Statistical Data Mining through Credal Decision Tree
Classifiers for Fault Prediction on Wind Turbine Blades Using
Vibration Signals

Joshuva Arockia Dhanraj1*, P Jayaraman2, Kuppan Chetty Ramanathan1, J Pravin Kumar3, T Jayachandran3
1Centre for Automation & Robotics (ANRO), Department of Mechanical Engineering,
Hindustan Institute of Technology and Science, Chennai 603103, India.
2Department of Mechanical Engineering, Prathyusha Engineering College, Chennai
602025, Tamilnadu, India.
3Department of Mechanical Engineering, Rajalakshmi Institute of Technology,
Chennai 600124, Tamilnadu, India.

* Corresponding author: joshuva1991@gmail.com

Abstract. In a wind turbine, blades are the most important component of wind capture in wind
turbines as they easily become unreliable due to environmental conditions. This paper
demonstrates the malfunction characterization of wind turbine blades by the use of vibration data
via the credal decision tree (CDT). The defects on the blades are replicated to model the defects
through machine learning. The extraction of functions (statistical functions) and the selection of
the component (algorithm of decision tree J48) were employed to identify the best framework
for defect classification. Using the credal decision tree, 82.67% of classification accuracy have
been obtained with the Kappa statistic of 0.792 and mean absolute error of 0.0768.

Keywords: Condition monitoring, credal decision tree (CDT), feature extraction, feature
selection, feature classification.

1. Introduction
As a result of all-inclusive environmental defilement increment, advancement towards manageable
vitality, and green power sources, similar to, wind vitality has expanded. Wind vitality is one of the
inexhaustible, productive sources, and an ideal alternator for traditional sources of vitality. The wind
turbine’s primary aim is to improve productivity by enhancing vitality recovery and the wind turbine's
power production varies habitually with diminishing winds [1]. Because of natural and awful
meteorological conditions, the cutting edges are presented to serious vibration that prompts harm. The
cutting edge disappointment causes a reduction in vitality yield. Henceforth, fault analysis is
consolidated into the wind turbine. In this examination, the flaw conclusion on a wind turbine edge was
helped out utilizing a vibration signal through the machine learning approach [2]. There are two forms
of approaches used to analyze the conditions: the traditional approach and machine learning. The
standard approach is used primarily in systems where time does not shift in the recurrence section. The
non-standard signal is produced by turning machines. Because wear and tear in the recurrence segments,
the segregation of faults using customary FFT technique are problematic. Consequently, it isn't preferred. In the machine learning approach, calculations can adapt persistently and adjust to changing circumstances. Specialists regularly resort to the machine learning approach for fault conclusion of mechanical frameworks [3].

Jin et al., [4] also performed a state control report on the use of SCADA data processing of wind turbine generators. This paper suggests an overarching method for identifying irregularities and diagnosing wind turbine problems. For the modeling of their regular behavior, historic data from SCADA obtained from safe wind turbines is used as a reference room. Solimine et al., [5] introduced an experimental analysis on passive acoustic tracking of wind turbine blades for structural health monitoring. The technique uses blade-internal microphones to track patterns, changes, or spikes in the frequency of the blade cavity by using a small network of airborne acoustic sensors, passive device induction, and intermittent measuring windows. A research campaign for the system for systemic health surveillance applications was conducted on a fatigue-tested, utilitarian wind turbine rotor.

Wang et al, [6] published an assessment of wind turbine condition control, based on the control and acquisition data analyses of temperature-dependent components. This paper suggests a framework for determining the state of wind turbines by evaluating the temperature parameters of the measurement control and data acquisition system based on existing studies. Based on the main temperature signals from WTs a prediction model of time-period regression is developed to represent their health in residues. To track the state of a wind turbine tip, Dolinsky and Krawczuk [7] conducted measurements of modal parameters using a statistical method. The findings of the proposed study confirm the usefulness of the modal analysis and the mathematical estimation in the identification of losses. Xiao et al., [8] also surveyed wind turbine condition surveillance and vibration analysis. This article demonstrates a spectrogram, scalogram, and bi-spectrum study of vibrations in operating wind turbines. The findings show several non-stationary stochastic properties and volatility of the mode-coupling in the wind turbine tower vibrations.

Many studies were conducted using simulation analyzes of the wind turbine blade failure and design analysis, but only a few were performed experimentally [9]. For analysis, a very limited set of defects was taken into account. This is particularly true for wind turbine blade fault diagnoses [10]. There is therefore a strong need to design a fault diagnostic system, which can cope with several failures with a machine learning approach in wind turbine blades. Throughout his research, he attempted to discover five different conditions of blade failure by machine learning approaches and statistical analysis.

- This research investigates five defects for the diagnosis of wind turbine blade defects (blade break, corrosion, hub blade loose link, the twist angle of pitch, and bend).
- Method for statistical extraction of vibration signals has been used to extract the features needed.
- The algorithm of the J48 decision tree for the collection of dominating features.
- This study was modeled as a problem of multi-class grouping and attempting to identify the problem utilizing a credal decision tree (CDT).

2. Experimental Studies

In Joshuva and Sugumaran (2020) [11] explained in detail the structure, deficiencies, and operating strategies of the experiment, and Figure 1a and Figure 1b shows the methodology of the proposed work and its experimental setup. The test sampling frequency was 12,000 Hz and each signal was 10,000 data in length. The DYTRAN 3055B1 was used along with the DAQ (NI-USB 4432) for data capture. Figure 3 and Figure 4 shows the simulated fault on the blade and its vibration pattern.
3. Feature Extraction using Descriptive Statistical Analysis
In this study, descriptive statistical characteristics were extracted and the function extraction procedure was set out in [12]. Statistical characteristics such as mode, maximum, standard error, sum, mean, median, range, skewness, kurtosis, standard deviation, and variance of samples.

4. Feature Selection using J48 Decision Tree
The feature selection process is used where the user chooses certain functions that most add to the forecast attribute or output automatically or manually. The absence of relevant characteristics in the data will minimize the model's precision and make the model learn from irrelevant characteristics. J48 decision tree algorithm was used as a method for choosing functions. The feature selection mechanism was comprehensively explained in [13].

5. Credal Decision Tree (CDT) Classifier
This approach is close to the C4.5 Quinlan algorithms [14] for credal decision trees (CDT). The key distinction is that CDT measures the likelihood values of the attribute and class component with unprecise probabilities. Like is the case for the CDT process, a credal establishes a new split condition in the ambiguity check. CDT thus claims the training package is not highly reliable since it may be influenced by classes or noises (see Moral-García et al., (2020)) [15]. CDT can therefore be considered a good noisy field tool. CDT is established by replacing the criterion for dividing the C4.5 info-gain ratio with the imprecise criterion (IIGR) divided [16].

6. Result and Discussion
For the good condition of the blade and other wind turbine blades, the vibration signals were recorded. There have been cumulative sets of 600 samples, of which 100 were from a good condition set. The mathematical parameters have been derived and are used to join the CDT group. The required performance of the CDT algorithm will be the corresponding state of the classified data [17].
For CDT it was defined as default parameters such as batch size (100) and "true" debug option. The CDT classifier criteria are shown in Figure 2 and the CDT production in Figure 3 [18]. The 10-fold cross-validation outcome for CDT is shown in Figure 4. For 600 incidents, 496 (82.67%) were categorized correctly and 104 (17.33%) remaining instances were graded incorrectly. Other metrics such as kappa statistics are present: absolute error, squared root average, relative absolute error, and comparatively squared root error values (Figure 3) [19]. The uncertainty confusion matrix for CDT is shown in Figure 4. The diagonal elements represent the properly categorized instances in the uncertainty matrix, and the other instances are incorrectly classified. This is a positive state in the first row of the uncertainty matrix (Figure 4). The first variable (the location (1, 1)) is the number of correctly defined instances.
The second part (the location (1, 2)) is the number of good instances wrongly categorized as bend faults (bend). The third factor (the location (1, 3)) reflects the number of reasonable instances mistakenly labeled as crack defects. The fourth factor (the location (1, 4)) indicates the number of successful cases wrongly identified as a losing loss (loose) by the hub-blade. The fifth factor (the location (1, 5)) is the number of successful cases mistakenly categorized as pitch angle twist fault state. The sixth factor (location 1, 6) is the number of successful cases wrongly identified as erosion failure (erosion). Similarly, the second row is the second state, that is bending fault. The third row reflects the third state data points, i.e. crack breakdown. The fourth row displays the fourth data point, i.e. the loose malfunction in the hub-blade. The fifth range represents the fifth status data points, i.e. twist twisting fault. The sixth row reflects sixth state data points, i.e. erosion defects.

The diagonal components speak precisely to grouped examples in the confusion matrix, and the other instances are misclassified. Figure 5 demonstrates the thorough precision of the team. The individual consistency is represented by a class in terms of true (TP) positive, false (FP) positive, performance, recall, F-measurement, Matthews MCC, ROC, precise recall curves (PRC) [20]. For a good classifier, the true positive figure (TP) would exceed 1 and the false positives rate (FP) should be approximately 0. From Figure 5, the TP average in most groups is below 1 and the FP rate is below 0. It supports the finding in the uncertainty confusion matrix shown in Figure 4.

**Figure 4.** Confusion matrix for CDT

| a  | b  | c  | d  | e  | f  | <-- classified as |
|----|----|----|----|----|----|------------------|
| 74 | 0  | 1  | 24| 1  | 0  | a = Good         |
| 0  | 0  | 0  | 0  | 9  |    | b = Bend         |
| 0  | 1  | 4  | 1  | 0  | 5  | c = crack60      |
| 21 | 0  | 9  | 69| 1  | 0  | d = Loose        |
| 0  | 0  | 0  | 0  | 97 | 3  | e = PAT          |
| 0  | 2  | 0  | 0  | 92 |    | f = Erosion      |

**Figure 5.** Class-wise accuracy of CDT

### 7. Conclusion
In the use of wind energy for daily life, the wind turbine is very significant. The goal of this analysis is to identify wind turbine blades utilizing Credal Decision Trees (CDT). CDT was validated by 10-fold cross-validation, and 82.67% were found to be the properly classified instance. The total mean absolute error in this classifier is 0.0768. Therefore, for condition monitoring, credal decision trees can be used to detect the conditions for blade fault by wind turbine blades.

### 8. References
[1] Porté-Agel F, Bastankhah M and Shamsoddin S 2020 Wind-turbine and wind-farm flows: a review *Boundary-Layer Meteorology* **174** 1-59.
[2] Du Y, Zhou S, Jing X, Peng Y, Wu H and Kwok N 2020 Damage detection techniques for wind turbine blades: A review Mechanical Systems and Signal Processing 141 106445.

[3] Maldonado-Correa J, Martín-Martínez S, Artigao E and Gómez-Lázaro E 2020 Using SCADA Data for Wind Turbine Condition Monitoring: A Systematic Literature Review Energies 13 3132.

[4] Jin X, Xu Z and Qiao W 2020 Condition Monitoring of Wind Turbine Generators Using SCADA Data Analysis IEEE Transactions on Sustainable Energy 20.

[5] Solimine J, Niezrecki C and Inalpolat M 2020 An experimental investigation into passive acoustic damage detection for structural health monitoring of wind turbine blades Structural Health Monitoring 3 1475921719895588.

[6] Wang X, Zhao Q, Yang X and Zeng B 2020 Condition monitoring of wind turbines based on analysis of temperature-related parameters in supervisory control and data acquisition data Measurement and Control 53 164-80.

[7] Dolinski L and Krawczuk M 2020 Analysis of Modal Parameters Using a Statistical Approach for Condition Monitoring of the Wind Turbine Blade Applied Sciences 10 5878.

[8] Xiao F, Tian C, Wait I, Yang Z, Still B and Chen GS 2020 Condition monitoring and vibration analysis of wind turbine Advances in Mechanical Engineering 12 1687814020913782.

[9] Coppes J, Braunisch V, Bollmann K, Storch I, Mollet P, Grünschachner-Berger V, Taubmann J, Suchant R and Nopp-Mayr U 2020 The impact of wind energy facilities on grouse: a systematic review Journal of Ornithology 1 1-5.

[10] Wang Y, Yu Y, Cao S, Zhang X and Gao S 2020 A review of applications of artificial intelligent algorithms in wind farms Artificial Intelligence Review 53 3447-500.

[11] Joshuva A, Kumar RS, Sivakumar S, Deenadayalan G and Vishnuvardhan R 2020 An insight on VMD for diagnosing wind turbine blade faults using C4.5 as feature selection and discriminating through multilayer perceptron Alexandria Engineering Journal.

[12] Liu Z, Japkowicz N, Wang R, Cai Y, Tang D and Cai X 2020 A Statistical Pattern based Feature Extraction Method on System Call Traces for Anomaly Detection Information and Software Technology 25 106348.

[13] Nagra AA, Han F, Ling QH, Abubaker M, Ahmad F, Mehta S and Apasiba AT 2020 Hybrid self-inertia weight adaptive particle swarm optimisation with local search using C4.5 decision tree classifier for feature selection problems Connection Science 32 16-36.

[14] Zhou F, Xue L., Yan Z and Wen Y 2020 Research on college graduates employment prediction model based on C4.5 algorithm Journal of Physics: Conference Series 1453 012033.

[15] Moral-García S, Mantas CJ, Castellano JG, Benítez MD and Abellán J 2020 Bagging of credal decision trees for imprecise classification Expert Systems with Applications 141 112944.

[16] Nguyen PT, Ha DH, Nguyen HD, Van Phong T, Trinh PT, Al-Ansari N, Le HV, Pham BT, Ho LS and Prakash I 2020 Improvement of Credal Decision Trees Using Ensemble Frameworks for Groundwater Potential Modeling Sustainability 12 2622.

[17] Mattei L, Antonucci A, Mauá DD, Facchini A and Llerena JV 2020 Tractable inference in credal sentential decision diagrams International Journal of Approximate Reasoning 125 26-48.

[18] Maria M and Yassine C 2020 Machine Learning Based Approaches for Modeling the Output Power of Photovoltaic Array in Real Outdoor Conditions Electronics 9 315.

[19] Brus VR, Voronova LI and Voronov VI 2020 Neural Network Classification of Cardiac Activity Based on Cardiogram Data for Driver Support System In2020 Systems of Signals Generating and Processing in the Field of on Board Communications 1-5.

[20] Rahman MA, Haque MM, Anjum A, Mollah MN and Ahmad M 2020 Classification of motor imagery events from prefrontal hemodynamics for BCI application InProceedings of International Joint Conference on Computational Intelligence 2020.