Fault Identification in Wind Turbines: A Data-Centric Machine Learning Approach

*Short Paper to CSCI-RTCS

1st Danielle Pinna
PPCIC - CEFET/RJ
Rio de Janeiro, Brasil
danielle.pinna@eic.cefet-rj.br

2nd Rodrigo Toso
Research - Microsoft
Boston, USA
rfran@microsoft.com

3rd Rafaelli Coutinho
PPCIC - CEFET/RJ
Rio de Janeiro, Brasil
rafaelli.coutinho@cefet-rj.br

4th Ana Isabel Pereira
CeDRI - IPB
Bragança, Portugal
apereira@ipb.pt

5th Diego Brandão
PPCIC - CEFET/RJ
Rio de Janeiro, Brasil
diego.brandao@cefet-rj.br

***Abstract*—The last few years have been marked by the transition of the world energy matrix, predominantly with wind and solar sources considered clean energies. Wind turbines, responsible for the energy conversion process, are complex and expensive equipment susceptible to several failures due to multiple factors. Monitoring turbine components can assist in detecting failures before they occur, reducing equipment maintenance costs. This work compares machine learning techniques in a data-centric approach to wind turbine failure detection. Preliminary results demonstrate the importance of feature selection in this problem.**

***Index Terms*—wind turbine, machine learning, fault classification**

I. Introduction

Wind energy is an essential resource of clean and renewable energy available in nature that benefits global and economic interests. One of the goals that Brazil committed to in the Paris Agreement, held in 2015, was to increase the use of alternative energy sources to reduce greenhouse gas emissions.

The wind energy industry in Brazil closed the year 2021 with 21.57GW of installed capacity, representing a growth of 21.53% over the previous year [1]. In the 2021 world ranking of wind capacity, Brazil ranks sixth in total installed wind capacity onshore and third in newly installed wind capacity [2], underscoring the growing development of this sector.

This growth also comes with challenges related to reducing Wind Turbines' operation and maintenance (O&M) costs, which are sophisticated, complex, and expensive systems. According to [3], the OM of turbines accounts for about 25% to 35% of generation costs.

The problems related to wind turbine maintenance are usually electrical system component failures and those arising from extreme weather conditions. Some component failures do not occur so frequently, but a single disturbance can cause hours of lost productivity or even shutdown of the turbine.

For this reason, the most effective way to reduce maintenance costs is to monitor the status of critical components and predict their malfunction before the system fails [4]. Thus, early fault diagnosis is a critical factor in significantly reducing maintenance costs.

Modern wind turbines have a monitoring system known as a Control and Data Acquisition System (SCADA) [5]. As the name suggests, this monitoring system is powered by multiple sensors that provide measurements of critical process variables. However, extracting answers that enable failure prevention from a large amount of data is not a simple task and requires the application of increasingly sophisticated methods [6].

Most works on wind turbine fault detection are based on operational and event datasets, such as those provided by SCADA [7]. This work aims to compare different machine learning techniques in a data-centric approach to aid in fault detection of real wind turbine data components extracted from SCADA.

This paper is organized into five more sections. Related works are presented in Section 2. Section 3 presents the theoretical background of machine learning techniques. The methodology is presented in Section 4. The results are discussed in section 5, and finally, section 6 presents the final considerations.

II. Related Works

Stetco *et al.* [7] provide an extensive literature review of machine learning (ML) models used in wind turbine monitoring, including the fault detection task. The analyzed steps of ML models were: data source, selection and extraction of *features*, model selection and validation, and decision making. The results obtained show that most models use SCADA data with classification models.

Before applying the prognosis algorithms for wind turbines, Marti-Puig et al. [5] stressed the importance of implementing a preprocessing step, which is often underestimated by not considering its significant impact on the final results. The

2769-5654/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00106

Authorized licensed use limited to: b-on: Instituto Politecnico de Braganca. Downloaded on January 17,2024 at 16:00:15 UTC from IEEE Xplore. Restrictions apply.

authors evaluated the impact of removing extreme values (outliers) and found that removing it is not a good practice, as these values are the least frequent turbine operating mode (failure states).

De Sa et al. [8] proposed a workflow to detect wind turbine failures in the high-dimensional dataset obtained by SCADA. The flow was divided into two stages, the first in the automatic feature selection, using the Non-dominated Sorting Genetic Algorithm II (NSGA-II), and the second in the semi-supervised classification based on a Soft-Label and in a binary support vector machine (SVM). The linear SVM results found by the authors outperform the reference.

Garan et al. [9] discuss that most papers in the literature are model-based and not data preprocessing-based, emphasizing that more effort should be put into the quality of the dataset in order to improve the performance of classification measures. They propose a data-centric methodology, where data-driven steps are performed iteratively in wind turbine fault detection.

III. Theoretical Foundation

In recent years, with the growing volume of data generated and greater complexity of the problems to be dealt with computationally, it has become necessary to use more sophisticated and autonomous computational tools. One example is machine learning (ML) algorithms that can model data and solve complex problems [10].

According to [11], as ML depends on its data, having high-quality data plays a decisive role in building reliable and robust models, as opposed to just a good training algorithm.

Machine learning problems can be classified into supervised and unsupervised. Supervised machine learning is one of the standard learning techniques, whose objective is to learn a predictive model from a dataset, composed of a target variable and a set of explanatory variables.

On the other hand, unsupervised learning seeks to identify similarities between input objects so that those with something in common are categorized together, such as cluster analysis.

In this study, the interest is in the supervised machine learning problem, more specifically in the binary classification problem to recognize the faults and the flawless operations of a wind turbine. The binary classification algorithm only deals with two classes, 0 or 1. Class 0 indicates a fault-free (healthy) observation and class 1 indicates fault (defective) observations.

A. Classification Methods

Logistic Regression, Naive Bayes and k-Nearest Neighbors are some of the main supervised machine learning classification algorithms, where the purpose of each algorithm is to develop a model capable of determining the class of examples that are not labeled.

1) Logistic Regression: It is a type of generalized linear model (GLM) used for binary classification. It aims to estimate discrete values (binary values such as 0/1, yes/no, true/false) based on a given set of explanatory variables. Normally, logistic regression uses a function "Sigmoid" (logistic function), which has an "S" curve, used for the binary classification that converts values to the interval [0,1], which can be interpreted as the probability that a given instance belongs or does not belong to a given class.

2) Naive Bayes: The Naive Bayes classifier is one of the methods developed for supervised classification problems, due to its simplicity and robustness. It is based on the application of Bayes' theorem with the "naive" assumption of conditional independence between each pair of features, given the value of the class variable [12].

3) k-Nearest Neighbors: It is a method based on the concept of distance, that is, on the proximity between the data, which uses information from the data of k-neighbors made by a classifier based on memory. According to [10], the base hypothesis is that similar data tend to be concentrated in the same region in the input space and, in the same way, data that are not similar will be distant from each other. The k parameter is user-defined and in classification problems, it is common to use odd values to avoid ties.

To measure the performance of classification algorithms and verify the model's ability to generalize to unseen data examples, several metrics can be used in the context of binary classification, such as the confusion matrix, accuracy, precision, sensitivity and F1-Score [13].

IV. Methodology

The adopted methodology, illustrated in Figure 1, consists of 4 main steps: 1) Data acquisition; 2) Data preprocessing; 3) Feature selection; and 4) Training and evaluation of supervised machine learning algorithms: Naive Bayes, Logistic Regression and k-Nearest Neighbors (KNN).

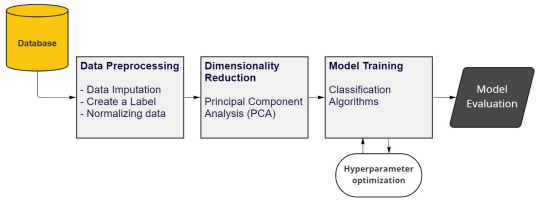

Fig. 1. Pipeline of the Adopted Methodology.

A. Data Description

The database used is provided by Energias de Portugal (EDP) [14]. This is one of the most complete free datasets available for wind resource analysis and wind turbine performance research [15]. The availability of this data was based on a challenge proposed by the company in which the objective was to detect failures in wind turbines. The records were extracted from a SCADA system consisting of 5 wind turbines measured in the years 2016 and 2017.

The information provided by EDP is:

- *Metmast*: Dataset of meteorological mast variables, measured every 10 minutes. Data is extracted from a single tower. It has 40 variables related to wind speed and direction (2 anemometric sensors), temperature, atmospheric pressure, humidity and precipitation.

Authorized licensed use limited to: b-on: Instituto Politecnico de Braganca. Downloaded on January 17,2024 at 16:00:15 UTC from IEEE Xplore. Restrictions apply.

- *Failures*: Dataset with the record of failure occurrences of the 5 wind turbine components, measured at the time of each occurrence. The faulty components are: Transformer, Generator Bearing, Hydraulic Group, Generator and Gearbox.
- *Logs*: Dataset of the history of normal and abnormal events that occurred in each turbine.
- *Signals*: Dataset of the SCADA system sensor variables for the most important components and production values of each turbine, read every 10 minutes. It has 81 variables related to wind speed and direction, generator, transformer, etc.
- *Locations*: Location of turbines, containing latitude and longitude.

The dataset *Failures* provides the history of failures that occurred in the years 2016 and 2017. Figure 2 presents the failures for each component of the turbines and as can be seen, the Hydraulic Group was the component that most failed in this period, containing 8 failures in total, followed by the Generator with 7 failures.

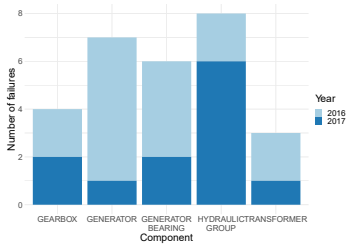

Fig. 2. Frequency of Failures in Wind Turbine Components.

B. Data Preprocessing

The application of data preparation techniques is important to improve data quality and to help machine learning algorithms to build models more faithful to the real distribution of data [10].

To create the dataset with sensor (*Signals*) and meteorological (*Metmast*) data, the bases were combined by the measurement time variable (*Timestamp*). For some instants of time, there was no measurement, and therefore, it was imputed so that the series was complete with all measurements every 10 minutes. This imputation occurred by repeating the values of the previous time. After this step, the failure data was included by the turbine ID and by the measurement time less than or equal to the failure time.

As the focus is on the classification problem, a field was created with the difference between the failure time and the measurement time, assigning a value of '1' to observations 60 days before the occurrence of failures and '0 ' for the others. This 60-day limit was chosen by the evaluation used in the EDP challenge, in which a failure up to 60 days in advance of the date of occurrence was counted as the correct classification.

Variables with little variance and named *offset* were removed from the dataset. Many of the variables contained information with minimum, maximum, average and variance values for the measurement time every 10 minutes, having a high correlation with each other. Thus, in order to remove redundant attributes, only the average values of each variable in the dataset were selected, totaling 60 features.

The preprocessing step also includes the normalization of the numerical data in order to eliminate the discrepancy of the measurement units between the variables.

C. Dimensionality Reduction

Often real data sets contain a large number of attributes, however, not all of them are informative for the process they are supposed to describe. In this way, it is necessary to apply resource selection techniques to identify variables that are actually useful for the analysis problem.

The mutual information [9], [16], extra trees classifier (ETC) [17] and principal component analysis (PCA) [9], [18], [19] are some of the methods found in the literature for the problem of fault detection in wind turbines.

PCA was applied to eliminate high correlation and reduce the dimensionality of multivariate data with minimal loss of information. The selection of the number of principal components was defined keeping 98% of the data variance.

D. Model Training

For the last step, the classifiers used were Naive Bayes, Logistic Regression and k-Nearest Neighbors (KNN), as an alternative to the results found in [9], which applied only Decision Trees (DT).

The database was divided into 80% for training and 20% for testing, maintaining the order of the dataset. To arrive at the most optimized model, the grid search method with cross-validation was used to define the hyperparameters. This method is effective because it configures a grid of hyperparameter values to train the model and thus choose the one with the best result by the metric F_1-Score, considering the cross-validation with 5 folds to generate an average estimate of hits and errors for each base.

V. Results and Discussions

The experiments were performed using computational routines implemented in Python version 3, on an Intel(R) Xeon(R) Gold 5120 CPU 2.20GHz, with 28 cores and 192*GB* of RAM. The libraries used were *pandas*[1], *numpy*[2] and *scikit-learn*[3].

Table I presents the performance metrics of the models in the test base: Accuracy, Precision, Recall and F_1-Score, in addition to the combination of hyperparameter values of each model that resulted in the highest F_1-Score and the number of main components selected for each wind turbine component.

The choice of the best algorithm tested for each component of the wind turbine was based on the F_1-Score metric. The results obtained in this work are compared with the results of the article benchmark by [9], which also used a data-centric

[1]https://pandas.pydata.org/
[2]https://numpy.org/
[3]https://scikit-learn.org/stable/

Authorized licensed use limited to: b-on: Instituto Politecnico de Braganca. Downloaded on January 17,2024 at 16:00:15 UTC from IEEE Xplore. Restrictions apply.

TABLE I
METRICS FOR TESTING DATA ON EACH TURBINE COMPONENT.

Gearbox

Model	Accuracy	Precision	Recall	F_1-*Score*	Hyperparameter
Naive Bayes	43.89%	0.03%	0.02%	0.03%	-
Logistic Regression	69.35%	61.80%	21.72%	**32.14%**	penalty='none', solver='newton-cg'
KNN	60.14%	8.93%	2.09%	3.39%	n_neighbors=1, weights='distance'

Number of principal components: 21

Generator Bearing

Model	Accuracy	Precision	Recall	F_1-*Score*	Hyperparameter
Naive Bayes	68.60%	73.55%	9.67%	17.09%	-
Logistic Regression	67.90%	62.38%	10.34%	17.73%	penalty='none', solver='lbfgs'
KNN	63.84%	39.35%	14.80%	**21.51%**	n_neighbors=1, weights='distance'

Number of principal components: 18

Transformer

Model	Accuracy	Precision	Recall	F_1-*Score*	Hyperparameter
Naive Bayes	76.03%	68.54%	34.94%	**46.29%**	-
Logistic Regression	73.86%	85.56%	13.92%	23.95%	penalty='none', solver='sag'
KNN	70.28%	41.90%	1.38%	2.67%	n_neighbors=9, weights='distance'

Number of principal components: 19

Generator

Model	Accuracy	Precision	Recall	F_1-*Score*	Hyperparameter
Naive Bayes	73.15%	91.42%	3.26%	6.30%	-
Logistic Regression	74.81%	96.99%	9.24%	16.88%	penalty='l2', solver='lbfgs'
KNN	75.69%	66.45%	24.49%	**35.79%**	n_neighbors=1, weights='distance'

Number of principal components: 18

Hydraulic Group

Model	Accuracy	Precision	Recall	F_1-*Score*	Hyperparameter
Naive Bayes	54.79%	34.03%	8.70%	13.86%	-
Logistic Regression	62.05%	95.86%	9.62%	**17.49%**	penalty='l1', solver='liblinear'
KNN	57.25%	38.06%	3.61%	6.60%	n_neighbors=1, weights='distance'

Number of principal components: 21

approach comparing different attribute selection techniques for each turbine component applying only the classifier of the Decision Tree.

In Table II it is observed that for the Transformer and Generator components of the wind turbine, the results obtained from the F_1-Score using the Naive Bayes and KNN model, respectively, exceeded the case study that used only the Decision Tree. For the other turbine components, the F_1-Score obtained was lower than the benchmark. The adopted methodology provided results in which other simpler classifiers are able to detect component failures with greater precision.

TABLE II
COMPARISON OF RESULTS WITH THE BENCHMARK.

	Decision Tree [9]	***Results obtained***	
Component	F_1-*Score*	Best Model	F_1-*Score*
Gearbox	**37.73%**	Logistic Regression	32.14%
Generator Bearing	**36.29%**	KNN	21.51%
Transformer	8.08%	Naive Bayes	**46.29%**
Generator	9.59%	KNN	**35.79%**
Hydraulic Group	**44.85%**	Logistic Regression	17.49%

VI. CONCLUSIONS AND FUTURE WORK

Predictive maintenance of machines that wear out over time is an important method to improve process efficiency. Wind turbines are complex systems that require maintenance. As data acquisition increases, so does the possibility of applying machine learning algorithms combining data-centric approaches to improve the quality of the data entering the models.

Fault detection was performed on a real SCADA system dataset during wind turbine monitoring. Dimensionality reduction was applied by creating a new set of variables with minimal loss of information. The initially proposed methodology was able to predict component failures, and two of these components outperformed the literature results. The next steps aim to test other algorithms, such as ensemble methods. Also, evaluate oversampling methods and other dimensionality reduction techniques that consider the time dependence of the data.

VII. ACKNOWLEDGEMENTS

This work has been supported by FCT - Fundação para a Ciência e a Tecnologia within the RD Units Project Scope Research Center in Digitalization and Intelligent Robotics (CeDRI) UIDB/05757/2020 and UIDP/05757/2020 and SusTEC (LA/P/0007/2021). The authors would like to thank the following Brazilian Agencies CAPES, CNPq, and FAPERJ.

REFERENCES

[1] ABEEólica "Annual Bulletin - Brazilian Wind Energy Association (in portuguese)", 2021. Available: https://abeeolica.org.br/energia-eolica/dados-abeeolica/. [Accessed: 24- Jul- 2022].

[2] GWEC "Global Wind Report - Global Wind Energy Council", 2022. Available: https://gwec.net/global-wind-report-2022/. [Accessed: 24- Jul- 2022].

[3] A. Blanco-M. et al., "Impact of target variable distribution type over the regression analysis in wind turbine data," International Conference and Workshop on Bioinspired Intelligence (IWOBI), 2017, pp. 1-7.

[4] S. Qin et al. "Ensemble learning-based wind turbine fault prediction method with adaptive feature selection," Communications in Computer and Information Science, pp. 572–582, 2017.

[5] P. Marti-Puig et al. "Effects of the pre-processing algorithms in fault diagnosis of wind turbines," Environmental Modelling & Software, pp. 119–128, 2018.

[6] B. Corley et al., "Combination of thermal modelling and machine learning approaches for fault detection in wind turbine gearboxes," Energies, vol. 14, no. 5, p. 1375, 2021.

[7] A. Stetco et al. "Machine learning methods for wind turbine condition monitoring: A Review," Renewable Energy, vol. 133, pp. 620–635, 2019.

[8] F. P. de Sa et al., "Wind turbine fault detection: A semi-supervised Learning Approach with Automatic Evolutionary Feature Selection," International Conference on Systems, Signals and Image Processing (IWSSIP), 2020.

[9] M. Garan et al., "A data-centric machine learning methodology: Application on predictive maintenance of wind turbines," Energies, vol. 15, no. 3, p. 826, 2022.

[10] P. Norvig, S. Russell, "Artificial Intelligence: A Modern Approach", Pearson Education, 2021.

[11] N. Sambasivan et al., "Everyone wants to do the model work, not the data work: Data Cascades in high-stakes ai," Proceedings of the CHI Conference on Human Factors in Computing Systems, 2021.

[12] H. Zhang, "Exploring conditions for the optimality of naïve Bayes," International Journal of Pattern Recognition and Artificial Intelligence, vol. 19, no. 02, pp. 183–198, 2005.

[13] R. Kohavi and F. Provost, "Glossary of terms," Glossary of Terms Journal of Machine Learning. Available: https://ai.stanford.edu/ ronnyk/glossary.html. [Accessed: 08-Jul-2022].

[14] EDP Open Data. Available: https://opendata.edp.com/pages/homepage/. [Accessed: 15-Aug-2021].

[15] D. Menezes et al., "Wind Farm and Resource Datasets: A comprehensive survey and overview," Energies, vol. 13, no. 18, p. 4702, 2020.

[16] R. L. Hu et al., "Using Domain Knowledge Features for Wind Turbine Diagnostics," 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016.

[17] E. Mammadov et al., "AI-enabled Predictive Maintenance of Wind Generators," IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), 2021, pp. 1-5.

[18] C. Velandia-Cardenas et al., "Wind turbine fault detection using highly imbalanced real SCADA data," Energies, vol. 14, no. 6, p. 1728, 2021.

[19] C. Correa-Jullian et al., "Exploring Quantum Machine Learning and feature reduction techniques for wind turbine pitch fault detection," Energies, vol. 15, no. 8, p. 2792, 2022.

Authorized licensed use limited to: b-on: Instituto Politecnico de Braganca. Downloaded on January 17,2024 at 16:00:15 UTC from IEEE Xplore. Restrictions apply.