Gearbox Failure Prediction in Wind Turbines

Jagadeesh Kovi Harshavardhan Reddy Dhoma Suraj Vamshi Muthyam

project-jagakovi-hdhoma-smuthyam

Abstract

This project aims to enhance the reliability and efficiency of wind turbines by minimizing unplanned failures and maximizing power generation. The primary objective is to develop an effective Failure Forecasting model for predicting Gearbox failure in wind turbines. The model will be trained utilizing historical data and will leverage machine learning techniques to consider various influencing factors, including temperature variations, wind direction, and Yaw angle. By doing so, this project seeks to contribute to the sustainability of renewable energy sources, ultimately improving the performance and longevity of wind turbine systems.

Keywords

Wind turbine, Gearbox failure, Failure forecasting model, Renewable energy, Machine learning

1 Introduction

Wind turbines have become essential in our renewable energy landscape, providing clean and sustainable power. However, they often face unplanned failures leading to downtime and reduced energy output. To address this issue, our project focuses on developing a Failure Forecasting model, specifically for predicting Gearbox failures.

Our main goal is to leverage machine learning and historical data to create an accurate model for anticipating Gearbox failures in wind turbines. These failures can significantly disrupt wind energy systems, incurring costly repairs and operational downtime. By predicting these failures, we can implement proactive maintenance strategies, reduce unexpected interruptions, and optimize energy production.

We consider several influencing factors, including temperature variations, wind direction, and Yaw angle, all of which play a vital role in determining the health and performance of wind turbines.

This project not only improves wind turbine efficiency and longevity but also aligns with our commitment to sustainable renewable energy sources. By minimizing unplanned failures and maximizing power generation, we contribute to a cleaner and more eco-friendly energy landscape. This introduction paves the way for a detailed exploration of our project's scope, methodology, and expected outcomes in the subsequent sections.

Previous work

The authors at frontier [3], focus on statistical learning-based approaches for fault diagnosis and anomaly detection. The next reference [2] discuss about the effects of vibration on the failure of the gearboxes. The paper [7] primarily focuses on experimental methods such as SEM imaging,

nanoindentation, and Hertzian stress calculations to analyze the damage and failure modes of the bearings. We refer the mentioned research to develop our ML algorithms.This paper [**?**] by Leahy et al. (2018) centers on the diagnosis and prediction of wind turbine faults using Support Vector Machines (SVM) based on SCADA (Supervisory Control and Data Acquisition) data.

2 Methods

To predict Gearbox failures in wind turbines, we have several machine learning tools at our disposal. Naive Bayes is simple and efficient, suitable for classification tasks. Support Vector Machines (SVM) are great for binary classification, while k-Nearest Neighbors (KNN) works well with limited data. Decision Trees provide clarity and help select important features, Random Forest is good at handling complex data relationships, and Logistic Regression is straightforward and easy to understand.

Each tool has its unique strengths. Naive Bayes is straightforward and efficient. SVM is great for binary choices, and KNN is good for small datasets. Decision Trees help us understand important features, Random Forest handles complex relationships, and Logistic Regression is simple and clear.

By using these tools, we aim to create a strong model for predicting Gearbox failures in wind turbines. Each tool plays a specific role, helping us make accurate predictions and improve the reliability of wind energy systems.

We will select the most appropriate algorithm based on experimentation, guided by the characteristics of the dataset and the specific requirements of predictive maintenance. The optimal approach to forecasting Gearbox failures will be chosen through thorough performance evaluation, ensuring the effectiveness of the selected model.

2.1 Data Collection

Our project utilized the SCADA (Supervisory Control and Data Acquisition) dataset sourced from Kaggle for its data collection. This dataset comprises operational and environmental parameters collected from wind turbines, providing valuable insights into the performance and health of these renewable energy systems. The dataset's origin from Kaggle ensures its credibility and accessibility to the wider data science community. Leveraging this SCADA dataset allowed for a comprehensive exploration of operational patterns, identification of failure events, and the development of predictive models for effective maintenance strategies in the realm of wind turbine management.

2.2 Data Description

- In our dataset, we have key parameters such as wind speed (m/s), representing the survival speed of commercial wind turbines, indicating how fast the air moves past a specific point."

- Power (kW) is another critical variable, denoting the output of the wind turbine. Wind power, in this context, describes the conversion of wind energy into mechanical power or electricity."

- We're also monitoring temperatures across various components: Gear oil, Ambient, Nacelle, Bearing, and Wheel Hub, providing insights into the thermal conditions of these crucial elements."

- Rotor Speed, a crucial part of the drivetrain, reflects the rotational speed of the wind turbine rotor, typically spinning between 8 to 20 rotations per minute."

- Wind direction, expressed in degrees, informs us about the direction of the wind, adding a directional component to our dataset."

- Generator Speed represents the rotational speed necessary for most generators to produce electricity, ensuring the turbine's generator can efficiently generate AC electricity."

2.3 Score Evaluation

1. Accuracy:
Formula: Accuracy = $\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$
Accuracy represents the overall correctness of the model and is calculated by dividing the number of correct predictions by the total number of predictions.
2. Precision:
Formula: Precision = $\frac{\text{True Positives}}{\text{True Positives + False Positives}}$
Precision focuses on the accuracy of positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positives. A high precision indicates a low false positive rate.
3. Recall (Sensitivity or True Positive Rate):
Formula: Recall = $\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$
Recall measures the ability of the model to capture all the relevant instances. It is the ratio of correctly predicted positive observations to the total actual positives. A high recall indicates a low false negative rate.
4. F1 Score:
Formula: F1 Score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$
The F1 Score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is particularly useful when the class distribution is imbalanced.

These formulas are applied to the confusion matrix, which is a table used to evaluate the performance of a classification algorithm. The confusion matrix contains four values: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

By calculating these values using the mentioned formulas, one can evaluate the performance of each algorithm in terms of accuracy, precision, and recall.

2.4 Model Development

The initial Principal Component Analysis (PCA) analysis resulted in the identification of three clusters (Figure 1), deviating from the expected two clusters for target classification. This unexpected outcome prompts a closer examination of the data and the PCA process to understand the underlying reasons for the observed cluster discrepancy.

In our efforts to develop a robust predictive model for wind turbine failure, we actively considered the utilization of several machine learning algorithms. The prospective models under consideration include Gaussian Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Logistic Regression. These models are chosen based on their inherent strengths and potential suitability for the classification task at hand. Gaussian Naïve Bayes is being considered for its computational efficiency and ability to handle numerous features. KNN, with its non-parametric nature and effectiveness in capturing localized patterns, shows promise in the context of wind turbine failure prediction. The interpretability of Decision Trees makes them a valuable candidate, while the ensemble approach of Random Forest enhances robustness. Logistic Regression, known for its simplicity and efficiency in binary classification, is also part of our deliberations.

The choice of the final model was based on the specific characteristics of the dataset and the overarching goals of the wind turbine failure prediction project. This approach allows for a thoughtful selection process based on the strengths and features of each model.

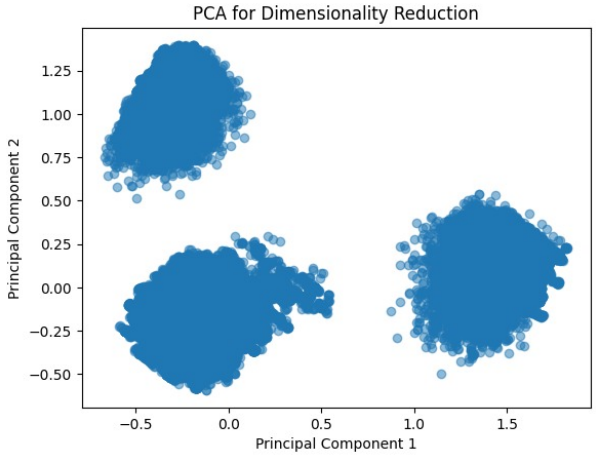

Figure 1: PCA for Dimensionality Reduction

3 Interpretation

To determine if these values are outliers, it's essential to understand the context of the data and the physical limitations of the variables. In the case of wind speed and rotor speed, negative values might indicate errors in data collection or recording, as these variables are typically non-negative in physical systems.

Here are some considerations:

1. Negative Wind Speed: - Wind speed should generally be non-negative. Negative values might be an anomaly or an error in data recording. It's advisable to investigate these cases and, if possible, correct or remove the erroneous data points.

2. Negative Rotor Speed: - Rotor speed is also expected to be non-negative in typical wind turbine systems. Having negative values might be indicative of a measurement issue or an error. It's recommended to review the data collection process and potentially remove or correct these data points.

From pairplots, we can deduce the following: While there is observable differentiation between the 'Failure' and 'No_Failure' categories in the pairwise plot, it's important to note that the plot is based on only 10% of the entire dataset. Additionally, we specifically chose a subset of features that we deem significant for this analysis.

4 Results

Gaussian Naïve Bayes achieved 95% accuracy, excelling in precision and recall.
K-Nearest Neighbors (KNN) showed a strong 94% accuracy, with notable precision and recall.
The Decision Tree model performed well with 91% accuracy, demonstrating balanced precision and recall.

Random Forest exhibited a high 95% accuracy, showcasing robust precision and recall.
Logistic Regression achieved a solid 94% accuracy, with strong precision and recall.
In summary, all models performed well, and the choice may depend on specific requirements for precision and recall.

5 Discussion

If there is a positive correlation between generator bearing temperature and gear oil temperature, it suggests that as one of these temperatures increases, the other is likely to increase as well. This positive correlation may indicate a thermal interaction between the generator bearing and the gear oil.

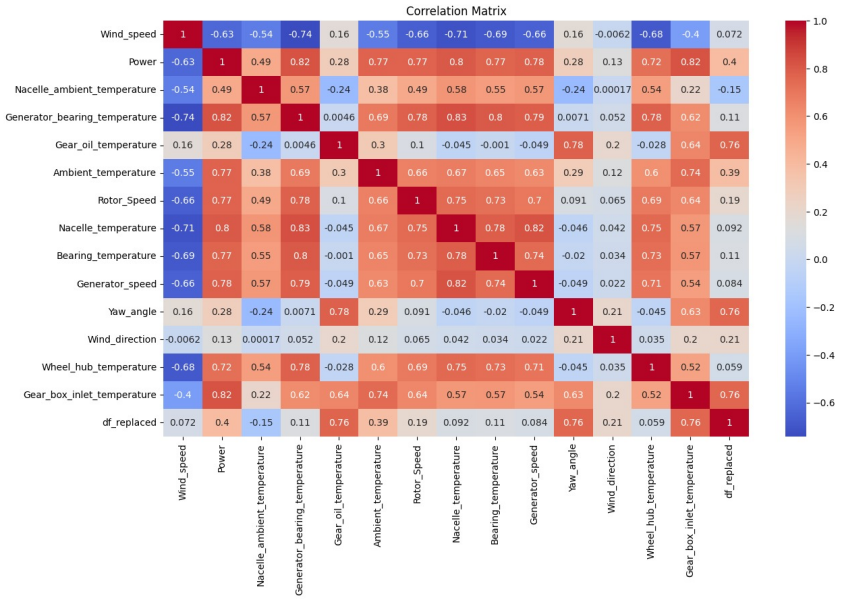

Figure 2: Correlation Matrix

Similarly, a positive correlation between generator bearing temperature and bearing temperature connectivity implies that as the generator bearing temperature increases, the connectivity related to bearing temperature also tends to increase. This could indicate that heat generated in the generator bearing affects the connectivity in some way.

The strength of the correlation coefficient provides information about the intensity of the relationship between variables. A correlation coefficient close to +1 indicates a strong positive correlation, while a coefficient close to -1 suggests a strong negative correlation. A coefficient near 0 indicates a weak or no linear correlation.

The distribution plot of wheel hub temperature, categorized by failure status, reveals a distinct pattern in the scaled temperature range of -50 to 100. Notably, there are no recorded failures within this interval. This observation may suggest a normal operating range where the system exhibits robust performance and is less prone to failures. Alternatively, it prompts an investigation into potential factors such as sensor calibration or data anomalies during this specific temperature range. Understanding this "safe" zone could inform the establishment of operational thresholds, aiding in failure prediction models. Further analysis is recommended to explore the reasons behind the change in failure status outside the -50 to 100 range, considering system characteristics and external influences on temperature dynamics.

6 Author Contribution

- **Data Gathering and Initial Exploration:**
  - Jagadeesh Kovi took the lead in gathering the dataset from Kaggle, ensuring its relevance to the project objectives.

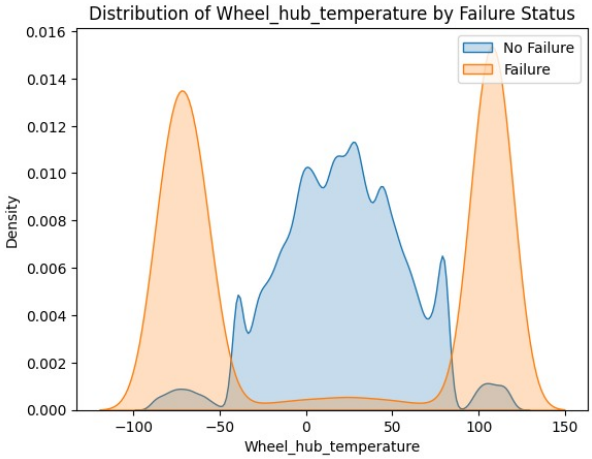

Figure 3: Distribution Plot of Wheel Hub Temperature

  - Along with Jagadeesh Kovi, Harshavardhan Reddy Dhoma conducted initial exploratory data analysis to understand the dataset's structure, identifying potential challenges and opportunities for preprocessing.

- **Preprocessing and Feature Engineering:**
  - Suraj Vamshi Muthyam is involved in preprocessing the gathered data. This involved handling missing values, cleaning outliers, and standardizing features.
  - He also contributed to feature engineering, identifying and creating relevant features that could enhance the performance of machine learning algorithms.
- **Algorithm Implementation and Evaluation:**
  - Suraj Vamshi Muthyam and Jagadeesh Kovi were responsible for implementing five different machine learning algorithms on the preprocessed data by carefully selecting and configuring each algorithm, considering the project's objectives.
  - Harshavardhan Reddy Dhoma conducted thorough model evaluation, including metrics such as accuracy, precision, recall, and F1 score to ensure a comprehensive assessment of each algorithm's performance.
- **Manuscript Drafting and Documentation:**
  - All three of us collaborated on drafting the manuscript.
  - Jagadeesh Kovi contributed to the introduction and previous work sections, setting the context for the project.
  - Harshavardhan Reddy Dhoma focused on detailing the dataset description and interpretation and feature engineering strategies.
  - Suraj Vamshi Muthyam took the lead in presenting the algorithm implementations, results, and discussions.

References

[1] Contributing Editor Andrea R. Aikin. Bearing and gearbox failures: Challenge to wind turbines. *https://www.stle.org/files/TLTArchives/2020/08_August/Feature.aspx?WebsiteKey = a70334df − 8659 − 42fd − a3bd − be406b5b83e5*, 10, 2022.

[2] James Carroll, Sofia Koukoura, Alasdair McDonald, Anastasis Charalambous, Stephan Weiss, and Stephen McArthur. Wind turbine gearbox failure and remaining useful life prediction using machine learning techniques. *Wind Energy*, 22(3):360–375, 2019.

[3] Shawn Sheng Effi Latiffianti and Yu Ding. Wind turbine gearbox failure detection through cumulative sum of multivariate time series data. *Frontiers*, 10, 2022.

[4] Alipujiang Jierula, Shuhong Wang, Tae-Min Oh, and Pengyu Wang. Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *Applied Sciences*, 11:2314, 03 2021.

[5] Kevin Leahy, R. Hu, Ioannis Konstantakopoulos, Costas Spanos, Alice Agogino, and Dominic O' Sullivan. Diagnosing and predicting wind turbine faults from scada data using support vector machines. *International Journal of Prognostics and Health Management*, 9, 02 2018.

[6] Koo Ping Shung. Accuracy, precision, recall or f1? *https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9*, 2018.

[7] H. Singh, R.V. Pulikollu, and W. et al. Hawkins. Investigation of microstructural alterations in low- and high-speed intermediate-stage wind turbine gearbox bearings. *Tribol Lett*, 65, 2017.

[8] Haderbieke M Guo L Cheng Z Tuerxun W, Xu C. A wind turbine fault classification model using broad learning system optimized by improved pelican optimization algorithm. *Machines*, 10(5):407, 2022.
[1] [8] [6] [5] [4]