to the actual measured values. We will begin with a basic model, which involves using multiple linear regression (MLR), decision tree regression, and random forest regression. Later on, we will compare the results obtained from the linear model with non-linear algorithm, namely, extreme gradient boosting (XGBoost).

3.3.1. Multi linear regression

Multiple Linear Regression (MLR) is a statistical model used to analyze the relationship between two or more input variables (predictors) and a single output variable response [32]. It is a linear approach where the output variable is represented as a linear combination of the input variables using regression coefficients [33]. The MLR model can be expressed by the following equation:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$

where y is the output variable, x_i are the input variables, β_i are the regression coefficients, β_0 is the intercept, p is the number of input variables, and ϵ is the error term. The goal of the MLR model is to find the best set of regression coefficients that limit the sum of squared errors (SSE) among the predicted and actual values of the output variable [33].

The MLR model utilizes the method of Ordinary Least Squares (OLS) which estimates the regression coefficients by calculating the partial derivative of SSE with respect to each regression coefficient and setting it to zero which minimizes the SSE[33]. The solution for the regression coefficients can be represented by the following equation:

$\hat{\beta} = (X^T X)^{-1} X^T y$

where $\hat{\beta}$ is the vector of estimated regression coefficients, X is the matrix of input variables, y is the vector of output variables, and $(X^T X)^{-1}$ is the inverse of the matrix $X^T X$.

The MLR model is a simple and efficient method to model the relationship among a couple of input features and a single output feature. However, it assumes that the connection among the input and output variables is linear, and that there is no multicollinearity between the input variables [33].

3.3.2. Decision Tree regression

Decision tree regression is a machine learning algorithm that works by recursively splitting the dataset into subsets, using a set of decision rules based on the input features, to predict the output variable [34]. The algorithm tries to find the best decision rule that splits the dataset in a way that minimizes the variance of the output variable inside every subset.

The decision tree is constructed in a top-down manner, where the algorithm starts with the entire dataset and recursively splits it into smaller subsets based on the values of the input features. The splitting process is achieved in a way that maximizes the homogeneity of the output variable within each subset. The

Homogeneity is normally measured with a metric inclusive of mean squared error (MSE) or mean absolute error (MAE)[35].

The decision tree continues constructing until criterion is match for example a maximum tree depth or a minimum number of samples required to split a node. The prediction for a given input instance is the average value of the output variable for all the training instances that fall within the same leaf node navigating the completed tree starting from root to leaf node corresponding to input instance[36].

The decision tree regression model can be represented mathematically as follows:

1. Given a dataset D = (x1, y1), (x2, y2), ..., (xn, yn), where x_i is the input feature vector and y_i is the corresponding output variable, the goal is to learn a decision tree that can predict the output variable y for a new input feature vector x.
2. The construction of a decision tree involves iteratively dividing the dataset into smaller subsets, based on the input feature values, utilizing a predetermined set of decision rules.
3. Each internal node of the tree represents a decision rule, and each leaf node represents a prediction for the output variable.
4. The decision tree is constructed in a top-down manner, where the algorithm starts with the entire dataset and recursively splits it into smaller subsets primarily based on the values of the input features.
5. The splitting process is done in a way that maximizes the homogeneity of the output variable within each subset, as measured by a metric such as MSE or MAE.
6. The decision tree is built until a stopping criterion is match, like a maximum tree depth or a minimum number of samples required to divide a node.
7. Once the tree construction is completed, it can predict the output variable for new input instances by navigating the tree from the root node to the leaf node corresponding to the input instance.
8. The prediction for a given input instance is the average value of the output variable for all the training instances that fall within the same leaf node.

Overall, the decision tree regression model is a powerful and interpretable machine learning algorithm that can be used to predict the output variable based on a set of input features.

3.3.3. Random forest regression

Random forest regression is a type of technique that constructs many decision trees by randomly selecting samples and features from the dataset and combines their predictions to produce a more accurate output[37].

The random forest algorithm follows the following steps:

1. Randomly choose 'k' features from a pool of 'm' features, where 'k' is significantly smaller than 'm'.

2. Determine the node D by evaluating the best split point among the selected 'k' features. Randomly select 'k' features from a set of 'm' features, where 'k' is much smaller than 'm'. Identify the node D by calculating the optimal split point from the chosen 'k' features.
3. Divide the node into sub nodes based on the optimal split.
4. Iterate through steps 1 to 3 until either the number of nodes surpasses the minimum threshold or no further enhancements can be achieved.
5. Construct the forest by repeating steps 1 to 4 for a total of "n" iterations, resulting in the creation of "n" individual trees.

The output of the random forest model is calculated by taking the average of the outputs of all the decision trees [38].

The formula for calculating the output of a random forest model is as follows:

$$y = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{2}$$

where 'y' is the predicted output variable, 'n' is the total number of decision trees, and 'fi' is the output of the i-th decision tree.

In summary, the random forest regression model is a powerful and flexible machine learning technique that can predict complex relationships between variables in a given dataset.

3.3.4. XGBoost regression

XGBoost is a powerful and widely-used machine learning algorithm based on the gradient boosting machine learning method, especially for predictive modeling tasks that require high performance found on idea of iteratively training weak models and combining them into a strong one[39]. It is widely used for supervised learning problems inclusive of regression and classification. It has received huge recognition in latest years.

XGBoost employs an ensemble of decision trees to make predictions. It uses a technique called boosting to iteratively improve the performance of the decision trees. At each iteration, the algorithm adds a new decision tree that tries to accurate the mistakes of the preceding trees. The final output is aggregation of the predictions of all the trees in the ensemble. Overall the algorithm works by creating an ensemble of decision trees, where each tree is built to correct the errors of the previous tree. XGBoost also incorporates regularization techniques to prevent over-fitting and improve the model's generalization performance[40].

The algorithm is optimized for speed and performance by implementing parallel processing, tree-pruning, and caching features. The objective function used in XGBoost is the aggregation of the loss function and a regularization term. The loss function calculates the difference between the predicted and actual values, while the regularization term prevents over-fitting. The regularization term is a penalty on the complexity of the model, calculated as the sum of the squares of the model parameters[39].

The XGBoost algorithm can be formulated as follows:

1. Given a training set of N examples (x1,y1), (x2,y2), ..., (xN,yN), where xi is a vector of input features, and yi is the corresponding output value.
2. Initialize the prediction value for each example to zero.
3. For each iteration m = 1,2,...,M, do the following:
   (a) Compute the negative gradient of the loss function with respect to the predicted values.
   (b) Train a decision tree on the negative gradient values as the target variable, with xi as the input variables.
   (c) Add the new decision tree to the ensemble by combining it with the previous trees.
   (d) Update the prediction value for each example by adding the contribution of the new decision tree.
4. The final prediction value for each example is the sum of the predictions of all the decision trees.

The XGBoost algorithm has several hyperparameters that can be tuned to improve its performance, such as the learning rate, number of iterations, maximum depth of the trees, and regularization parameters. It has been shown to achieve state-of-the-art performance on many benchmark datasets and is widely used in various applications, including predictive modeling, anomaly detection, and natural language processing[39].

3.4. Evaluation metrics of models

In this study, we utilized four metrics to evaluate the effectiveness of the temperature predictive regression models discussed in section3. These metrics are: coefficient of determination (R-Squared), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The corresponding formulas for these metrics are shown below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{6}$$

Here,n is the number of observations,y_i represents the measured value, $\hat{y}_i$ represents the predicted value, m represents the number of instances of data in the test set, and $\bar{y}$ represents the mean of the measured value.

R2, also known as the goodness of fit, defines the degree to which the regression model fits the observed values. A value almost 1 states a better fit, while

a value almost 0 states a poor fit. The RMSE, MAE, and MAPE metrics, on the other hand, indicate the accuracy of the prediction model. Smaller values for these metrics imply higher accuracy.

While RMSE is sensitive to errors, MAE is more robust to outliers. However, MAPE cannot handle extremely small observed values close to zero or zero, but RMSE and MAE can handle such. Therefore, we selected RMSE, MAE, and MAPE, in addition to R2, to complement each other based on their strengths and weaknesses.

3.5. Post processing

After training and evaluating our model, we proceeded to utilize the mean and standard deviation of the output features from the training phase, specifically the Generator bearing temperature and Gearbox bearing temperature. This information was used to calculate the Z-score for the predicted temperature values generated by our model. By applying this approach, we were able to effectively assess and identify any outliers present in the Z-scores of our model's predictions.

To visualize these outliers and detect abnormal behavior, we presented various charts displaying the predicted Z-scores. We manually applied a predefined range to identify data points that fell beyond the fault threshold or control limits. Any shifts in the average were also taken into consideration. These outliers or signals beyond the control limits were considered indicative of abnormal behavior.

The model's effectiveness in detecting faults in wind turbines without failure data was further validated using a different wind turbine dataset that included maintenance logs. This validation process allowed us to uncover real faults and assess the model's ability to accurately identify incidents and anomalies.

By leveraging the Z-score calculations and analyzing the predicted values, we were able to gain insights into the abnormal behavior of the system. This post-processing approach provided valuable information for fault detection and helped in identifying potential maintenance or operational issues in wind turbines.

4. Results

In this section, we will present the practical implementation of our proposed methodology using two distinct wind farms as case studies: the La Haute Borne wind farm in Meuse, France, operated by ENGIE [41], and a wind farm managed by EDP (Energias de Portugal) in the West African Gulf of Guinea [42][43]. The La Haute Borne dataset does not include any records of failures or maintenance logs. In contrast, the EDP dataset contains both operational data and documented instances of failures. The objective of our study is to demonstrate the effectiveness of our prediction model, even in scenarios where failure data is unavailable in a wind farm.

Our study aims to demonstrate the effectiveness of our prediction model even in the absence of failure data in a wind farm. This is particularly relevant

for newly installed wind turbine's running for a short period. In such cases, we can leverage the short period of operational data to predict when the WT is likely to fail using the algorithm proposed in this paper.

We begin by showcasing the results on the ENGIE dataset, followed by the validation of our proposed methodology on EDP data, which contains maintenance data, to demonstrate the algorithm's ability to detect faults.

1. Engie Dataset
   (a) Data cleaning
   The data set available for analysis consists of SCADA data recorded every 10 minutes between January 1, 2017, and January 11, 2018, for four turbines. There are 136 sample variables, and 34 unique parameters were recorded, with basic statistics such as minimum, maximum, mean, and standard deviation. To ensure the accuracy of the data, variables with minimum, maximum, and standard deviation were removed since the average values captured the most relevant information. Any variables with NaN values were also removed because filling these values could result in misleading wind turbine conditions. The data cleaning subsection of Section3 was followed, and instances with missing input or output variables were excluded. After this step, the clean data was ready for model training, and the input parameters were extracted to construct the input dataset. The variables making up the model for each wind turbine component were selected based on Table 1. The chosen output variables for the wind turbine models of its components (gearbox and generators) across the four turbines in the training phase are shown graphically in Figures 1 and 2.
   Since there were no maintenance records available, it was assumed that the turbine was in normal condition throughout its operation. Therefore, all the data for each of the eight models were selected to analyze their behavior in the training phase. The input dataset was standardized as described in Section III, and the entire dataset was split into a training set and a testing set. The first 70% of the data were used for training, and the last 30% were reserved for testing to prevent data shuffling, which could result in data leakage since the dataset was composed of time series.
   (b) Model processing
   To ensure the development of a reliable fault detection model, all the algorithms discussed in Section III were employed. This included Multi Linear Regression, Decision Tree, Random Forest, and XGBoost algorithms for each regression model. The selection of the best model was based on a combination of performance metrics outlined in Section III.
   The performance of the algorithms varied; in most of the models, XGBoost outperformed others. The results indicated in Table 2 and Table 3 that MLR had the poorest performance in all eight models, suggesting that the relationship between the variables was non-linear.

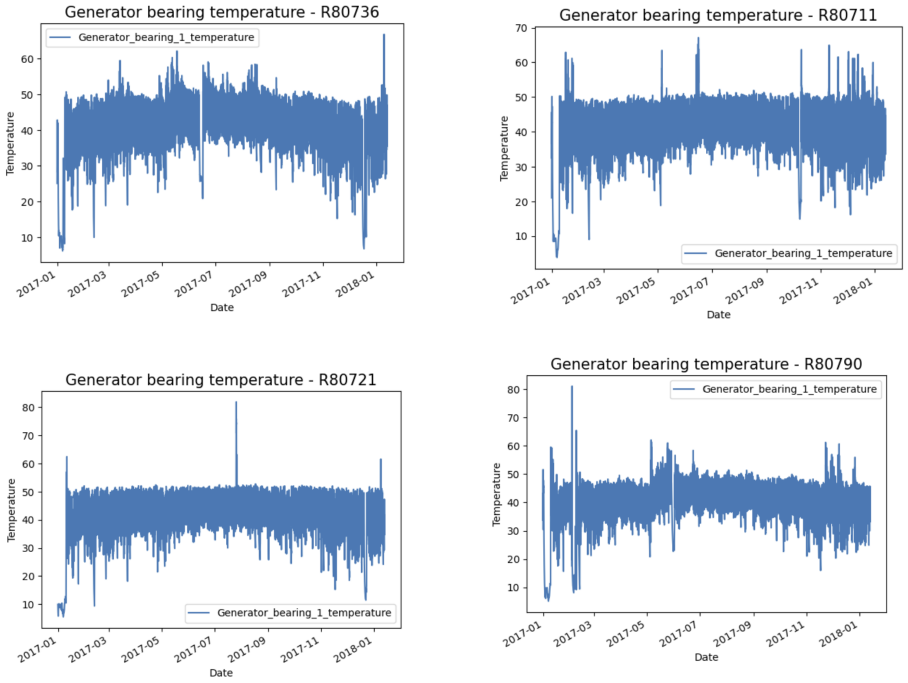

Figure 1: Engie turbines - Generator bearing temperature

Therefore, using multiple algorithms helped to enhance the accuracy and robustness of the fault detection model, enabling the detection of faults with a high level of precision.

i. Generator model

In this study,we tested four wind turbines applying four different machine learning models, and their performance metrics were recorded and presented in Table 2.

A. Wind Turbine R80736

The results showed that XGBoost and Random Forest performed the best, achieving MSE values of 1.83 and 1.86 respectively, indicating higher accuracy. The Decision Tree algorithm also performed well with an MSE of 3.50. MLR had the lowest accuracy with an MSE of 3.65. The models' predictions were relatively close to the actual values, with MAPE values ranging from 1.00 to 1.41. Considering the statistical analysis of "Generator Bearing Temperature" and its correlation with the predicted output shown in Figure 3, a left-skewed distribution was observed, highlighting the need to consider the range of -2.90 to 1.90 for outliers. These findings suggest that Random Forest and XGBoost algorithms

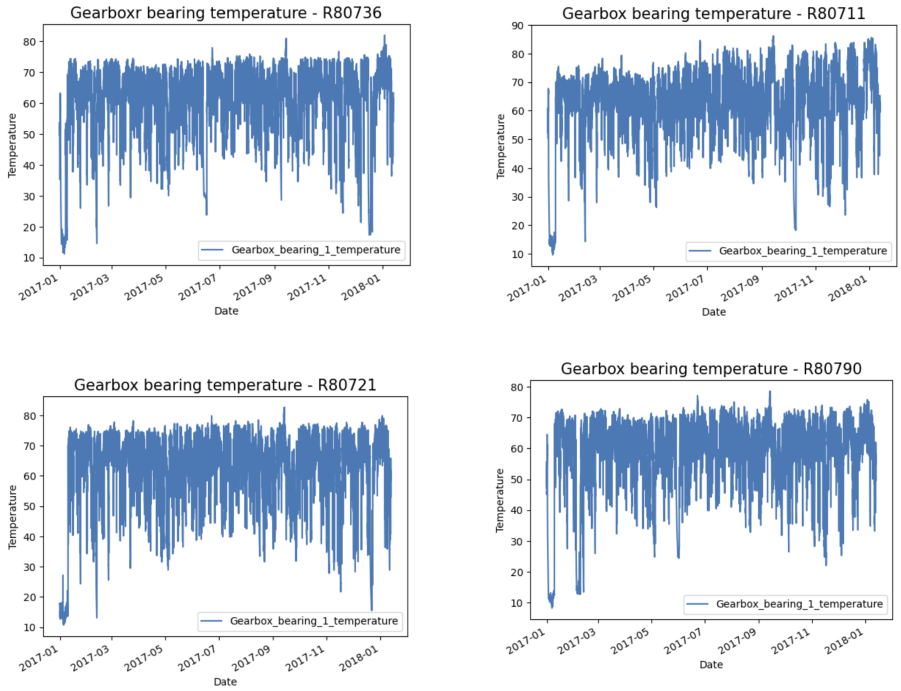

Figure 2: Engie turbines - Gearbox bearing temperature

Table 2: Model accuracy for generator bearing temperature prediction - Engie

Turbine	Model	Accuracy	MSE	RMSE	MAE	MAPE
R80736	MLR	0.89	3.65	1.91	1.18	1.27
	DT	0.90	3.50	1.87	1.02	1.41
	RF	0.94	1.86	1.36	0.76	1.00
	XG	0.95	1.83	1.35	0.76	1.03
R80721	MLR	0.90	3.51	1.87	1.19	0.94
	DT	0.93	2.44	1.56	0.83	0.76
	RF	0.96	1.35	1.16	0.63	0.54
	XG	0.96	1.33	1.15	0.64	0.54
R80711	MLR	0.83	5.41	2.33	1.33	1.17
	DT	0.82	5.51	2.35	1.05	1.01
	RF	0.90	3.03	1.74	0.78	0.71
	XG	0.90	3.02	1.74	0.79	0.74
R80790	MLR	0.88	3.93	1.98	1.14	0.93
	DT	0.87	4.38	2.09	1.00	0.87
	RF	0.93	2.39	1.55	0.74	0.64
	XG	0.93	2.39	1.55	0.74	0.62

Table 3: Model accuracy for gearbox bearing temperature prediction - Engie

Turbine	Model	Accuracy	MSE	RMSE	MAE	MAPE
R80736	MLR	0.97	1.96	1.40	0.76	0.45
	DT	0.98	1.60	1.26	0.68	0.34
	RF	0.99	0.87	0.93	0.51	0.26
	XG	0.99	0.90	0.95	0.53	0.27
R80721	MLR	0.98	2.17	1.47	0.75	0.38
	DT	0.98	1.56	1.25	0.60	0.30
	RF	0.99	0.87	0.93	0.45	0.21
	XG	0.99	0.89	0.95	0.47	0.24
R80711	MLR	0.98	1.54	1.24	0.63	0.41
	DT	0.99	1.01	1.00	0.50	0.39
	RF	0.99	0.55	0.74	0.36	0.31
	XG	0.99	0.57	0.76	0.38	0.30
R80790	MLR	0.98	1.62	1.27	0.65	0.49
	DT	0.98	1.31	1.14	0.57	0.37
	RF	0.99	0.61	0.78	0.41	0.28
	XG	0.99	0.65	0.81	0.43	0.30

are recommended for more accurate predictions in wind turbine modeling. These insights have been leveraged to identify potential dates of increased failure likelihood and recorded, as demonstrated in Figure 21.

B. Wind Turbine R80721
The performance of four models for wind turbine R80721 was evaluated and found to have high accuracy. The MSE values ranged from 1.33 to 3.51, indicating close predictions to the actual values. The XGBoost and Random Forest models performed the best with MSE values of 1.35 and 1.33 respectively, followed by Decision Tree with an MSE of 2.44, and MLR with an MSE of 3.51. In terms of accuracy and precision, Random Forest and XGBoost outperformed the other models, making them recommended choices for turbine modeling. The analysis of the input feature "Generator Bearing Temperature" and its correlation with the predicted output feature revealed a left-skewed distribution shown in Figure 4. It is important to consider outliers within the range of -2 to 1.78 to avoid incorporating erroneous data that could lead to failures. These findings have aided in identifying specific dates when failures are more likely to occur, as depicted in Figure 21

C. Wind Turbine R80711
All four algorithms used for modeling wind turbine R80711 achieved high levels of accuracy. The MSE values ranged from 3.02 to 5.41, and the MAPE values ranged from 0.74

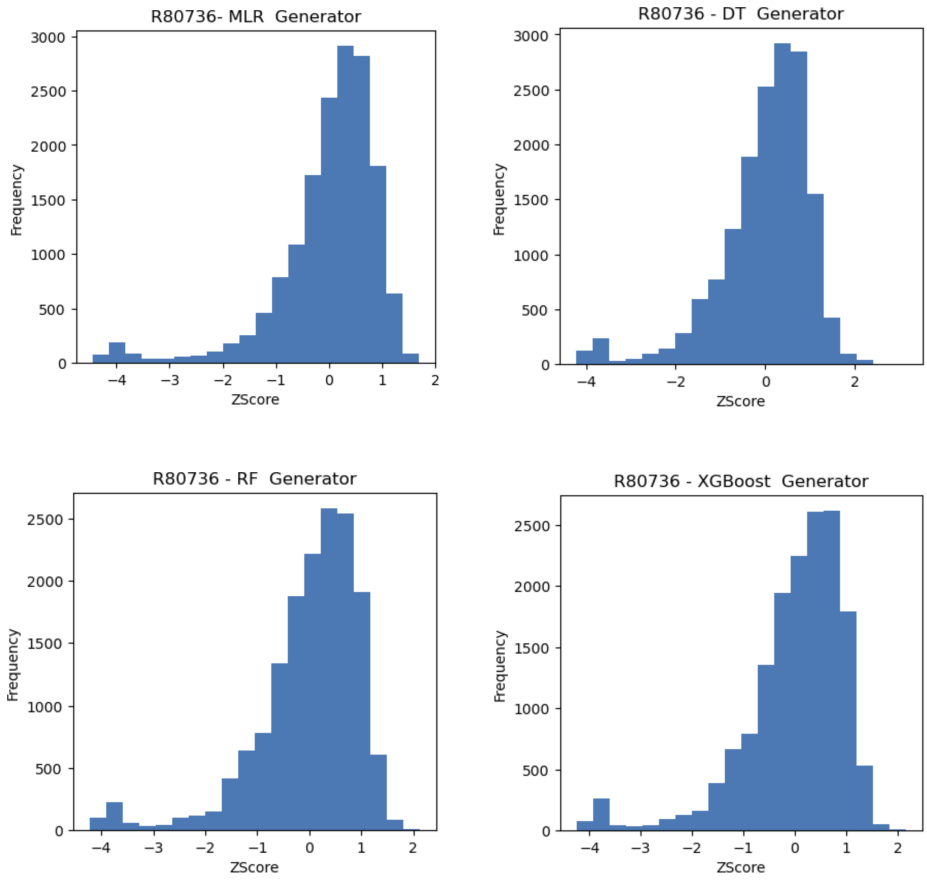

Figure 3: R80736 - ZScore of Generator bearing temperature on various models

to 1.17. The XGBoost and Random Forest algorithms performed the best, with MSE values of 3.02 and 3.03 respectively. The Decision Tree algorithm also showed good performance with an accuracy of 0.90. However, the MLR algorithm had the lowest accuracy, with an MSE of 5.41 and accuracy of 0.83. Based on these results, it is recommended to use the Random Forest or XGBoost algorithms for building turbine models due to their superior accuracy and precision. The graph in Figure 5 indicates a left-skewed distribution, suggesting the need to identify and exclude outliers within the range of -2.20 to 1.80 to prevent turbine failures. The outliers predicted dates of failures shown in Figure 21.

D. Wind Turbine R80790

RF and XG models showed the highest R2 values of 0.93,

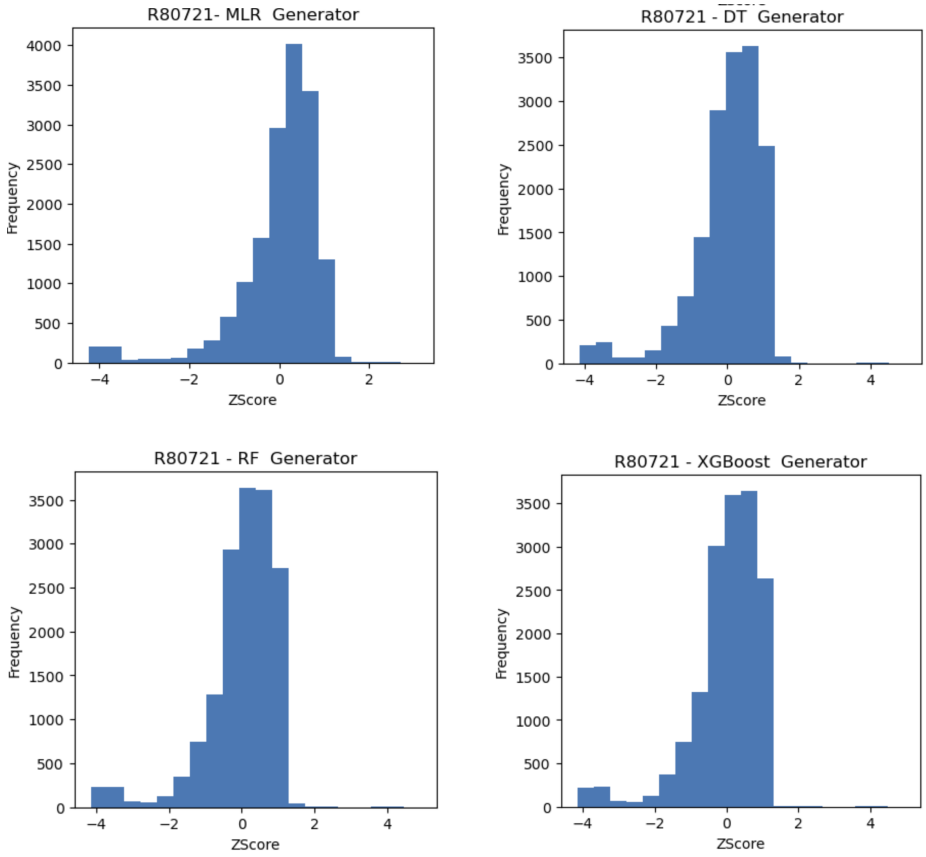

Figure 4: R80721 - ZScore of Generator bearing temperature on various models

indicating their ability to explain a large portion of the data variance. RF and XG models also outperformed in terms of MSE, MAE, RMSE, and MAPE values, exhibiting lower errors compared to MLR and DT models. The graphical representation in Figure 6 revealed a left-skewed distribution, highlighting the need to consider outliers within the range of -2 to 2 to avoid incorporating faulty data that could lead to failures. The dates of failures predicted as shown in Figure 21. Overall, RF and XG models demonstrated superior accuracy and precision in predicting the behavior of wind turbine R80790.

ii. Gearbox model

The table 3 provided presents the performance metrics of different machine learning models for the four wind turbine in gearbox model. The dates of failures for the gearbox model predicted as shown in Figure 21.

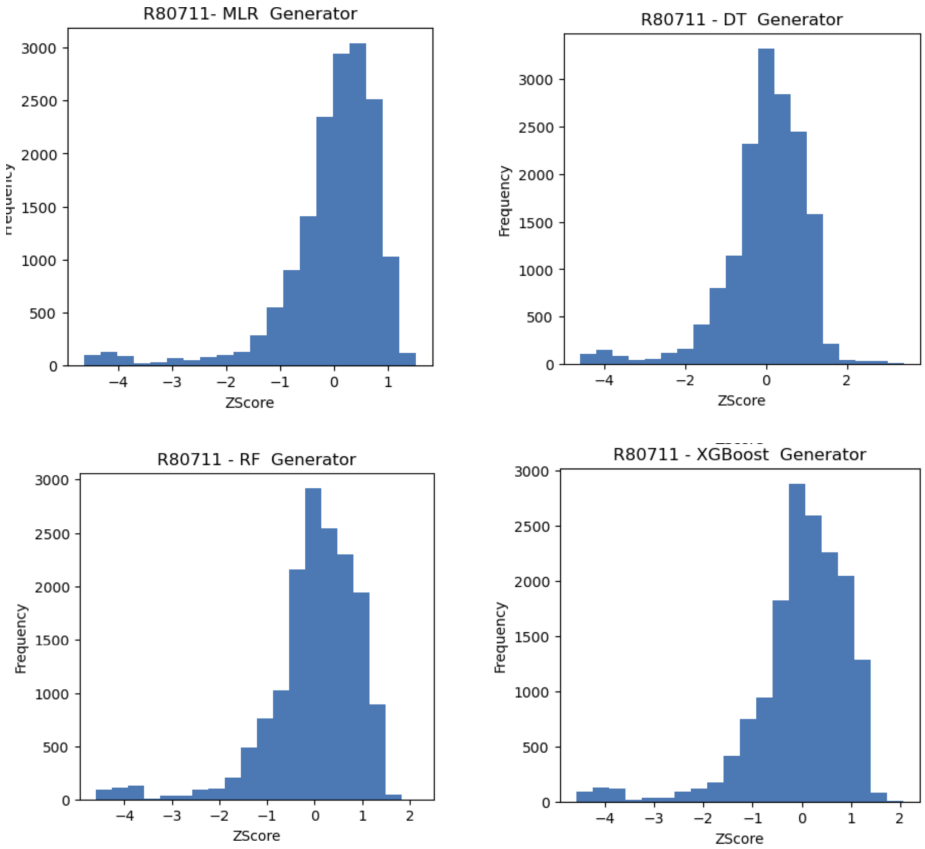

Figure 5: R80711 - ZScore of Generator bearing temperature on various models

A. Wind Turbine R80736
All four models demonstrate exceptional accuracy, with scores ranging from 0.97 to 0.99, indicating precise predictions of turbine gearbox performance. The RF and XG models outperform MLR and DT models in terms of error metrics, exhibiting lower MSE, RMSE, MAE, and MAPE values. Among them, the RF model achieves the lowest error metrics, closely followed by the XG model. Therefore, either the RF or XG model is recommended for accurate predictions of wind turbine R80736 gearbox performance. The graphical representation in Figure 7 reveals a left-skewed distribution, highlighting the importance of considering outliers within the range of -2.94 to 1.54 to avoid incorporating faulty data that could lead to failures.

B. Wind Turbine R80721
The Random Forest and XGBoost models consistently out-

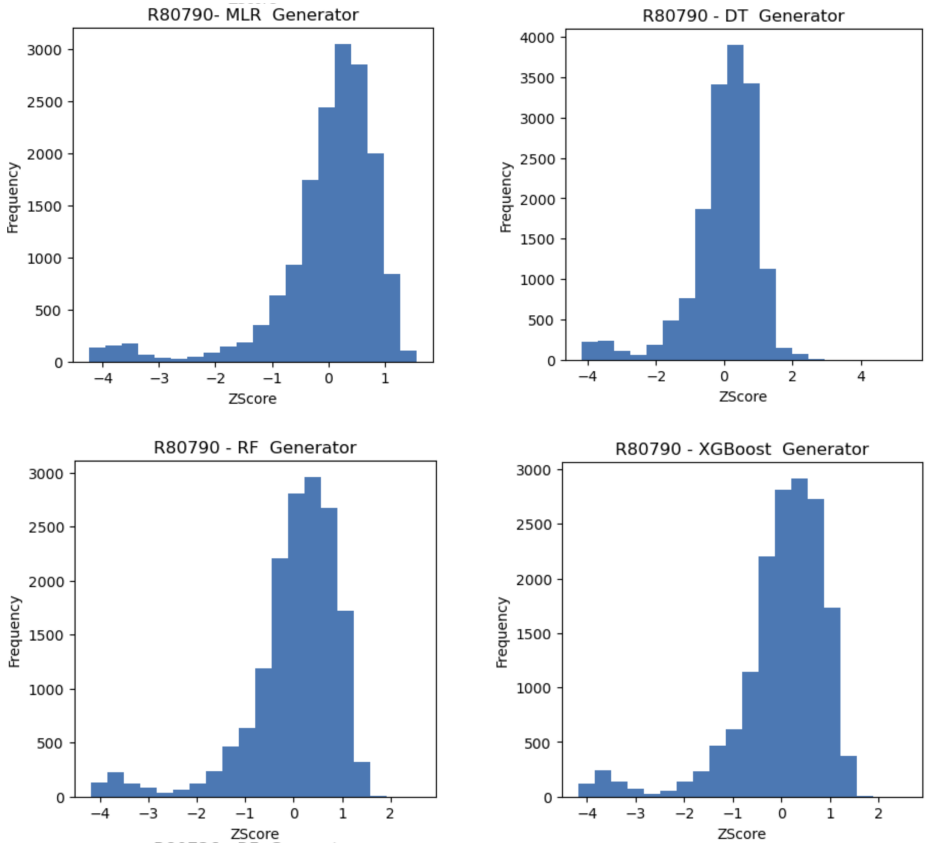

Figure 6: R80790 - ZScore of Generator bearing temperature on various models

performed the MLR and DT models, achieving lower MSE, RMSE, MAE, and MAPE values. They demonstrated higher accuracy and precision in predicting the turbine's gearbox behavior, as indicated by their high R-squared values ranging from 0.98 to 0.99. Therefore, the RF and XG models are recommended as the top-performing models for this specific turbine application. The graphical representation in Figure 8 illustrates the output ZScore of predicted temperature for the different models. The graph displays a left-skewed distribution, suggesting the need to consider outliers within the range of -2.73 to 1.54. This will help avoid incorporating faulty data that could lead to failures, allowing for the identification of potential dates when failures are likely to occur.

C. Wind Turbine R80711

The results shows that all four models have very high accu-

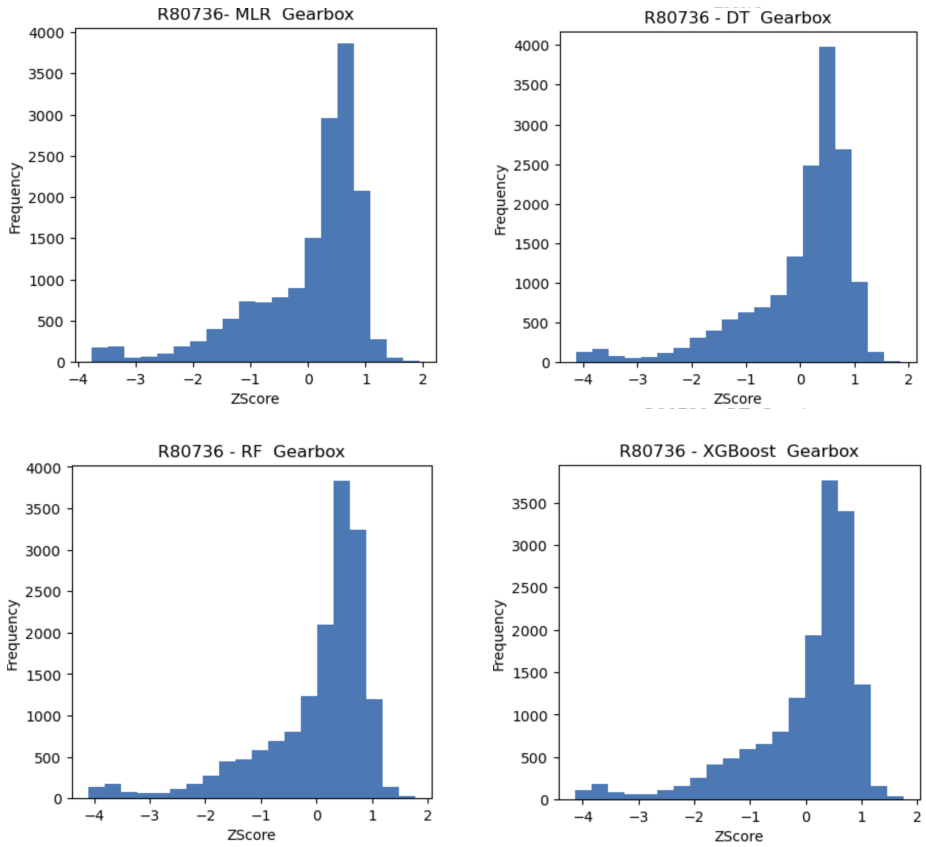

Figure 7: R80736 - ZScore of Gearbox bearing temperature on various models

racy, with all models having a coefficient of determination (R-squared) value close to 1. The Decision Tree model has the lowest MSE and RMSE values, while the Random Forest and XGBoost models have the lowest MAE and MAPE values. This suggests that the Decision Tree model may be the best model for minimizing overall prediction error, while the Random Forest and XGBoost models may be better suited for minimizing the absolute prediction error. The statistical analysis conducted on the input feature "Generator Bearing Temperature" and its correlation with the predicted output feature "Generator Bearing Temperature" for various machine learning models is presented in Figure 9. The graph depicts a left-skewed distribution, indicating that considering outliers within the range of -2.94 to 1.94 may result in the inclusion of erroneous data, potentially leading to failures. These findings can be utilized to identify potential

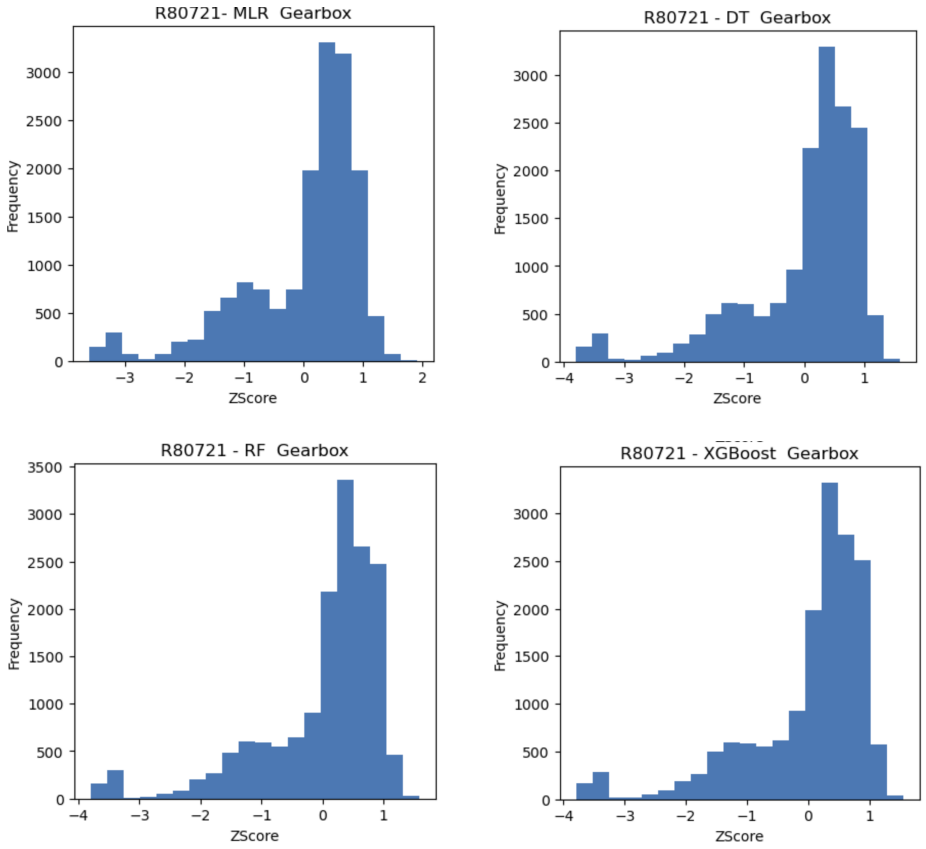

Figure 8: R80721 - ZScore of Gearbox bearing temperature on various models

dates when failures are more likely to occur.

D. Wind Turbine R80790

All four machine learning models exhibit high accuracy levels, with R-squared values ranging from 0.98 to 0.99, indicating their ability to explain a significant portion of the turbine data variance. The RF and XG models outperform the MLR and DT models, with lower MSE and RMSE values, indicating more accurate predictions. The MAE and MAPE values are also low across all models, further indicating their close proximity to the actual turbine output values. Overall, the RF and XG models demonstrate slightly higher accuracy than the MLR and DT models. Figure 10 presents the statistical analysis of the "Generator Bearing Temperature" input feature and its correlation with the predicted output feature for different machine learning models. The graph shows a left-skewed distribution, emphasizing the im-

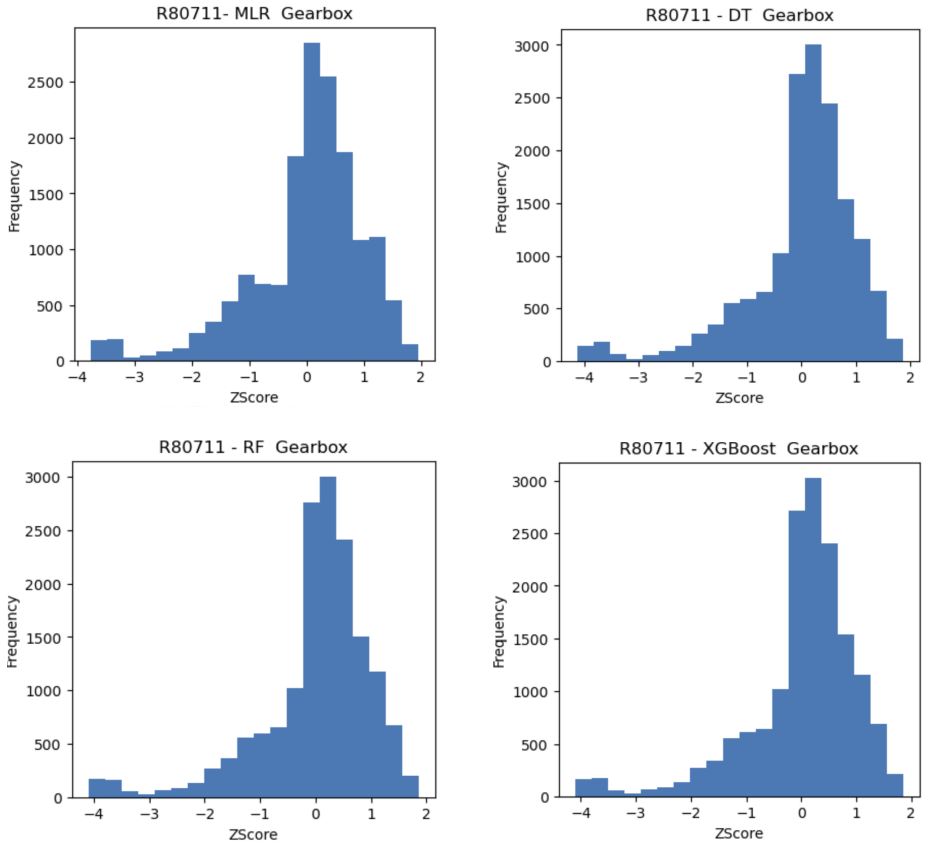

Figure 9: R80711 - ZScore of Gearbox bearing temperature on various models

portance of considering outliers within the range of -2.84 to 1.40 to avoid incorporating erroneous data that may lead to failures. These insights can help identify dates when failures are more likely to occur.

2. EDP Dataset
   (a) Data cleaning
   The historical SCADA data for the year 2017 was collected every 10 minutes, recording 83 sample variables for four turbines. Along with the SCADA data, a failure logbook for the same year was available. Before selecting a suitable dataset for training, it was necessary to analyze the failure data. It was important for the selected dataset to include all the variables (both input and output) required to define the normal operation of the wind turbines. The data cleaning process, described in the subsection of Section 3, was followed using a step-wise method. After the data was thoroughly cleaned and prepared for model training, the input parameters needed for the models were

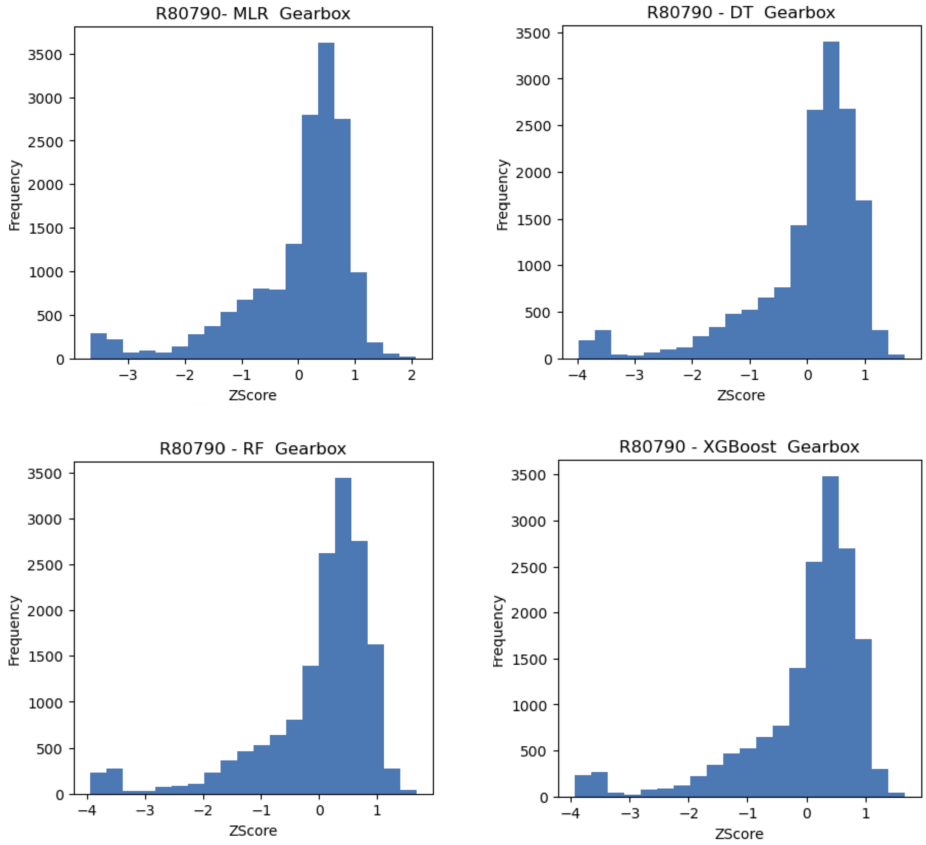

Figure 10: R80790 - ZScore of Gearbox bearing temperature on various models

extracted from the cleaned SCADA data. These extracted parameters were then used to form the input dataset for further analysis. The input variables for each component of the wind turbines (such as gearbox and generators) were selected based on Table 1. Figure 11 and 12 provides a graphical representation of the chosen output variables needed to define the models for the wind turbine components across the two turbines during the training phase.

To construct the input dataset, a set of variables necessary for each wind turbine component was selected, along with the corresponding output variable. The input dataset was standardized as discussed in Section 3. After standardization, the entire dataset of the input dataset and the output variable was split into a training set and a testing set. The data split was performed by selecting the first 70% of the data for training and the last 30% for testing. This method of data splitting ensured that there was no shuffling of the data, considering that the dataset consists of time series data. Randomly

selecting data could lead to data leakage.

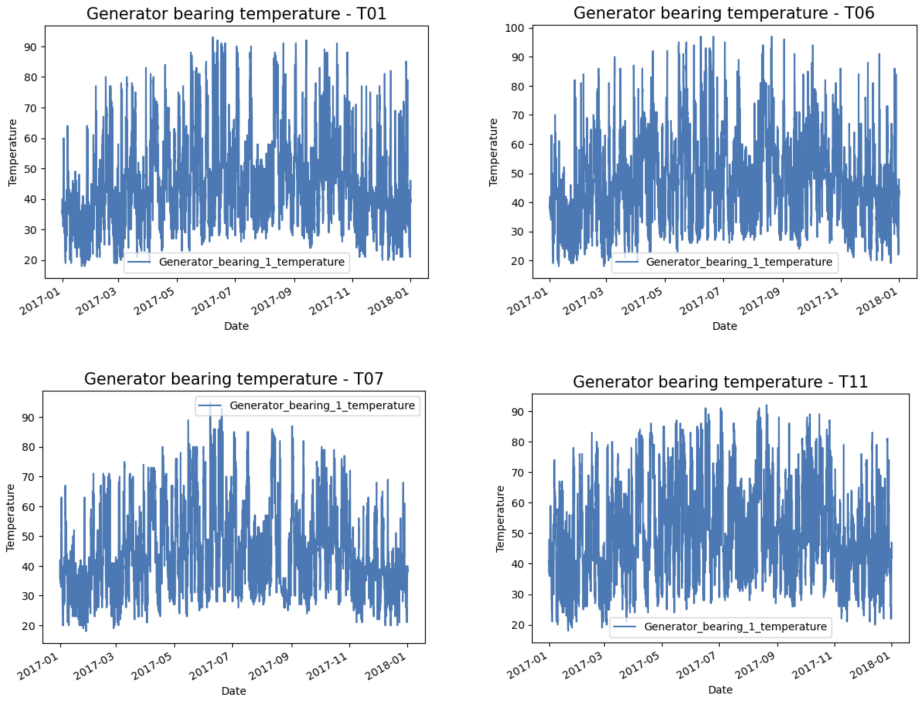

Figure 11: EDP turbines - Generator bearing temperature

Turbine	Model	Accuracy	MSE	RMSE	MAE	MAPE
T01	MLR	0.95	5.81	2.41	1.22	0.24
	DT	0.94	7.64	2.76	1.21	0.27
	RF	0.96	4.33	2.08	0.98	0.21
	XG	0.96	4.26	2.06	0.98	0.21
T06	MLR	0.95	5.65	2.38	1.05	0.92
	DT	0.92	9.21	3.03	1.16	1.04
	RF	0.95	5.23	2.29	0.94	0.78
	XG	0.96	4.86	2.21	0.90	0.70
T07	MLR	0.99	1.62	1.27	0.74	0.19
	DT	0.98	1.84	1.36	0.70	0.19
	RF	0.99	1.03	1.02	0.55	0.15
	XG	0.99	0.99	0.99	0.55	0.15
T11	MLR	0.91	10.82	3.29	1.99	0.76
	DT	0.87	16.61	4.08	2.01	0.89
	RF	0.92	9.43	3.07	1.67	0.77
	XG	0.93	8.98	3.00	1.67	0.78

Table 4: Model accuracy for generator bearing temperature prediction - EDP

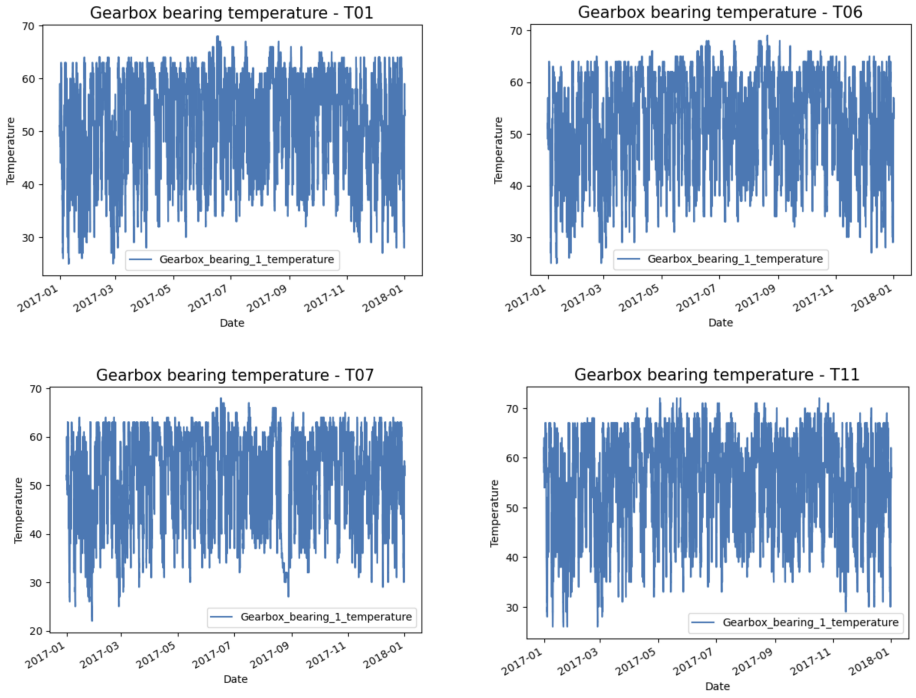

Figure 12: EDP turbines - Gearbox bearing temperature

Turbine	Model	Accuracy	MSE	RMSE	MAE	MAPE
T01	MLR	0.98	0.89	0.95	0.54	0.15
	DT	0.98	0.84	0.92	0.44	0.13
	RF	0.99	0.48	0.69	0.37	0.11
	XG	0.99	0.46	0.68	0.37	0.11
T06	MLR	0.98	0.93	0.97	0.56	0.39
	DT	0.98	0.84	0.92	0.43	0.30
	RF	0.99	0.47	0.69	0.36	0.25
	XG	0.99	0.47	0.68	0.37	0.25
T07	MLR	0.98	0.84	0.91	0.53	0.15
	DT	0.98	0.76	0.87	0.41	0.12
	RF	0.99	0.42	0.65	0.34	0.10
	XG	0.99	0.43	0.65	0.35	0.10
T11	MLR	0.98	1.01	1.00	0.59	0.22
	DT	0.98	0.89	0.94	0.47	0.18
	RF	0.99	0.52	0.72	0.39	0.15
	XG	0.99	0.51	0.71	0.39	0.15

Table 5: Model accuracy for gearbox bearing temperature prediction - EDP

(b) Model processing
The two models were built in the same manner as described in Section

3. The results indicated in Tables 4 and 5 that MLR had the poorest performance The Tables 4 and 5 provided presents the performance metrics of different machine learning models for the four wind turbine in generator and gearbox model. The dates of failures for the gearbox model predicted as shown in Figure 22.

i. Generator model

A. Wind Turbine T01

Among the machine learning algorithms, Random Forest (RF) and XGBoost (XG) show the highest accuracy with an accuracy score of 0.96. They also have the lowest MSE and RMSE values, indicating better performance in predicting the generator's behavior. The Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) values are also relatively low for RF and XG. In Figure 13, the statistical analysis performed on the input feature "Generator Bearing Temperature" and its correlation with the predicted output feature is visualized. The graph displays a right-skewed distribution, underscoring the importance of managing outliers within the range of -2.84 to 1.40. Incorporating outliers beyond this range may introduce erroneous data, potentially causing failures. Leveraging these findings, potential failure dates were identified and compared with the actual failure dates from the logs dataset which includes the actual failure date of 2017-08-11. This indicates that the fault detection algorithm successfully anticipated the generator damage before the incident and generated multiple alarms in advance.

B. Wind Turbine T06

RF and XG again perform well in terms of accuracy, with an accuracy score of 0.96. They have the lowest MSE and RMSE values, suggesting better predictive performance. The MAE and MAPE values are also lower for RF and XG compared to the other algorithms. The correlation between the input feature "Generator Bearing Temperature" and the predicted output feature is explored in the statistical analysis presented in Figure 14. The graph exhibits a right-skewed distribution, highlighting the significance of considering outliers within the range of -2.5 to 1.22. Including outliers beyond this range may introduce erroneous data, leading to failures. The analysis further identifies potential dates when failures are more likely to occuras shown in Figure 22. Among the range of dates obtained from the predictions, the actual failure date of 2017-08-19 from the logs dataset is included. This demonstrates the fault detection algorithm's ability to forecast generator damage ahead of time, issuing multiple alarms prior to the occurrence.

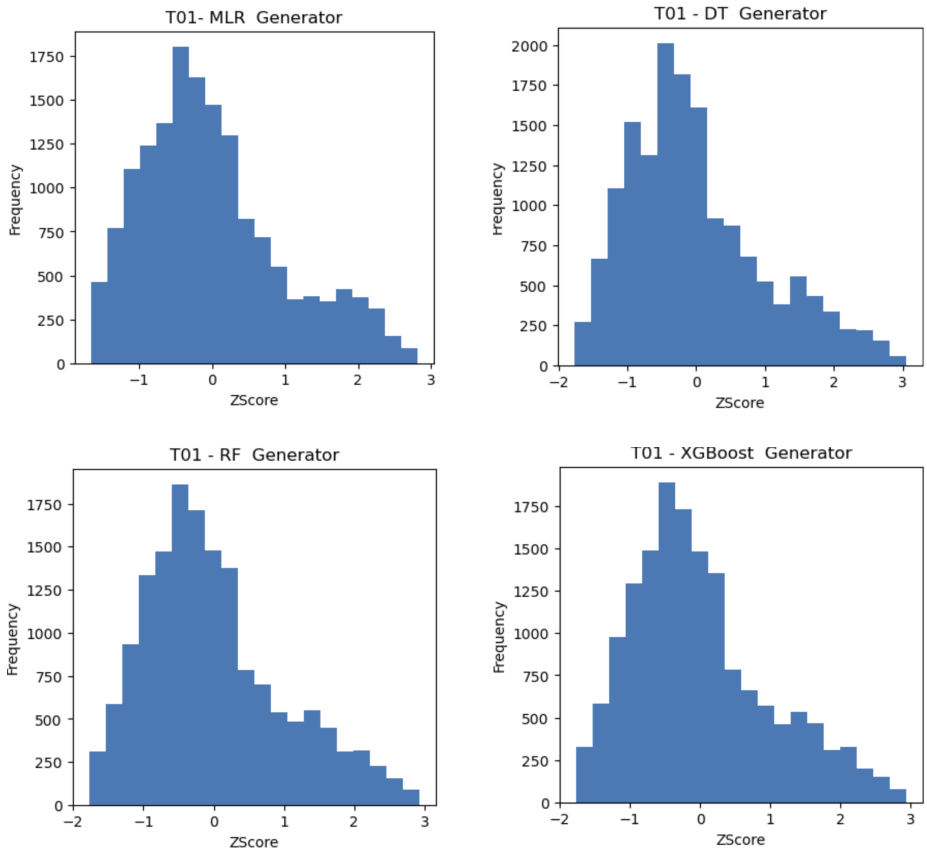

Figure 13: T01 - ZScore of Generator bearing temperature on various models

C. Wind Turbine T07

RF and XG demonstrate the highest accuracy with an accuracy score of 0.99. They have the lowest MSE, RMSE, MAE, and MAPE values, indicating superior performance in predicting the generator's behavior. Figure 15 showcases the statistical analysis conducted on the relationship between the input feature "Generator Bearing Temperature" and the predicted output feature. The graph visualizes a right-skewed distribution, emphasizing the need to account for outliers within the range of -1.86 to 1.77. Including outliers outside this range could introduce erroneous data and contribute to failures. As shown in Figure 22, the logs dataset reveals that the actual failure dates of 2017-06-17, 2017-08-20 and 2017-08-21 aligns with the range of dates provided by the predictions. This signifies the fault detection algorithm's effectiveness in anticipating generator damage well in advance,

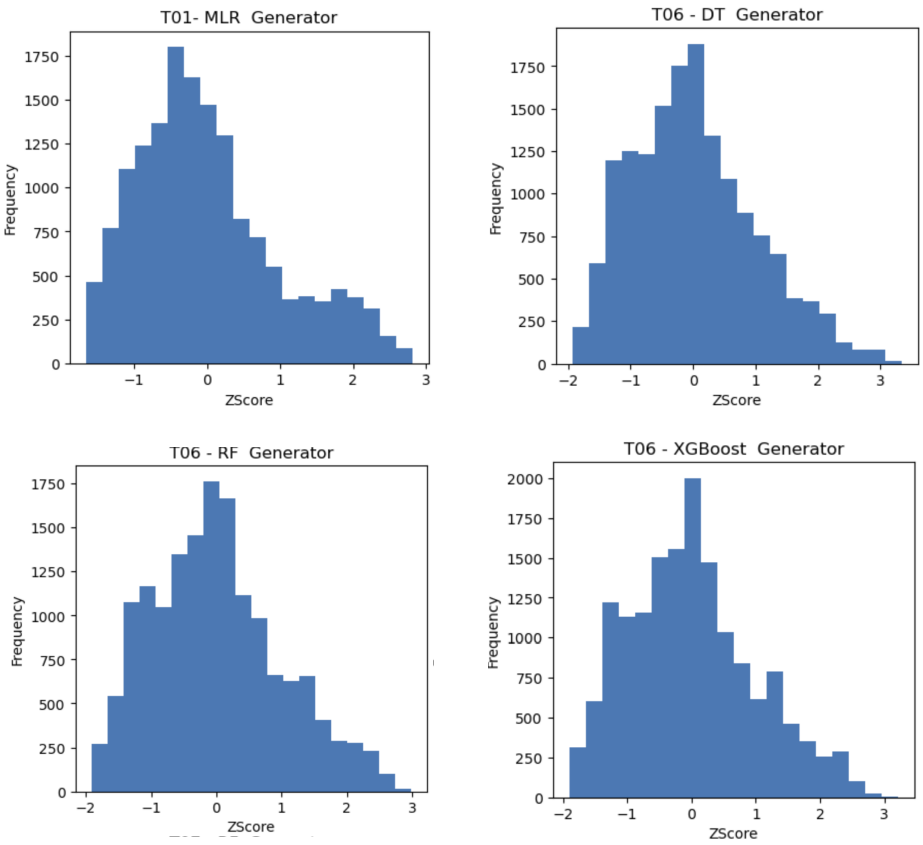

Figure 14: T06 - ZScore of Generator bearing temperature on various models

generating multiple alarms beforehand.

D. Wind Turbine T11

RF and XG exhibit higher accuracy compared to the other algorithms, with an accuracy score of 0.92 and 0.93, respectively. They also have lower MSE, RMSE, MAE, and MAPE values, indicating better performance in predicting the generator's behavior. Figure 16 showcases analysis conducted on the relationship between the input feature "Generator Bearing Temperature" and the predicted output feature. The graph visualizes a right-skewed distribution, emphasizing the need to account for outliers within the range of -1.86 to 1.77. Including outliers outside this range could introduce erroneous data and contribute to failures. As shown in Figure 22, the recorded failure date of 2017-04-26 and 2017-09-12 coincides with the range of dates derived from the predictions. This highlights the fault detection algorithm's capa-

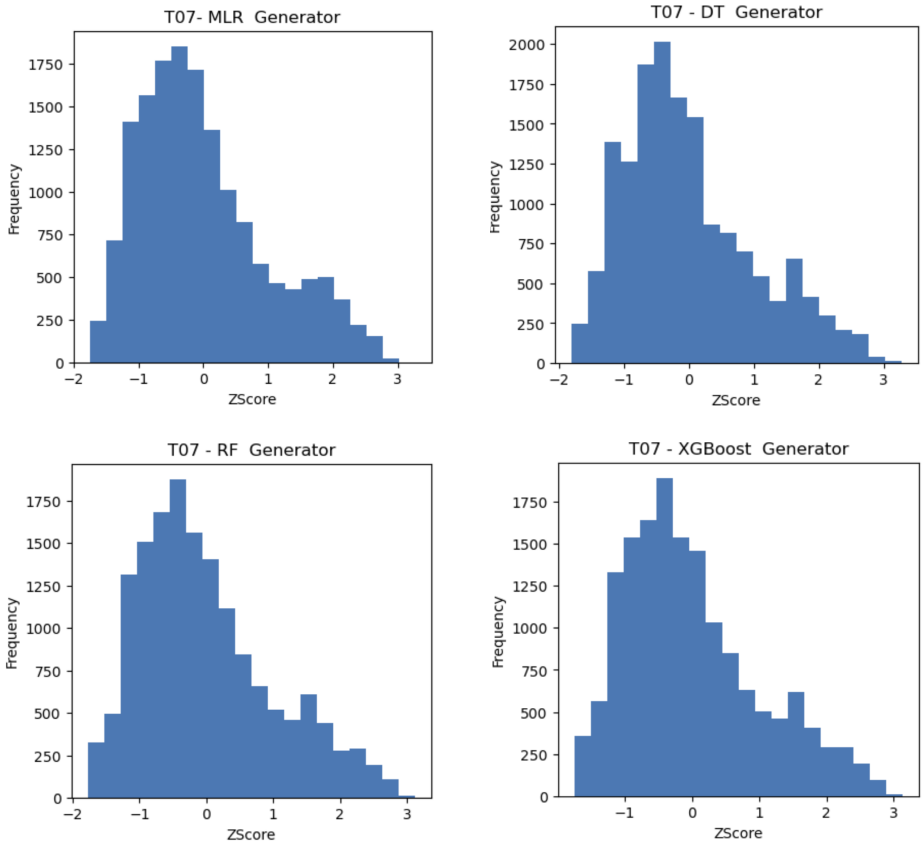

Figure 15: T07 - ZScore of Generator bearing temperature on various models

bility to predict generator damage ahead of time, issuing multiple alarms in anticipation of the incident.

ii. Gearbox model

A. Wind Turbine T01

RF and XG demonstrate the highest accuracy with an accuracy score of 0.99. They have the lowest MSE, RMSE, MAE, and MAPE values, indicating better performance in predicting the gearbox's behavior. Figure 17 illustrates the statistical analysis performed on the input feature "Gearbox Bearing Temperature" and its correlation with the predicted output feature. The graph displays a left-skewed distribution, suggesting that the range of -2.29 to 1.41 should be considered for outliers to avoid incorporating faulty data, which could lead to failures. These insights have been leveraged to identify potential dates of increased failure likelihood and compared with actual failure dates recorded in the

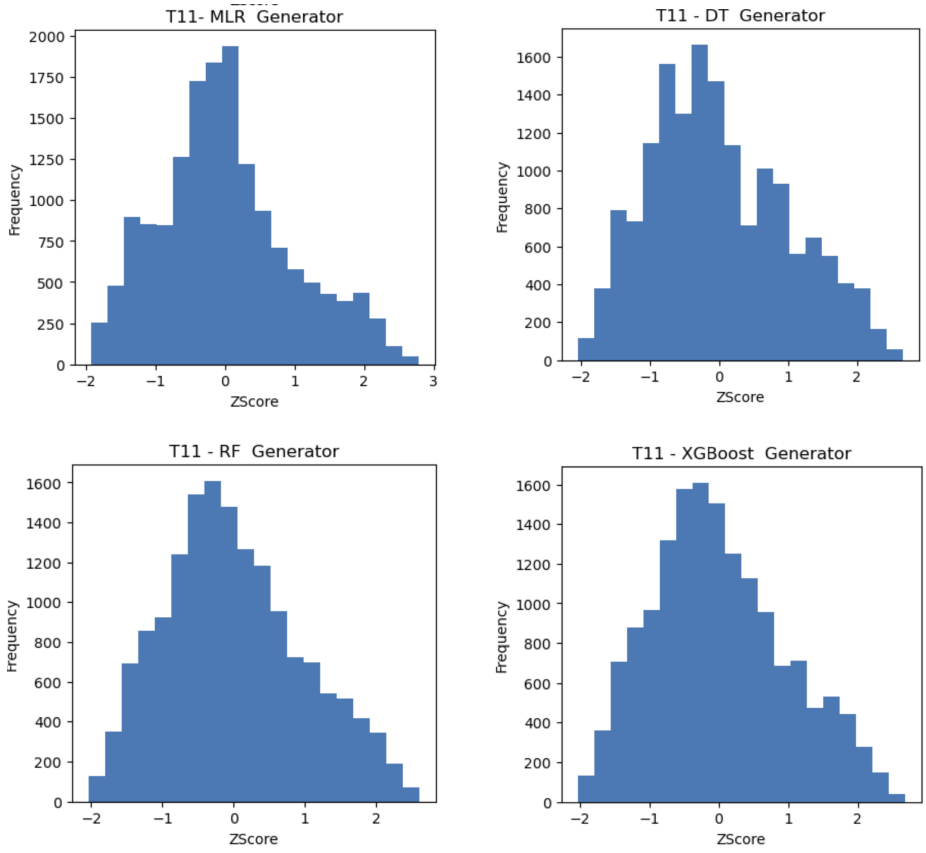

Figure 16: T11 - ZScore of Generator bearing temperature on various models

logs dataset, as demonstrated in Figure 22 including date of 2017-08-11. This indicates that the fault detection algorithm successfully forecasted the gearbox damage in advance, providing multiple alarms prior to the occurrence.

B. Wind Turbine T06

RF and XG exhibit the highest accuracy with an accuracy score of 0.99. They have the lowest MSE, RMSE, MAE, and MAPE values, indicating better performance in predicting the gearbox's behavior. The analysis in Figure 18 examines the statistical relationship between the input feature "Gearbox Bearing Temperature" and the predicted output feature. The graph showcases a left-skewed distribution, indicating that it is crucial to consider outliers within the range of -1.94 to 0.87. Including such outliers may introduce erroneous data, potentially resulting in failures. As depicted in Figure 22, the logs dataset confirms that the actual failure date

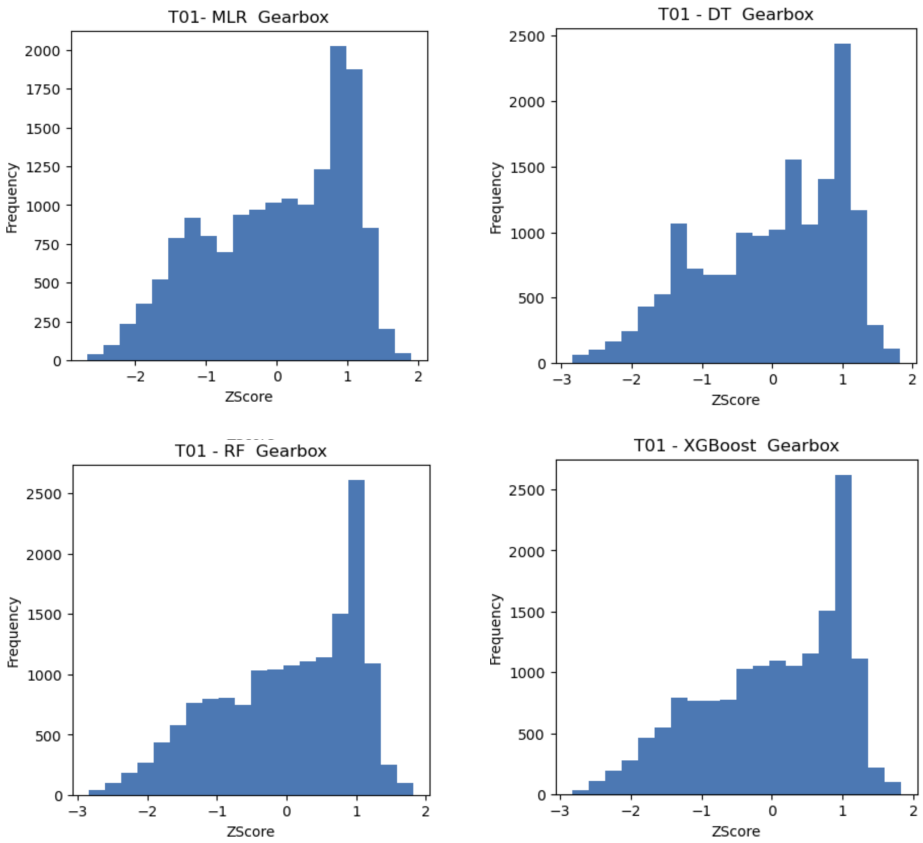

Figure 17: T01 - ZScore of Gearbox bearing temperature on various models

of 2017-08-19 and 2017-10-17 is included. This showcases the fault detection algorithm's ability to anticipate gearbox damage ahead of time, generating multiple alarms as a pre-emptive measure.

C. Wind Turbine T07

RF and XG show the highest accuracy with an accuracy score of 0.99. They have the lowest MSE, RMSE, MAE, and MAPE values, indicating better performance in predicting the gearbox's behavior. Figure 19 presents the statistical analysis carried out on the "Gearbox Bearing Temperature" input feature and its correlation with the predicted output. The graph exhibits a left-skewed distribution, underscoring the significance of accounting for outliers within the range of -2.29 to 1.41. Including outliers beyond this range could introduce faulty data, leading to failures. The findings have been instrumental in identifying dates with a higher proba-

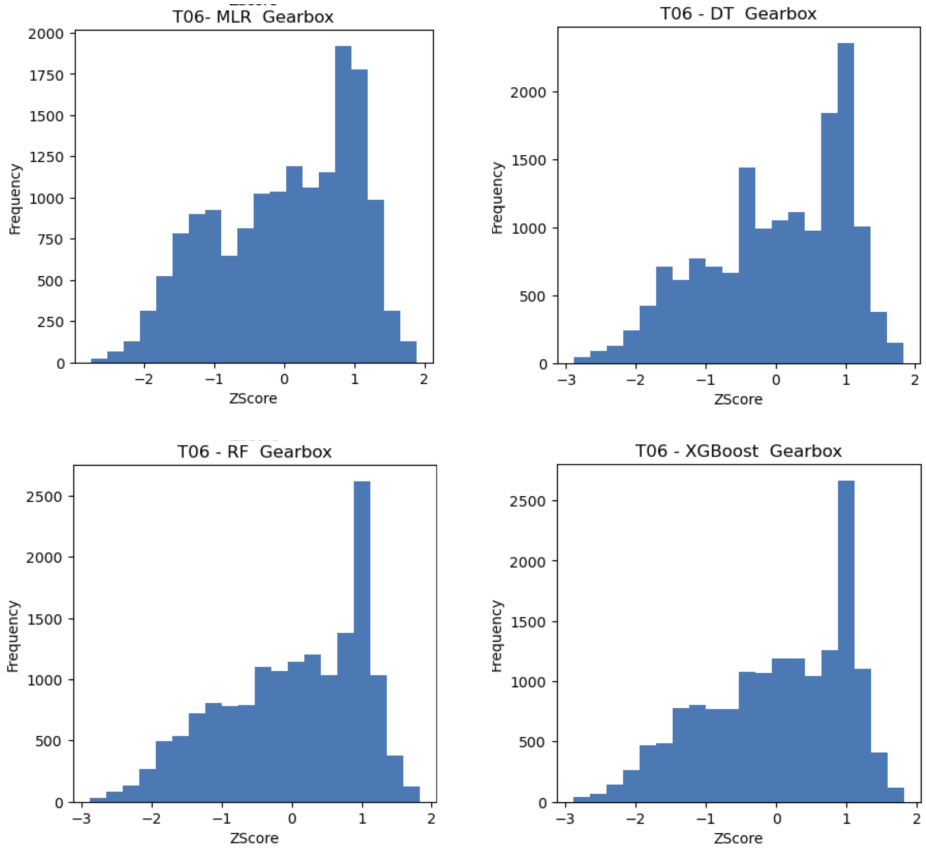

Figure 18: T06 - ZScore of Gearbox bearing temperature on various models

bility of failure occurrence, which have been compared with the actual failure dates from the logs dataset in Figure 22. The logs dataset contains the actual failure dates of 2017-06-17, 2017-08-20 and 2017-10-19, which coincides with the range of dates obtained from the predictions. This signifies the fault detection algorithm's proficiency in predicting gearbox damage well in advance, issuing multiple alarms as an early warning system.

D. Wind Turbine T11

RF and XG demonstrate the highest accuracy with an accuracy score of 0.99. They also have the lowest MSE, RMSE, MAE, and MAPE values, indicating better performance in predicting the gearbox's behavior. The statistical analysis depicted in Figure 20 focuses on the input feature "Gearbox Bearing Temperature" and its relationship with the predicted output feature. The graph showcases a left-skewed

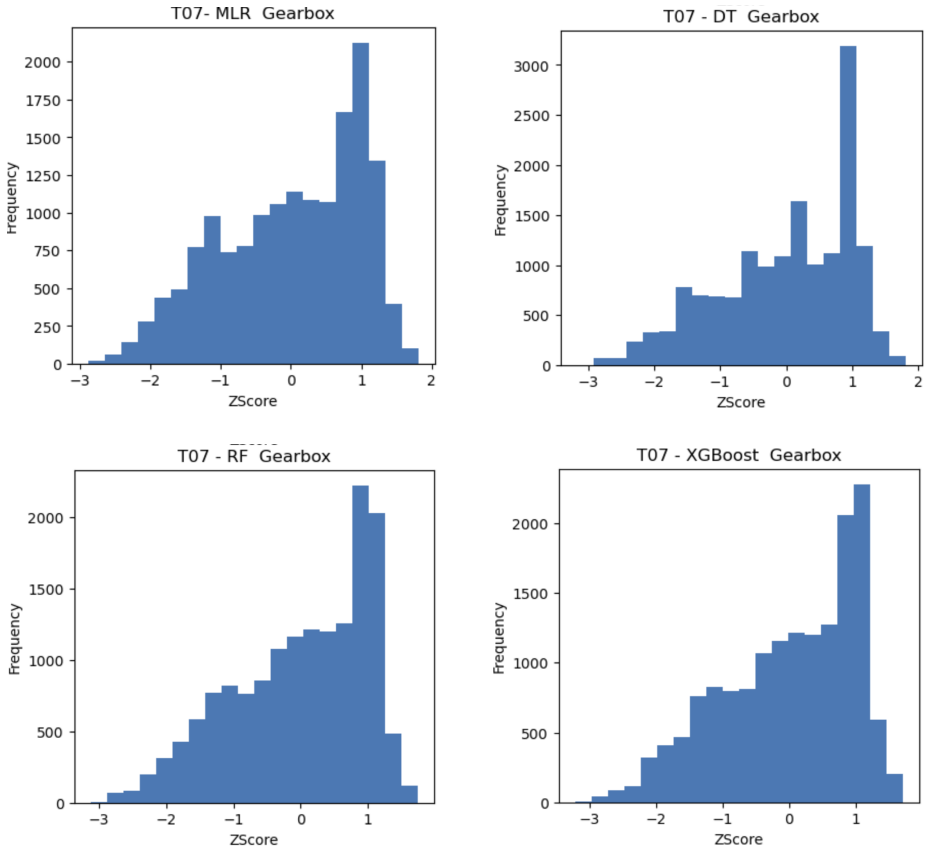

Figure 19: T07 - ZScore of Gearbox bearing temperature on various models

distribution, emphasizing the need to consider outliers within the range of -2.29 to 1.41. Including outliers outside this range may introduce erroneous data that could lead to failures. The actual failure date of 2017-04-26, extracted from the logs dataset, is among the range of dates provided by the predictions. This validates the fault detection algorithm's accuracy in anticipating gearbox damage beforehand, triggering multiple alarms prior to the incident.

5. Effectiveness of our fault detection approach

The fault detection algorithm in this study was developed through a systematic process consisting of three main steps: data acquisition and preprocessing, model processing, and post-processing. By employing machine learning algorithms such as MLR, DT, RF, and XGBoost, the cleaned data was utilized

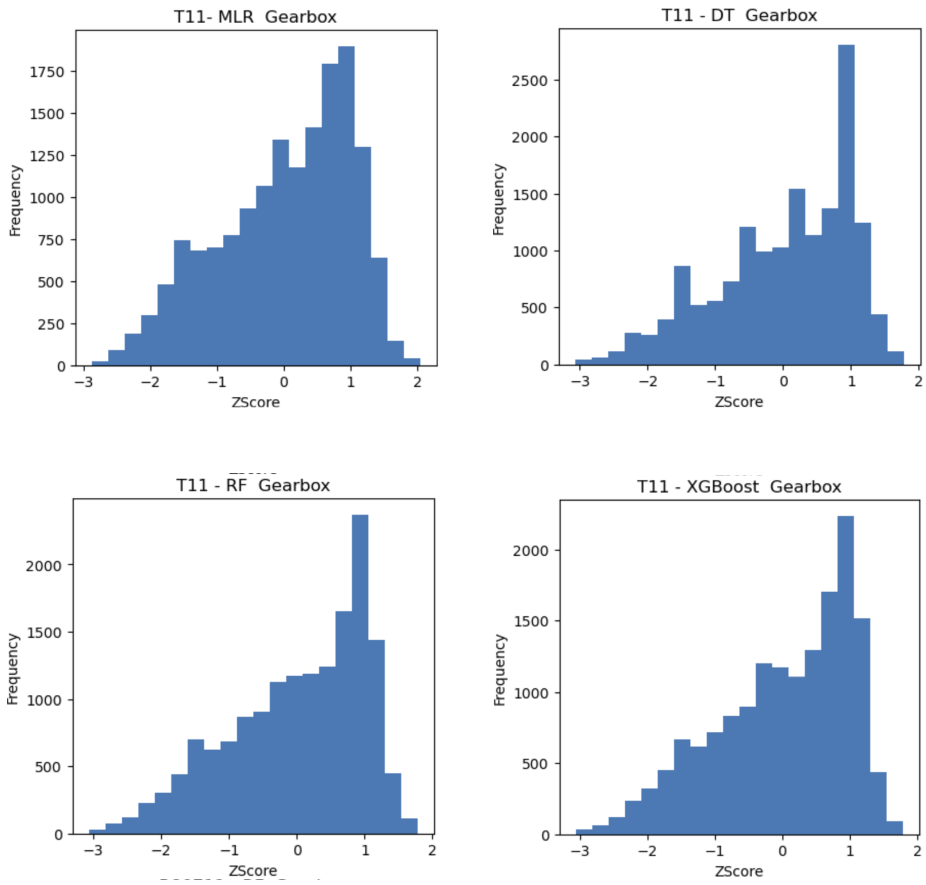

Figure 20: T11 - ZScore of Gearbox bearing temperature on various models

to identify the best-performing model based on rigorous performance metrics including R-Squared, RMSE, MAE, and MAPE.

In the post-processing stage, the model was employed to predict the output variable by considering the mean and standard deviation of the input variable. The predicted output was then compared to the actual historical records, and any data points falling outside the predetermined fault threshold range were indicative of a fault in the wind turbine.

To demonstrate the effectiveness of our fault detection approach, we presented two case studies utilizing SCADA data from operational wind farms. In the first case study, we were able to gain valuable insights into the potential failure occurrences of the wind turbine, even without prior knowledge of what failure specifically entailed. In the second case study, we validated our approach, successfully predicting the WT fault before its actual occurrence, as confirmed by the failure logs of the wind farm as shown in Figure 23

R80711

	Generator	Gearbox	Predicted
0	2017-01-06	2017-01-05	2017-01-05
1	2017-08-29	2017-01-06	2017-01-06
2	nan	2017-09-13	2017-08-29
3	nan	nan	2017-09-13

R80790

	Generator	Gearbox	Predicted
0	2017-01-07	2017-01-03	2017-01-03
1	2017-06-22	2017-01-04	2017-01-04
2	nan	2017-01-06	2017-01-06
3	nan	2017-02-07	2017-01-07
4	nan	2017-09-13	2017-02-07
5	nan	nan	2017-06-22
6	nan	nan	2017-09-13

R80736

	Generator	Gearbox	Predicted
0	2017-01-07	2017-01-04	2017-01-04
1	2017-06-22	2017-09-13	2017-01-07
2	nan	nan	2017-06-22
3	nan	nan	2017-09-13

R80721

	Generator	Gearbox	Predicted
0	2017-01-07	2017-01-05	2017-01-05
1	2017-07-25	nan	2017-01-07
2	2017-07-26	nan	2017-07-25
3	nan	nan	2017-07-26

Figure 21: Engie turbines - Predicted dates of failures

T01

	Generator	Gearbox	Predicted	Actual
0	2017-08-11	2017-08-11	2017-08-11	2017-08-11

T06

	Generator	Gearbox	Predicted	Actual
0	2017-08-19	2017-08-19	2017-08-19	2017-08-19
1	nan	2017-10-17	2017-10-17	2017-10-17

T07

	Generator	Gearbox	Predicted	Actual
0	2017-06-17	2017-06-17	2017-06-17	2017-06-17
1	2017-08-20	2017-08-20	2017-08-20	2017-08-20
2	2017-08-21	2017-10-19	2017-08-21	2017-08-21
3	nan	nan	2017-10-19	2017-10-19

T11

	Generator	Gearbox	Predicted	Actual
0	2017-04-26	2017-04-26	2017-04-26	2017-04-26
1	2017-09-12	nan	2017-09-12	2017-09-12

Figure 22: EDP turbines - Predicted vs Actual dates of failures

Through our comprehensive methodology and accurate fault detection algorithm, we have provided a reliable solution for anticipating and identifying faults in wind turbines, which can significantly enhance operational efficiency and minimize potential downtime.

6. Conclusion

This research paper presents a comprehensive system for monitoring and detecting anomalies in wind turbine gearbox and generator using SCADA data and various machine learning algorithms including Multi Linear Regression (MLR), Decision Tree Regression (DT), Random Forest Regression (RF), and extreme

	Generator	Gearbox	Predicted	Actual
0	2017-01-25	2017-01-25	2017-01-25	2017-01-25
1	2017-04-26	2017-06-17	2017-04-26	2017-04-26
2	2017-06-17	2017-08-11	2017-06-17	2017-06-17
3	2017-08-11	2017-08-19	2017-08-11	2017-08-11
4	2017-08-19	2017-08-20	2017-08-19	2017-08-19
5	2017-08-20	2017-09-12	2017-08-20	2017-08-20
6	2017-09-12	2017-10-17	2017-09-12	2017-08-21
7	2017-09-16	2017-10-18	2017-09-16	2017-09-12
8	nan	2017-10-19	2017-10-17	2017-09-16
9	nan	nan	2017-10-18	2017-10-17
10	nan	nan	2017-10-19	2017-10-18
11	nan	nan	nan	2017-10-19

Figure 23: EDP data - Predicted vs Actual dates of failures

gradient boosting (XGBoost). The system utilizes the mean and standard deviation of the output features from the training phase, specifically the Generator bearing temperature and Gearbox bearing temperature. This information was used to calculate the Z-score for the predicted temperature values generated by our model. We effectively assess and identify any outliers present in the Z-scores of our model's predictions. This model evaluates the deviations between the predicted temperature and the recorded temperature, enabling effective fault detection.

To evaluate the performance and applicability of the proposed method, two real case studies involving eight different wind turbines were conducted. The results showed that MLR exhibited the lowest performance among all models, while XGBoost consistently outperformed the other models in building generator and gearbox models for the listed wind turbines.

The effectiveness of our fault detection algorithm was demonstrated by successfully detecting faults in wind turbines that had no failure logs. By predicting when faults are likely to occur, our algorithm can assist asset managers of newly installed wind farms in planning for early intervention to prevent catastrophic damage. This dynamic data-driven maintenance strategy offers significant cost savings compared to the traditional static time-based maintenance approach.

Moving forward, our research will focus on exploring the use of statistical process control (SPC) techniques to assess the sensitivity level of deviations and calculate the remaining useful life (RUL) using models such as long short-term memory (LSTM). Additionally, we will investigate the application of streaming data for real-time fault detection and utilize deviation signatures from control charts to carry out fault diagnosis, specifically identifying the subcomponents of main components that are prone to failure. Collaboration with domain experts will be crucial in establishing data requirements and defining the normal behavior of these subcomponents, aiming to build a robust system that optimizes the

operation of the wind energy sector while ensuring cost-effectiveness.

In conclusion, this research paper introduces a comprehensive system for monitoring and detecting faults in wind turbines, leveraging SCADA data and advanced machine learning algorithms. The proposed fault detection algorithm demonstrates its effectiveness and applicability, providing valuable insights for asset managers and maintenance crews in the wind energy sector. The future steps outlined will further enhance the system's capabilities, enabling more accurate Remaining Useful Life calculations and real-time fault detection for optimized operations.

Acknowledgement

"I firmly believe that success is not achieved in isolation but through the assistance and support of others. I am profoundly grateful for the unwavering guidance and backing provided by my professor, Mr. Antorweep Chakravorthy, during my master's program at the University of Stavanger. At every stage of my study, he imparted knowledge and played an indispensable role in the accomplishment of this thesis.

I extend my heartfelt appreciation to all my colleagues and participants at Tietoevry Create Norway, who afforded me the opportunity to conduct my research. I would like to express my special thanks to Mr. Amund Haugeseth for his mentorship and facilitation throughout the entire thesis writing process. Their valuable support and industry expertise significantly contributed to shaping my research and enabling the completion of this thesis.

Words cannot adequately convey the depth of my gratitude to my family members, whose unwavering encouragement in all my endeavors has inspired me to pursue my dreams. They have supported me unconditionally and have been a source of moral support and motivation during both favorable and challenging times. It is their boundless love and unwavering support that have paved the path for my entire journey."

References

[1] L. Xiang, X. Yang, A. Hu, H. Su, P. Wang, Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks, Applied Energy 305 (2022) 117925.

[2] Norway, IEA Wind TCP - Global Wind Energy Research Collaboration, accessed: 2023-04-15.
URL `https://iea-wind.org/about-iea-wind-tcp/members/norway/`

[3] Wind energy, International Renewable Energy Agency, accessed: 2023-04-15.
URL `https://www.irena.org/Energy-Transition/Technology/Wind-energy`

[4] Z. Xu, J. Wei, S. Zhang, Z. Liu, X. Chen, Q. Yan, J. Guo, A state-of-the-art review of the vibration and noise of wind turbine drivetrains, Sustainable Energy Technologies and Assessments 48 (2021) 101629. doi:10.1016/j.seta.2021.101629.

[5] W. Teng, Y. Liu, Y. Huang, L. Song, Y. Liu, Z. Ma, Fault detection of planetary subassemblies in a wind turbine gearbox using tqwt based sparse representation, Journal of Sound and Vibration 490 (2021) 115707. doi:10.1016/j.jsv.2020.115707.

[6] T. Wang, Q. Han, F. Chu, Z. Feng, Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review, Mechanical Systems and Signal Processing 126 (2019) 662–685. doi:10.1016/j.ymssp.2019.02.051.

[7] W. Teng, X. Ding, Y. Zhang, Y. Liu, Z. Ma, A. Kusiak, Application of cyclic coherence function to bearing fault detection in a wind turbine generator under electromagnetic vibration, Mechanical Systems and Signal Processing 87 (2017) 279–293. doi:10.1016/j.ymssp.2016.10.026.

[8] The MathWorks, Predictive maintenance, part 1: Introduction video, The MathWorks, accessed: Aug. 21, 2022.
URL https://U.K.mathworks.com/videos/predictive-maintenance-part-1-introduction-1545827554336.html

[9] A. Stetco, F. Dinmohammadi, X. Zhao, V. Robu, D. Flynn, M. Barnes, J. Keane, G. Nenadic, Machine learning methods for wind turbine condition monitoring: A review, Renewable Energy 133 (2019) 620–635. doi:10.1016/j.renene.2018.10.047.

[10] A. F. Dakhil, W. M. Ali, A. A. Abdulredah, Predicting prior engine failure with classification algorithms and web-based iot sensors, in: Proceedings of the Emerging Technologies in Computing, Communications and Electronics (ETCCE), 2020, pp. 1–6. doi:10.1109/etcce51779.2020.9350895.

[11] K. Leahy, R. L. Hu, I. C. Konstantakopoulos, C. J. Spanos, A. M. Agogino, Diagnosing wind turbine faults using machine learning techniques applied to operational data, in: Proceedings of the IEEE International Conference on Prognostics and Health Management (ICPHM), 2016, pp. 1–8. doi:10.1109/ICPHM.2016.7542860.

[12] Y. Liu, Z. Wu, X. Wang, Research on fault diagnosis of wind turbine based on scada data, IEEE Access 8 (2020) 185557–185569. doi:10.1109/access.2020.3029435.

[13] Y. Wang, X. Ma, P. Qian, Wind turbine fault detection and identification through pca-based optimal variable selection, IEEE Transactions on Sustainable Energy 9 (4) (2018) 1627–1635. doi:10.1109/tste.2018.2801625.

[14] M. Beretta, J. Cárdenas, C. Koch, J. Cusidó, Wind fleet generator fault detection via scada alarms and autoencoders, Applied Sciences 10 (23) (2020) 8649. doi:10.3390/app10238649.

[15] A. Verma, A. Kusiak, Fault monitoring of wind turbine generator brushes: A data-mining approach, Journal of Solar Energy Engineering 134 (2) (2012) 1–5. doi:10.1115/1.4005624.

[16] Z. Liu, C. Xiao, T. Zhang, X. Zhang, Research on fault detection for three types of wind turbine subsystems using machine learning, Energies 13 (2) (2020) 460. doi:10.3390/en13020460.

[17] Y. Zhao, D. Li, A. Dong, D. Kang, Q. Lv, L. Shang, Fault prediction and diagnosis of wind turbine generators using scada data, Energies 10 (8) (2017) 1210. doi:10.3390/en10081210.

[18] J. Chatterjee, N. Dethlefs, Deep learning with knowledge transfer for explainable anomaly prediction in wind turbines, Wind Energy 23 (8) (2020) 1693–1710. doi:10.1002/we.2510.

[19] B. Manobel, F. Sehnke, J. A. Lazzás, I. Salfate, M. Felder, S. Montecinos, Wind turbine power curve modeling based on gaussian processes and artificial neural networks, Renewable Energy 125 (2018) 1015–1020. doi:10.1016/j.renene.2018.02.081.

[20] M. Schlechtingen, I. F. Santos, S. Achiche, Using data-mining approaches for wind turbine power curve monitoring: A comparative study, IEEE Transactions on Sustainable Energy 4 (3) (2013) 671–679. doi:10.1109/tste.2013.2241797.

[21] P. Trizoglou, X. Liu, Z. Lin, Fault detection by an ensemble framework of extreme gradient boosting (xgboost) in the operation of offshore wind turbines, Renewable Energy 179 (2021) 945–962. doi:10.1016/j.renene.2021.07.085.

[22] H. Chen, H. Liu, X. Chu, Q. Liu, D. Xue, Anomaly detection and critical scada parameters identification for wind turbines based on lstmae neural network, Renewable Energy 172 (2021) 829–840. doi:10.1016/j.renene.2021.03.078.

[23] J. Fu, J. Chu, P. Guo, Z. Chen, Condition monitoring of wind turbine gearbox bearing based on deep learning model, IEEE Access 7 (2019) 57078–57087. doi:10.1109/access.2019.2912621.

[24] R. Orozco, S. Sheng, C. Phillips, C. Phillips, Diagnostic models for wind turbine gearbox components using scada time series data, IEEE Xplore, accessed: Aug. 6, 2022 (Aug. 2018).
URL `https://ieeexplore.ieee.org/document/8448545`

[25] H. S. Dhiman, D. Deb, J. Carroll, V. Muresan, M.-L. Unguresan, Wind turbine gearbox condition monitoring based on class of support vector regression models and residual analysis, Sensors 20 (23) (2020) 6742. doi:10.3390/s20236742.

[26] P. Qian, X. Tian, J. Kanfoud, J. Lee, T.-H. Gan, A novel condition monitoring method of wind turbines based on long short-term memory neural network, Energies 12 (18) (2019) 3411. doi:10.3390/en12183411.

[27] F. Castellani, D. Astolfi, F. Natili, Scada data analysis methods for diagnosis of electrical faults to wind turbine generators, Applied Sciences 11 (8) (2021) 3307. doi:10.3390/app11083307.

[28] A. Santolamazza, D. Dadi, V. Introna, A data-mining approach for wind turbine fault detection based on scada data analysis using artificial neural networks, Energies 14 (7) (2021) 1845. doi:10.3390/en14071845.

[29] N. M. Khan, G. M. Khan, P. Matthews, Ai based real-time signal reconstruction for wind farm with scada sensor failure, in: Proceedings of the IFIP Advances in Information and Communication Technology, 2020, pp. 207–218. doi:10.1007/978-3-03049186-418.

[30] L.-L. Li, X. Zhao, M.-L. Tseng, R. R. Tan, Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm, Journal of Cleaner Production 242 (2020) 118447. doi:10.1016/j.jclepro.2019.118447.

[31] M. Yesilbudak, Implementation of novel hybrid approaches for power curve modeling of wind turbines, Energy Conversion and Management 171 (2018) 156–169. doi:10.1016/j.enconman.2018.05.092.

[32] D. C. Montgomery, E. A. Peck, G. G. Vining, Introduction to linear regression analysis, John Wiley & Sons, 2021.

[33] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, Multivariate data analysis (8. baskı), Eight Edition, Cengage: Learning EMEA (2019).

[34] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees. wadsworth & brooks, Cole Statistics/Probability Series (1984).

[35] J. R. Quinlan, Introduction of decision trees, Machine learning 1 (1986) 81.

[36] T. Hastie, R. Tibshirani, J. H. Friedman, J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Vol. 2, Springer, 2009.

[37] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[38] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R news 2 (3) (2002) 18–22.

[39] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[40] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[41] E. OpenData, La haute borne data 2013-2016 (2023).
URL https://opendata-renewables.engie.com/explore/dataset/la-haute-borne-data-2013-2016/export/?refine.wind_turbine_name=R80711

[42] E. OpenData, Open data (2023).
URL https://opendata.edp.com/open-data/en/data.html

[43] D. Menezes, M. Mendes, J. A. Almeida, T. Farinha, Wind farm and resource datasets: A comprehensive survey and overview, Energies 13 (18) (2020) 4702.