End-to-End Deep Learning Approach for Wind Turbine Bearing Fault Detection From Acoustic Data

Gabriel Luiz Barros De Oliveira
gabriel_barros@atlantico.com.br
Instituto Atlântico
Fortaleza, CE, Brazil

Lucas Alves
lucas_alves@atlantico.com.br
Instituto Atlântico
Fortaleza, CE, Brazil

Cleilton L. Rocha
cleilton_rocha@atlantico.com.br
Instituto Atlântico
Fortaleza, CE, Brazil

João Pedro B. Lima
joao_brasil@atlantico.com.br
Instituto Atlântico
Fortaleza, CE, Brazil

Paulo Mesquita
abner@atlantico.com.br
Instituto Atlântico
Fortaleza, CE, Brazil

Alex Trajano
alex_ferreira@atlantico.com.br
Instituto Atlântico
Fortaleza, CE, Brazil

ABSTRACT

Wind energy generation through wind turbines has become increasingly attractive as a clean and renewable energy source. Effective maintenance of wind turbines is crucial, as failure can result in significant economic losses and damage to equipment due to unplanned downtime. Nevertheless, ensuring effective maintenance remains challenging because these systems usually operate in severe and remote environments. Considering that rolling bearings are among the most necessary mechanical components in wind turbines, accurately detecting faults in these bearings is important for ensuring the regular and reliable operation of the equipment. This paper proposes a deep learning-based monitoring architecture that utilizes acoustic signals emitted from rolling bearings to detect faults in wind turbines. Using real-world data collected via microphones and a Raspberry Pi system, we constructed a structured and manually annotated dataset. A convolutional neural network (CNN) model was trained on mel-spectrogram representations to distinguish between healthy and faulty operational states. The system achieved promising performance, with 83.62% accuracy, 87.40% F1-score, 95% AUC-ROC, and 98% precision-recall AUC. The proposed end-to-end pipeline integrates data acquisition, preprocessing, classification, and confidence-based decision thresholds, making it suitable for deployment in operational monitoring scenarios. These results demonstrate the viability of audio-based fault detection as a scalable and non-invasive solution for predictive maintenance in wind energy systems.

KEYWORDS

Audible noise; Fault detection; Artificial Intelligence; Wind Turbine

1 INTRODUCTION

The climate crisis, primarily driven by greenhouse gas emissions from human activities, has disrupted natural cycles, leading to global warming and climate change [7]. In response, the global shift towards for sustainable energy has accelerated the deployment of wind turbines as part of a cleaner energy matrix [24]. However, this growth has underscored significant operational challenges. The remote and harsh environments in which wind turbines operate make maintenance a costly and complex task. In addition, failures in components, such as gearboxes and bearings, are leading causes of unanticipated downtime [28, 30], often resulting in financial losses. Operation and maintenance (O&M) expenses can represent up to one-third of a wind farm's total life-cycle cost, making reliability and predictive maintenance crucial to the economic viability of wind energy systems [31].

Wind turbine bearings are essential for efficient wind energy production, and the system's dependability and operational uptime are directly impacted by their condition. Bearing failures are among the leading causes of wind turbine downtime and costly repairs [28]. These components are subjected to significant stress due to challenging and variable operating conditions, such as changing wind speeds, heavy loads, and extreme temperatures, which makes them susceptible to failure [11]. The main reasons for bearing failures include fatigue, contamination, misalignment, physical wear, overheating, and excessive load [14].

When a bearing's raceway or rolling element has an issue, such as a crack, spall, or insufficient lubricant, it causes repeated impacts each time the damaged section passes through the load zone. These impacts produce distinctive vibration and acoustic patterns that can be reliably detected by sensors, enabling effective condition monitoring of the bearing [35, 37].

Therefore, mitigating these problems through early detection or prevention of faults is essential and increasingly relies on advanced technological strategies [28]. The literature presents a wide range of approaches for detecting wind turbine faults, such as monitoring temperature, electrical signals, and oil conditions [30]. In particular, bearing defects have been effectively detected through the analysis of vibration and acoustic emission (AE) signals [34]. This strong correlation between bearing degradation and distinctive signal patterns underpins our work, enabling the monitoring of these critical components by analyzing their audio and vibration signatures.

With the advancement of artificial intelligence technologies, intelligent systems have become increasingly prevalent across various industrial sectors [27], acting both as facilitators and as solutions to complex problems. In this context, state-of-the-art machine learning techniques, particularly those driven by deep learning, have enabled the development of highly specialized models capable of recognizing intricate patterns in defined scenarios. Deep learning architectures construct hierarchical representations by stacking

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

non-linear modules, each of which transforms input data into progressively more abstract features. This compositional structure allows the learning of highly complex functions [20].

Among the diverse applications of deep learning is audio classification, a task that involves analyzing and categorizing audio signals [36]. Audio classification spans various domains, including environmental sound identification [18], and has also been explored for defect detection by recognizing distinctive audio signatures associated with mechanical faults [15].

This is particularly promising for wind turbine monitoring, where components such as bearings emit characteristic "beating" or grinding sounds when deteriorating [9, 30, 34]. These sounds can be captured via acoustic sensors and used to train deep learning models that distinguish between healthy and faulty states [14]. Compared to traditional vibration-based monitoring, which can be noisy and less interpretable, audio-based systems offer a complementary and often earlier indication of component degradation, enabling continuous, non-intrusive condition monitoring and supporting predictive maintenance strategies [35].

In this paper, we present a deep learning pipeline designed to monitor wind turbine operation by analyzing emitted sounds and detecting mechanical defects, specifically, bearing wear. Our approach employs an optimized Convolutional Neural Network (CNN) classifier to automatically learn and identify the acoustic signatures associated with early-stage faults. To achieve this goal, we created a dataset by collecting real-world audio recordings from operational wind turbines in the field. A specialized device was designed using the Raspberry Pi 5 and microphones to collect and process audio data. All recordings were annotated by specialists, resulting in a total of 172 minutes of labeled audio. Out of this, 106 minutes displayed defective operation, while 66 minutes showed normal operation. Labeled audio samples were included in this dataset to facilitate the training and evaluation of supervised models. We evaluated model performance using standard metrics, including accuracy, F1-score, AUC-ROC, and precision-recall. Our approach achieved outstanding results: 83% accuracy, 87% F1-score, 95% AUC-ROC, and 98% on the precision-recall curve. These results demonstrate that the use of a supervised approach, specifically the selection of an optimized CNN classifier, was highly effective for detecting wind turbine faults.

The paper is organized as follows. Section 2 presents the fundamental concepts of supervised learning and deep learning relevant to this work. Section 3 reviews related studies on wind turbine fault detection. Section 4 details the proposed method, including data collection, labeling, pre-processing, model design, training, and the classification pipeline. Section 5 presents and discusses the experimental results. Finally, Section 6 concludes the paper and outlines directions for future work.

2 BACKGROUND

This section outlines the fundamental concepts of supervised learning and deep learning relevant to this work.

2.1 Supervised Learning and Deep Learning

Supervised learning is a fundamental paradigm in machine learning, where a model is trained on labeled data to associate inputs with their corresponding outputs [26]. Deep learning has emerged as a powerful approach within supervised learning, producing state-of-the-art results in a variety of domains, including computer vision, speech recognition, natural language processing, machine translation, and biomedical data analysis [6].

Convolutional Neural Networks (CNNs) are a prominent class of deep learning models [5, 21], particularly effective for classification tasks involving spatial or temporal data [20]. Deep CNNs leverage multiple layers of convolutional operations to automatically learn hierarchical representations of data, enabling highly accurate classification and detection capabilities [19]. Their ability to capture local patterns and progressively abstract features makes them particularly well-suited for detecting subtle signal characteristics indicative of faults in complex mechanical systems.

One challenge in deploying deep learning models, including CNNs, in embedded or resource-constrained environments is their computational and memory demand [23]. This has motivated research into model compression, quantization, and low-precision representations to make deep learning more efficient and feasible for real-time or low-power applications.

In this work, we adopt a supervised learning approach using a CNN architecture to address the binary classification problem of identifying fault conditions in wind turbine bearings based on labeled acoustic data. The model was designed to balance predictive accuracy with computational efficiency to allow potential deployment in edge devices, while retaining the benefits of a full-precision CNN during training and inference.

2.2 Evaluation Metrics in Supervised Learning

In supervised learning, the evaluation of model performance is a fundamental step that determines not only predictive accuracy but also the reliability and robustness of the system under realistic conditions. Especially in industrial fault detection, where datasets are often imbalanced and the cost of misclassification is asymmetric, choosing appropriate evaluation metrics is critical to ensure a faithful representation of model behavior [12, 17, 33]. Metrics derived from the confusion matrix have become a standard tool to characterize performance across multiple perspectives, balancing precision, sensitivity, and overall predictive consistency.

The most common metric, **accuracy**, measures the proportion of correctly classified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

However, accuracy alone may be misleading in imbalanced datasets, as a model predicting only the majority class can achieve high accuracy while failing to detect rare but crucial faults [12].

To address this, additional metrics have been introduced to evaluate specific aspects of predictive behavior. **Precision** focuses on the reliability of positive predictions,

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

reflecting the system's ability to avoid false alarms, which in industrial monitoring directly impacts maintenance efficiency and operational trust [10]. Conversely, **recall** (or sensitivity),

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

measures the model's capacity to identify all true fault cases. High recall is especially valuable in safety-critical contexts, where undetected anomalies can cause severe equipment failures [32, 33].

Balancing these two objectives, the **F1-score** combines precision and recall into a single harmonic mean:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

providing a scalar summary that remains interpretable even when false positives and false negatives carry similar importance [33].

Because accuracy and F1-score can still be biased toward the majority class, **Balanced Accuracy** and the **Matthews Correlation Coefficient (MCC)** have been proposed as more robust alternatives. Balanced accuracy averages sensitivity and specificity:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \tag{5}$$

while MCC considers all elements of the confusion matrix:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{6}$$

MCC behaves as a correlation coefficient between true and predicted classifications, offering a balanced measure even under strong class imbalance [4].

Further refinements, such as **Informedness** and **Markedness**, extend this evaluation framework by quantifying informed decision-making and prediction reliability. Informedness, defined as

$$\text{Informedness} = \text{Recall} + \text{Specificity} - 1, \tag{7}$$

represents the probability that predictions are informed rather than random [29], whereas Markedness,

$$\text{Markedness} = \text{Precision} + \text{NPV} - 1, \tag{8}$$

captures the likelihood that predicted labels are correct, emphasizing both the precision of positive predictions and the reliability of normal classifications [29].

Beyond these threshold-dependent measures, the **Receiver Operating Characteristic (ROC)** curve and the corresponding **Area Under the Curve (AUC)** are widely used to evaluate a model's discriminative capacity independently of the decision threshold. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for varying thresholds, while the AUC summarizes this relationship into a single scalar value, providing a threshold-independent indicator of model separability [8].

Together, these metrics form the theoretical foundation upon which the effectiveness of deep learning models can be assessed. They complement the architectural and optimization techniques discussed earlier by providing a multidimensional understanding of how well a model generalizes, discriminates, and maintains reliability in real-world fault detection scenarios.

3 RELATED WORK

Memari et al. (2024) review inspection methods for wind turbine blades (WTBs) and present recent advances that integrate deep learning for automated defect detection using image-based data. The study emphasizes the role of drones in data acquisition and explores multiple sensing modalities, including infrared thermography and LiDAR, providing a comprehensive overview of aerial inspection, structural integrity assessment, and computer vision techniques. The comparison among neural architectures and feature extraction methods highlights trade-offs between accuracy, scalability, and automation. In contrast, our work diverges from visual-based inspection and instead focuses on audio-based fault detection, targeting internal mechanical components that are not directly observable through imaging.

Ferreira da Silva et al. (2025) propose a non-invasive method that leverages acoustic signals collected within the turbine nacelle to detect mechanical anomalies using artificial intelligence. Their system employs an unsupervised learning strategy based on autoencoders trained on healthy operational spectrograms. Faults are detected when reconstruction quality degrades significantly, after which the data are passed to supervised classifiers (Conditional Autoencoders and SVMs) for fault type identification. Although this work also utilizes acoustic data, our method differs in both its architectural design and methodological scope. Rather than adopting a multi-stage hybrid approach, we employ a fully supervised, end-to-end pipeline that directly maps raw audio-derived representations to operational state predictions.

Zhang et al. (2024) propose an aeroacoustic noise analysis (AAN) approach to identify cracking and debonding faults along the trailing edge of wind turbine blades. While their method effectively detects surface-level defects through the analysis of aerodynamic noise, it does not address internal mechanical components. Conversely, our approach specifically targets bearing degradation by employing convolutional neural networks to learn discriminative spectral features from audio recordings.

In summary, prior works have primarily focused on visual inspection or hybrid acoustic pipelines combining unsupervised and supervised methods. As shown in Table 1, our proposed system integrates data acquisition, pre-processing, and classification within a single CNN-based framework. This end-to-end design eliminates handcrafted feature engineering, simplifies deployment, and achieves robust performance across multiple metrics, demonstrating its suitability for real-world wind turbine monitoring. Direct numerical comparisons with prior methods are limited due to differences in approach (visual vs. acoustic), methodology (hybrid vs. end-to-end), and target component specificity.

Table 1: Comparison between related works and this study

Work	Data	Target	Main Contribution
Memari et al. [24]	Image	Blades	Survey on deep learning and drone-based inspection.
Ferreira da Silva et al. [9]	Acoustic	Internal Mechanical	Hybrid acoustic approach with unsupervised anomaly detection.
Zhang et al. [38]	Acoustic	Blades	Aeroacoustic noise analysis for crack identification.
This work	**Acoustic**	**Bearings**	**End-to-end supervised CNN for internal fault detection.**

4 PROPOSED METHOD

This section outlines the proposed method employed in this work, encompassing all stages from data collection to model implementation for detecting bearing faults in wind turbines.

4.1 Data Acquisition

We conducted our data acquisition process, targeting the collection of audio samples to support supervised model training and evaluation, using wind turbines with known bearing defects which were already expected to emit the noises characteristic of the fault in question. However, the characteristic fault noises are not constant but appear only periodically. This operational variability allows us to collect samples of non-defective operation even from known faulty turbines. Specifically, during periods when the fault noise is absent, the acoustic signature is similar to that of a turbine in a perfect bearing condition. This enables us to collect both non-defective and defective data, which is essential for building a robust and varied dataset and for building a model that can identify the differences between these audio files.

To this end, data collection was conducted according to the architecture shown in Figure 1, which comprises three layers: (1) physical, (2) gateway, and (3) cloud.

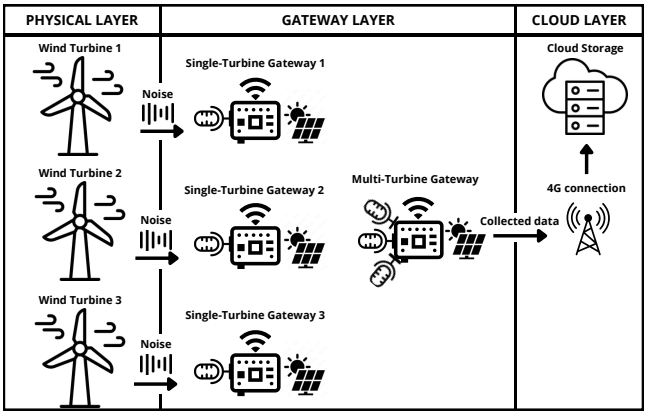

Figure 1: Data collection architecture

The first layer represents the three defective physical wind turbines, which continuously emit both normal and faulty acoustic signals. The second layer consists of autonomous gateways responsible for collecting the acoustic signals emitted by the wind turbines in the field. Each gateway is built using a Raspberry Pi 5 and specialized microphones, powered by solar panels, and equipped with a 4G module for data transmission, as well as a micro SD card for local data storage. It is essential to note that there are two types of gateways: single-turbine gateways and multi-turbine gateways. The first method collects acoustic signals from a single wind turbine, while the second gathers signals from all three turbines simultaneously. This strategy enhances data variability and improves the dataset for model training. Finally, the third layer is responsible for transmitting the collected data via a 4G connection to the cloud, enabling storage and further processing of the data as needed.

To support the development and evaluation of the classification models, a structured dataset was compiled from the WAV audio samples collected. Each entry contains metadata, such as file name, duration, sampling rate, and an associated class label, which is initially unset due to the absence of manual annotation during the early stages. The dataset comprises approximately 115 GB of audio, segmented into 30-second recordings, offering a substantial and representative sample of real-world operational and faulty conditions in wind turbines. The duration of 30 seconds was adopted as the standard sample length, as it represents a balanced trade-off between annotation efficiency and informational completeness. This duration is short enough to allow practical manual labeling while still being sufficiently long to capture the characteristic noise patterns associated with bearing defects. Moreover, shorter audio segments facilitate data transmission from remote collection devices by reducing file size and bandwidth requirements, ensuring efficient and reliable transfer of large-scale recordings for centralized processing. This dataset serves as the foundation for the supervised training and validation of fault detection models.

4.2 Data Labeling

To ensure the reliability of the labeled dataset used for supervised training, a standardized manual labeling protocol was established. This process involved three main stages: understanding the fault characteristics in data, performing exploratory analysis of the collected audio, and defining objective labeling criteria.

This standardization was critical for maintaining consistency across samples, allowing the labeling task to be distributed among multiple annotators without compromising quality. The resulting dataset, containing labeled instances of both normal and faulty audio segments, provided a solid foundation for training and evaluating machine learning models for fault detection.

From the total recorded audio corpus (approximately 381.8 hours), 346 audio samples were randomly selected, each lasting 30 seconds, amounting to a total of 2.88 hours. These samples were systematically organized into a structured dataset for manual annotation by 5 expert evaluators, following the established protocol. Only a subset of the available recordings was used, given the considerable effort required for manual labeling. Additionally, a portion of the collected data was discarded after being deemed contaminated by specialists, as it contained excessive noise.

Label	Quantity	Duration (h)
Normal	133	1.11
Defective	213	1.77

Table 2: Distribution of audio samples

The annotation process required approximately three hours in total, accounting for both the sample size and the time-intensive nature of manually listening to and analyzing each audio recording. Following this methodological approach, the 346 selected audio samples were systematically classified into two categories: normal (non-defective) and defective. After expert evaluation, the annotated dataset comprised 133 audio samples (1.11 hours) classified as normal and 213 as defective (1.77 hours), as shown in Table 2.

Although the total volume of audio collected exceeded 381 hours, only a subset of approximately 2.88 hours was manually annotated

and used in the model's initial training and validation. This decision was motivated by the high cost of expert annotation, since detecting the subtle acoustic signatures of faults requires careful listening and technical judgment by experts. In addition, the aim of this stage of the work was to validate the viability of the approach in a controlled scenario with high-quality data before scaling up to a more extensive annotation effort. Therefore, the initial sample, although small, was sufficient to provide robust empirical evidence of the efficacy of the proposed method, and an expansion of the annotated base is considered a natural future step to increase the generalization of the model.

The significant class imbalance present in the dataset is a direct consequence of the data acquisition methodology. The dataset was populated from turbines with known defects. Consequently, while occasional normal operation samples are present, the data collection process inherently favors the procurement of defective samples, resulting in a skewed distribution.

4.3 Pre-processing

Audio samples underwent standardized pre-processing before model training. Normalization procedures ensured uniform duration, sampling rate, and mono-channel configuration across all recordings. Each segment was transformed into a mel-spectrogram representation (Figure 2) with 64 mel bands and a 1024-sample FFT window. This time-frequency representation provides a two-dimensional input suitable for convolutional neural network processing while preserving relevant acoustic features for fault detection.

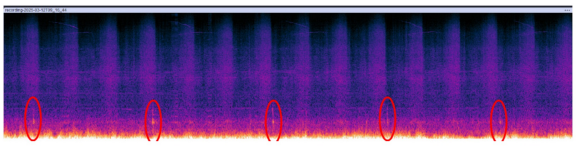

Figure 2: Defect noise

4.4 Model Building

With a solid dataset structured and labeled, we began the process of building and implementing a deep learning model that would be able to correctly identify the patterns that occur in turbine failures. Following an in-depth analysis of our requirements, the classification architecture emerged as the most suitable solution for our challenge. We opted for this supervised learning model because of the well-defined nature of the defect we aim to identify, which allows us to focus the model's intelligence on a single, specific task: distinguishing between a normal and a faulty conditions. By simplifying the problem to a choice between two classes, we ensure that the solution is efficient, accurate, and capable of providing binary and conclusive answers, which are crucial for effective decision-making. To achieve this, we implemented a Convolutional Neural Network (CNN) with a binary output. This architecture provides a straightforward yet powerful solution for distinguishing between normal operation and failure states based on audio spectrograms. The core design of our CNN was adapted from the architecture described by Ketan D. (2021)[1] in his work on audio classification, tailoring it to our specific needs.

The model is a sequential architecture, i.e., the data flows through a series of layers in a specific order. The input is a 2D representation of the audio (mel-spectrogram), that passes through 4 core convolutional blocks [1], each designed to extract increasingly complex features from the input audio data. Each block consists of 3 main components: the 2d convolutional layer itself (conv2d), which applies a filter over the input to produce a feature map, the rectified linear unit (ReLU) [25], which is an activation function that introduces non-linearity into the network, allowing it to learn more complex patterns, and the batch normalization layer (BatchNorm2d) [16] which normalizes the activations of the previous layer, helping to stabilize and speed up the training process. It is important to note that all convolutional layers, shown in Table 3, use Kaiming Normal initialization [13] for their weights and initialize their biases to zero.

Layer	Description
conv1	The first convolutional layer. It takes in 1 input channel and outputs 8 channels. Kernel: 5×5, Stride: 2×2
conv2	The second layer takes the 8 channels from the previous layer and outputs 16 channels. Kernel: 3×3, Stride: 2×2
conv3	The third layer takes the 16 channels and outputs 32 channels. Kernel: 3×3, Stride: 2×2
conv4	The final convolutional layer takes the 32 channels and outputs 64 channels. Kernel: 3×3, Stride: 2×2

Table 3: Description of convolutional layers

In addition, it is important to note that after the convolutional blocks, there are final classification layers. The global average pooling layer (GAP) [2], takes the average value of each feature map and reduces its dimensions to a single value. In this case, it converts the output from a 2D feature map to a size of 1x1 for each channel. This is an efficient way to flatten the feature maps while preserving spatial information. After that, the output from GAP passes through a flattening process, where the tensor is reshaped into a 2D tensor, preparing it for the fully connected layer.

Then, data is conducted to a dropout layer [3] that is applied with a probability of 0.5. This means that during training, 50% of the input features will be randomly set to zero. This is a regularization technique that helps prevent the model from overfitting to the training data.

The final layer is a linear (fully connected) layer. It takes the 64 features from the previous layers and maps them to a final output of 2 features. These two output features represent the model's prediction for the audio sample, likely corresponding to two distinct classes (normal and abnormal).

[1]https://medium.com/data-science/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5

4.5 Model Training

As previously discussed, the decision to rely on a fully supervised learning approach, supported by manually labeled data, was motivated by the need for a reliable and validated ground truth specific to the problem domain. Having established this methodological direction, we proceeded to implement a systematic and rigorous training process, aimed at ensuring the highest possible quality of the model learning.

As shown in the section 4.2, approximately 350 audio samples were carefully annotated by domain experts. This process involved a thorough auditory and spectral analysis of each recording to accurately identify and classify the subtle acoustic signatures associated with bearing defects. From this curated dataset, approximately 100 samples were selected for training purposes. This specific number was determined empirically, as preliminary experiments indicated that increasing the size of the training set beyond this point yielded similar returns, with model performance metrics converging and showing no statistically significant improvement. Furthermore, constraining the training set size served as a regularizing measure, mitigating the risk of overfitting to spurious correlations or dataset-specific noise that a larger, highly imbalanced set might introduce. At the same time, the remainder served as an independent validation set, enabling us to evaluate the model's generalization capacity on previously unseen data.

Although the resulting dataset was modest in size, the high degree of precision and consistency in its labeling proved to be a crucial factor in the model's effectiveness. In particular, the curated dataset enabled the detection algorithm to learn the nuanced faint acoustic patterns that characterize real defect conditions, while avoiding the pitfalls of overfitting to irrelevant noise or artifacts.

To mitigate the risk of overfitting, which is an inherent challenge when training on relatively small datasets, the training procedure incorporated an early stopping mechanism. This mechanism continuously monitored the validation loss over the course of training and halted the learning process if no improvement in validation performance was observed over a predetermined number of epochs. This safeguard ensured that the model's learning remained generalizable, preventing it from memorizing idiosyncrasies of the training data at the expense of performance on new data.

Furthermore, the model was trained with 15 complete epochs on the training set, using the CrossEntropyLoss function [22] to handle the binary classification task. Optimization was carried out using the Adam algorithm, with an initial learning rate set at 0.001 and L2 regularization (weight decay) with a factor of 1e-4, in order to mitigate the risk of overfitting. To improve learning stability, a dynamic learning rate scheduler (OneCycleLR) was used, with a linear annealing strategy across epochs. This combination of loss function, adaptive optimizer, explicit regularization and scheduler made it possible to achieve a good convergence rate and robust performance even in a limited number of epochs and with a relatively small data set.

4.6 Classification Pipeline

With the model trained and validated, a classification pipeline was developed to process audio data periodically and generate comprehensive fault reports. This pipeline is engineered for robustness, utilizing two parallel instances of the trained model to classify each audio file. For each file, a total of 30 classifications are made by each model instance. This prediction count threshold was determined empirically through extensive experimentation to mitigate statistical noise and enhance prediction reliability.

To ensure the integrity of the results and minimize false positives, a set of minimum confidence thresholds was established based on the consistency between the predictions of the two model instances. For consistent results, where both instances agree on the classification, we require a minimum average confidence of 55%. In cases of inconsistent results, where the instances diverge, a higher minimum average confidence of 65% is required to report a classification. In cases where these confidence levels are not met, the audio is classified as inconclusive. This data-driven approach to parameter tuning, which sought to optimize detection efficacy while minimizing computational overhead and latency, was essential for creating a system that is both robust and efficient for practical deployment in operational settings.

5 RESULTS AND DISCUSSION

With the model architecture and curated dataset established, we proceeded to evaluate the effectiveness of the proposed wind turbine bearing monitoring system. This section presents the experimental results and discusses the model's capability to distinguish between normal and defective operational states. The analysis is framed within the broader context of wind turbine condition monitoring, emphasizing the implications of the proposed approach for predictive maintenance and asset reliability.

To ensure a comprehensive understanding of the data and to conduct the experiments systematically, the evaluation process was divided into two distinct stages. The first stage consisted of an initial assessment using a single instance of the model during the training phase, with a reduced dataset, aimed solely at verifying the model's ability to learn and capture the underlying patterns. The second stage involved a full evaluation of the complete prediction pipeline, applying the model to the entire dataset and computing all the performance metrics previously described.

Even at this preliminary stage, the model demonstrated strong predictive performance, achieving an accuracy of 83.33%, an F1-score of 87.94%, an AUC-ROC of 95% (Figure 3), and an area under the Precision–Recall Curve of 98% (Figure 4). These results highlight the model's robustness and its capacity to maintain high discriminative power, even under moderate levels of noise and variability in the acoustic inputs.

The initial evaluation of the classification model demonstrated promising results during the experimental validation phase. The model achieved 11 true positives (TP) and 4 true negatives (TN), with only 1 false positive (FP) and 2 false negatives (FN), reflecting relatively low error rates. These outcomes indicate an effective ability to discriminate between the two target classes in a controlled experimental setting, providing confidence in the model's fundamental predictive capability.

The complete classification pipeline was implemented and evaluated in its final deployment configuration (second analytical layer) using the full dataset. Performance analysis yielded 133 true positives (TP) and 61 true negatives (TN), with 30 false positives (FP)

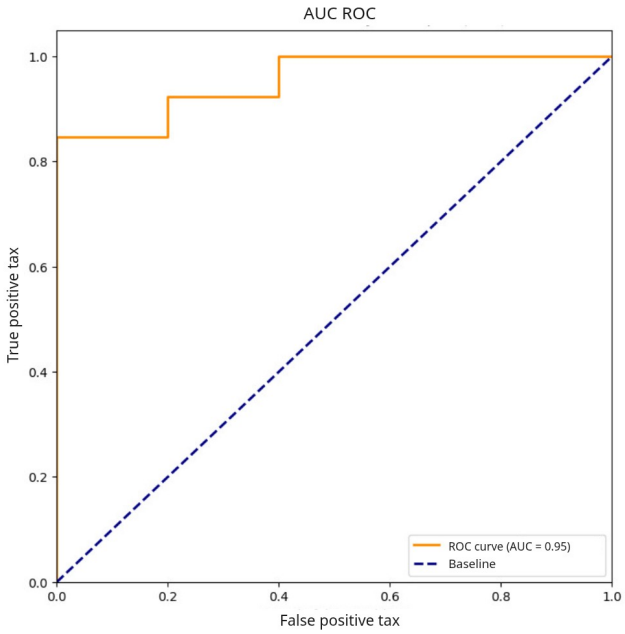

Figure 3: AUC - ROC

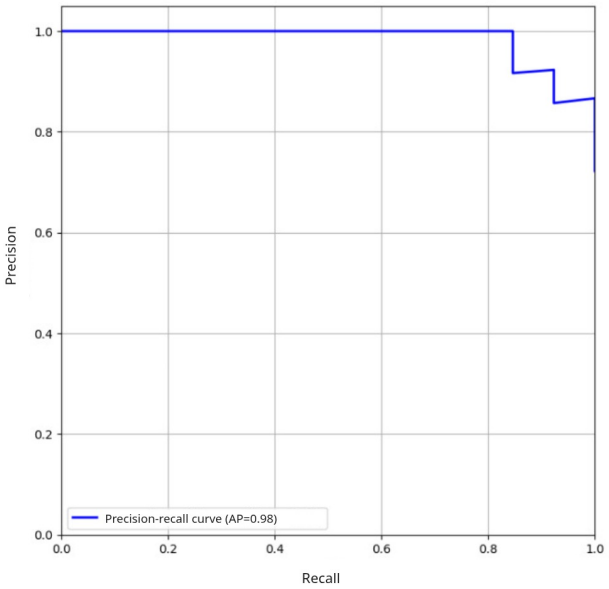

Figure 4: Precision-recall curve

and 8 false negatives (FN). These results, presented in the confusion matrix (Figure 5), demonstrate effective discriminative capability and balanced detection across both classes. The notably low false negative rate indicates particularly robust sensitivity, which is advantageous for applications where missing positive cases carries greater consequences than false alarms.

In this scenario, 6 audio samples (2.5% of the dataset) were classified as inconclusive, primarily due to low signal-to-noise ratios or

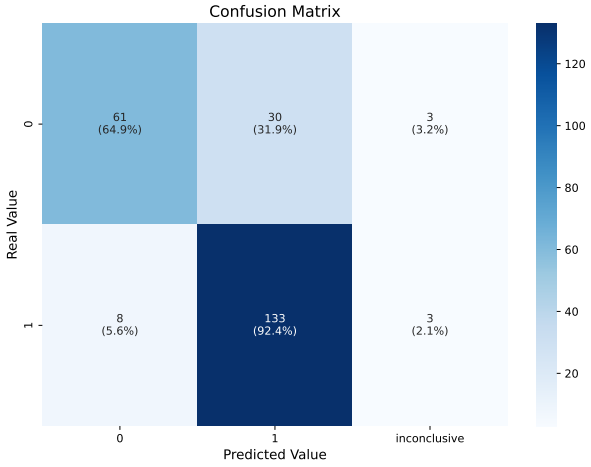

Figure 5: Confusion matrix from complete pipeline

ambiguous acoustic patterns. These samples were excluded from traditional binary metric calculations.

Based on the validated data, the final performance metrics of the deployed pipeline were as follows, shown in Table 4

Metric	Result
Accuracy	83.62%
Precision	81.59%
Recall	94.32%
F1-score	87.40%
Specificity	67%
Balanced Accuracy	80.70%
MCC	65.50%
Informedness	61.32%
Markedness	70%

Table 4: Results

The final evaluation of the classification system yielded an accuracy of approximately 83.6%, indicating that the majority of predictions aligned with the ground truth. While accuracy offers an initial overview of model performance, its interpretation in fault detection scenarios must be nuanced, particularly given class imbalance.

The precision score of approximately 81.6% reveals that the vast majority of positive classifications were indeed correct. This implies a low rate of false alarms, which is a critical property for deployment in industrial environments, where unnecessary maintenance actions can lead to significant operational costs. In parallel, the recall score reached a notably high value of 94.32%, emphasizing the effectiveness of the solution in capturing true fault occurrences.

The F1-score reached 87.40%, demonstrating a favorable balance between precision and recall. In binary fault detection context, the F1-score quantifies the trade-off between correctly identifying actual fault conditions (recall) and avoiding false alarms by misclassifying normal operations as faults (precision). This high F1-score

indicates that the model performs well on both dimensions simultaneously, ensuring reliable classification performance by neither overlooking true faults nor generating excessive false positives.

In complement to recall, the specificity score was computed and yielded a value of 67%, indicating its ability to correctly identify non-fault conditions. While lower than sensitivity, this result remains acceptable in fault detection applications, where prioritizing the identification of true faults often justifies a moderate trade-off in false alarms. The balanced accuracy, which averages sensitivity and specificity, reached 80.70%. This metric provides a fairer assessment of overall discriminative capability by accounting for performance on both classes equally, making it particularly valuable for imbalanced classification tasks.

To account for all confusion matrix components, the Matthews Correlation Coefficient (MCC) was calculated, resulting in a value of 65.5%. The observed value indicates a moderate-to-strong positive correlation between predicted and true labels, reinforcing the model's suitability for deployment in a real-world detection system.

The informedness and markedness metrics provide complementary perspectives on decision quality. Informedness (61.32%) reflects the probability that predictions exceed chance performance, integrating sensitivity and specificity into a prevalence-independent score. Markedness (70%) quantifies prediction reliability by combining precision and negative predictive value (NPV). Together, these metrics validate the system's informativeness and reliability for operational fault diagnosis.

These results confirm the model's effectiveness in detecting bearing-related faults, while maintaining a relatively low false positive rate, an important aspect for real-world deployment, as it reduces unnecessary inspections and operational costs. The high recall ensures that critical faults are not missed, while the high F1-score reflects a solid balance between detection sensitivity and prediction precision.

Inconclusive classifications accounted for only a small portion of the dataset and did not significantly affect the overall system performance. Most conclusive predictions achieved high confidence scores (above 0.65), supporting the model's reliability under real operational conditions.

6 THREATS TO VALIDITY

This section discusses potential threats to the validity of this study's results and conclusions, categorized following established research guidelines.

Internal Validity: The primary internal threat stems from the significant class imbalance in the dataset, which increases the risk of models learning biased patterns rather than genuine defect signatures. While regularization techniques (dropout, early stopping, weight decay) were employed to mitigate overfitting, the limited proportion of labeled data (2.88 hours out of 381 collected) remains a constraint. Environmental variations were addressed through normalization, though residual confounding factors may persist.

External Validity: The generalizability of findings is limited by dataset characteristics, including collection from a limited number of similar turbines and reliance solely on acoustic data. Future integration of multi-modal sensing (vibration, temperature) could enhance applicability across diverse operational scenarios and turbine types.

Construct Validity: The binary fault classification (normal or defective) may not adequately capture gradual degradation processes. Additionally, the empirically determined confidence thresholds (55%, 65%) and the focus on fault detection without severity estimation represent construct limitations for practical maintenance applications.

Conclusion Validity: The modest dataset size and inherent class imbalance limit the statistical robustness of conclusions. Although multiple complementary metrics (F1-score, MCC, Informedness) were used, results should be interpreted as preliminary feasibility indicators rather than definitive benchmarks, necessitating validation on larger, more balanced datasets.

7 CONCLUSION

The results from our evaluation demonstrate that the proposed model is highly effective at distinguishing between audio signals with and without fault-indicating noises. The model achieved a robust performance with an accuracy of 83%, an AUC ROC of 95%, an F1-score of 87%, and a precision-recall curve of 98%. While some classifications were inconclusive, our analysis shows that these instances do not compromise the overall integrity and long-term application of the system. The high confidence and accuracy of the conclusive classifications outweigh the instances deemed inconclusive, ensuring the system remains a reliable asset for fault detection. Furthermore, the model proved to be exceptionally robust when tested with data from enriched audio sources, demonstrating its resilience to variations in microphones, background noise, and adverse environmental scenarios. In summary, the developed system provides a mature and functional solution for wind turbine bearing monitoring. The combination of a robust model and the use of a representative dataset makes it a viable and effective tool for enhancing predictive maintenance and improving asset reliability. As a direction for future work, it is worth highlighting the potential of further exploring semi-supervised learning approaches in this context. Such methods would enable the exploitation of large volumes of unlabeled data, reducing the dependency on exhaustive manual labeling. To complement this and further reinforce the robustness of the findings, a comprehensive k-fold cross-validation study should be undertaken. This would provide a more reliable estimate of model performance and generalization by mitigating the variance associated with a single data split and reducing the risk of overfitting. Together, these strategies could lead to models with improved generalization, stability, and specialization across diverse operational scenarios.

ACKNOWLEDGMENTS

The authors thank the Instituto Atlântico for its essential research support, particularly for providing the infrastructure and data that made this study possible.

REFERENCES

[1] Arohan Ajit, Koustav Acharya, and Abhishek Samanta. 2020. A review of convolutional neural networks. In *2020 International Conference on Emerging Trends In Information Technology And Engineering (ic-ETITE)*. IEEE, 1–5.

[2] Aiman Al-Sabaawi, Hassan M Ibrahim, Zinah Mohsin Arkah, Muthana Al-Amidie, and Laith Alzubaidi. 2020. Amended convolutional neural network with global average pooling for image classification. In *International Conference on Intelligent Systems Design and Applications*. Springer, 171–180.

[3] Pierre Baldi and Peter J Sadowski. 2013. Understanding dropout. *Advances in neural information processing systems* 26 (2013).

[4] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* 21, 1 (2020), 6.

[5] Yan Martins B Gurevitz Cunha, José Matheus C Boaro, Daniel de Sousa Moraes, Pedro Cutrim dos Santos, Polyana Bezerra da Costa, Antonio José Grandson Busson, Julio Cesar Duarte, and Sérgio Colcher. 2024. An Ensemble Approach to Facial Deepfake Detection Using Self-Supervised Features. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC, 37–44.

[6] Shaveta Dargan, Munish Kumar, Maruthi Rohit Ayyagari, and Gulshan Kumar. 2020. A survey of deep learning and its applications: a new paradigm to machine learning. *Archives of computational methods in engineering* 27, 4 (2020), 1071–1092.

[7] Rajvikram Madurai Elavarasan, Rishi Pugazhendhi, Muhammad Irfan, Lucian Mihet-Popa, Irfan Ahmad Khan, and Pietro Elia Campana. 2022. State-of-the-art sustainable approaches for deeper decarbonization in Europe–An endowment to climate neutral vision. *Renewable and Sustainable Energy Reviews* 159 (2022), 112204.

[8] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[9] Mathaus Ferreira da Silva, Juliano Emir Nunes Masson, Murillo Ferreira dos Santos, William Rodrigues Silva, Iuri Wladimir Molina, and Gabriel Miguel Castro Martins. 2025. Audible Noise Evaluation in Wind Turbines Through Artificial Intelligence Techniques. *Sensors* 25, 5 (2025), 1492.

[10] Peter Flach. 2012. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press.

[11] Zafar Hameed, Young-Sun Hong, YM Cho, SH Ahn, and Chul Ki Song. 2009. Condition monitoring and fault detection of wind turbines and related algorithms: A review. *Renewable and Sustainable energy reviews* 13, 1 (2009), 1–39.

[12] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*. 1026–1034.

[14] Jürgen Herp, Mohammad H Ramezani, Martin Bach-Andersen, Niels L Pedersen, and Esmaeil S Nadimi. 2018. Bayesian state prediction of wind turbine bearing failure. *Renewable Energy* 116 (2018), 164–172.

[15] Maximilian Hoh, Alfred Schöttl, Henry Schaub, and Franz Wenninger. 2022. A generative model for anomaly detection in time series data. *Procedia Computer Science* 200 (2022), 629–637.

[16] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*. pmlr, 448–456.

[17] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 245–251.

[18] Baljinder Kaur and Jaskirat Singh. 2021. Audio classification: Environmental sounds classification. (2021). https://hal.science/hal-03501143/document

[19] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review* 53, 8 (2020), 5455–5516.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[21] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. 1989. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems* 2 (1989).

[22] Li Li, Miloš Doroslovački, and Murray H Loew. 2020. Approximating the gradient of cross-entropy loss function. *IEEE access* 8 (2020), 111626–111635.

[23] Xiaofan Lin, Cong Zhao, and Wei Pan. 2017. Towards accurate binary convolutional neural network. *Advances in neural information processing systems* 30 (2017).

[24] Majid Memari, Praveen Shakya, Mohammad Shekaramiz, Abdennour C Seibi, and Mohammad AS Masoum. 2024. Review on the advancements in wind turbine blade inspection: Integrating drone and deep learning technologies for enhanced defect detection. *IEEE Access* (2024).

[25] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 807–814.

[26] Vladimir Nasteski. 2017. An overview of the supervised machine learning methods. *Horizons. b* 4, 51-62 (2017), 56.

[27] Otacílio de A Ramos Neto, Rafael C Chaves, Alysson P Nascimento, and Ruan D Gomes. 2024. Middleware para aplicações distribuídas de vídeo com suporte à computação na borda na Indústria 4.0. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC, 215–222.

[28] Han Peng, Hai Zhang, Yisa Fan, Linjian Shangguan, and Yang Yang. 2022. A review of research on wind turbine bearings' failure analysis and fault diagnosis. *Lubricants* 11, 1 (2022), 14.

[29] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).

[30] Wei Qiao and Dingguo Lu. 2015. A survey on wind turbine condition monitoring and fault diagnosis—Part II: Signals and signal processing methods. *IEEE Transactions on Industrial Electronics* 62, 10 (2015), 6546–6557.

[31] Christine Röckmann, Sander Lagerveld, and John Stavenuiter. 2017. Operation and maintenance costs of offshore wind farms and potential multi-use platforms in the Dutch North Sea. In *Aquaculture Perspective of Multi-Use Sites in the Open Ocean: The Untapped Potential for Marine Resources in the Anthropocene*. Springer International Publishing Cham, 97–113.

[32] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10, 3 (2015), e0118432.

[33] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management* 45, 4 (2009), 427–437.

[34] Bouno Toshio, Yuji Toshifumi, Hamada Tsugio, and Hideaki Toya. 2005. Failure forecast diagnosis of small wind turbine using acoustic emission sensor. *KIEE International Transaction on Electrical Machinery and Energy Conversion Systems* 5, 1 (2005), 78–83.

[35] Xin Wang, Dongxing Mao, and Xiaodong Li. 2021. Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. *Measurement* 173 (2021), 108518.

[36] Khalid Zaman, Melike Sah, Cem Direkoglu, and Masashi Unoki. 2023. A survey of audio classification using deep learning. *IEEE Access* 11 (2023), 106620–106649.

[37] Pinjia Zhang and Delong Lu. 2019. A survey of condition monitoring and fault diagnosis toward integrated O&M for wind turbines. *Energies* 12, 14 (2019), 2801.

[38] Yang Zhang, Maciej Radzieński, Sidney Xue, Ziping Wang, and Wiesław Ostachowicz. 2024. A method for detecting cracks on the trailing edge of wind turbine blades based on aeroacoustic noise analysis. *Structural Health Monitoring* (2024), 14759217241303452.