Supervised Prediction of Remaining Useful Life of Drills from Acoustic Signals

Raphael Barbosa[1], **Sergio Chevtchenko**[2], **Saeed Afshar**[2], **Naqib Ibnul**[2], **Gustavo Callou**[1], **Ermeson Andrade**[1]

[1] Universidade Federal Rural de Pernambuco (UFRPE) – Recife – PE – Brasil

[2]Western Sydney University – Penrith – NSW – Australia

{raphael.bmelo, gustavo.callou, ermeson.andrade}@ufrpe.br

{s.chevtchenko, s.afshar, n.ibnul}@westernsydney.edu.au

Abstract. *Early failure detection in cutting tools is a critical challenge in automated manufacturing environments, where unplanned interruptions lead to significant operational costs. This study investigates the supervised prediction of imminent drill failures using acoustic signals collected during machining operations. Ultrasonic recordings obtained from microphones positioned inside a Computer Numerical Control (CNC) machine are segmented per hole and labeled according to failure occurrence within different prediction horizons (fail_in_1, fail_in_3, fail_in_5), indicating failures occurring within 1, 3, and 5 future drilling operations, respectively. Two approaches are evaluated: classical machine learning using statistical and spectral features with a Random Forest classifier, and deep learning using Convolutional Neural Networks (CNN) applied to log-Mel spectrograms. To ensure realistic generalization, data splitting is performed at the tool level, preventing leakage between training and test sets. Results show that acoustic signals exhibit discriminative patterns for early failure detection. The CNN achieves higher F1-scores across all horizons, while the Random Forest provides more stable performance and higher AUC values, highlighting the trade-off between sensitivity and robustness in predictive maintenance systems.*

1. Introduction

Early failure detection in cutting tools is a central problem in modern manufacturing systems, especially in automated production scenarios, where unplanned interruptions result in operational losses, increased costs, and degradation of final product quality [10, 3]. In the context of drilling processes, the progressive wear of drills can evolve into sudden failures, such as breakage or jamming, compromising both the machined part and the equipment involved. These issues are particularly critical in CNC machining environments, where operations are highly automated and continuous, amplifying the impact of unexpected tool failures on production efficiency and system reliability. This problem can also be interpreted within the context of Cyber-Physical Systems (CPS), where physical processes are monitored and analyzed through sensor data and computational models, enabling intelligent decision-making in industrial environments.

Traditionally, tool condition monitoring is performed using physical sensors directly coupled to the system, such as force, vibration, or electric current sensors. Al-

though effective, these methods present limitations related to instrumentation cost, intrusiveness, and scalability in complex industrial environments [7, 3]. In this context, the use of acoustic signals emerges as a non-invasive alternative, capable of capturing relevant information about the interaction between tool and material during the machining process. Recent reviews indicate that data-driven monitoring, including the use of acoustic sensors, has become one of the main research directions in the field [8].

Acoustic signals, especially in the ultrasonic range, contain components sensitive to variations in the cutting regime, enabling the identification of changes associated with wear and imminent failure. Recent studies demonstrate that acoustic emission sensors can be used for real-time monitoring, showing high sensitivity to events related to degradation and failures in machining processes [13, 4]. In addition, industrial applications show that these signals allow wear detection under real operating conditions, reinforcing their practical feasibility [11]. This capability makes acoustic sensing particularly suitable for non-invasive monitoring in industrial environments, where minimal interference with the machining process is required.

Classical models such as Random Forest can capture nonlinear relationships from statistical and spectral features extracted from signals. On the other hand, deep learning approaches such as convolutional neural networks allow automatic learning of representations directly from data in the time-frequency domain. Recent works show that CNNs applied to acoustic signals or representations such as Mel spectrograms achieve high performance in monitoring and diagnostic tasks, reducing dependence on manual feature engineering and improving generalization capability [13, 6, 1]. This approach is particularly relevant in industrial scenarios, where signals exhibit high variability and noise.

Despite these advances, direct prediction of Remaining Useful Life (RUL) remains challenging due to variability in machining processes and the presence of noise in the data. This limitation has been discussed in recent reviews, which highlight the need for more robust modeling, validation, and generalization strategies in real applications [8]. In this context, approaches based on classification of future failure in discrete horizons emerge as a more robust alternative, allowing greater practical applicability. This formulation also enables a more direct interpretation of model outputs in terms of failure events, simplifying their use in practical monitoring and maintenance scenarios.

In this context, this work investigates the prediction of imminent drill failures using ultrasonic acoustic signals collected from real machining experiments, formulating the problem as a binary classification task across multiple prediction horizons. Specifically, the labels fail_in_1, fail_in_3, and fail_in_5 are defined, indicating the occurrence of failure within 1, 3, and 5 future drilling operations, respectively, allowing control over the level of anticipation in detection. This formulation enhances practical applicability in predictive maintenance systems, where different decision horizons imply distinct intervention strategies. Two complementary approaches are investigated: (i) a Random Forest model using statistical and spectral features extracted from acoustic signals, and (ii) a CNN model applied to log-scale Mel spectrograms. These models were selected to compare a feature engineering-based approach, known for its robustness and interpretability, with a deep learning-based approach capable of automatically extracting discriminative representations from data. The results show that acoustic signals exhibit consistent discriminative patterns for failure anticipation, with clear differences between models in

terms of sensitivity and stability, highlighting direct implications for their adoption in industrial environments. The main contributions of this work are:

- A non-invasive approach based on ultrasonic acoustic signals for cutting tool monitoring;
- A failure classification formulation across multiple prediction horizons (fail_in_1, fail_in_3, and fail_in_5);
- A comparison between Random Forest and CNN models applied to acoustic data;
- An analysis of the trade-off between sensitivity (recall) and robustness (precision and AUC);
- A tool-level evaluation protocol to assess model generalization (drill_id split).

The remainder of this article is organized as follows. Section 2 presents related work on tool monitoring and the use of acoustic signals in predictive maintenance. Section 3 describes the theoretical background, addressing concepts of failure prognosis, remaining useful life, and machine learning techniques applied to the problem. Section 4 details the experimental methodology, including dataset, signal acquisition and segmentation, feature extraction, label definition, and models used. Section 5 presents experimental results and discussion. Finally, Section 6 presents conclusions and directions for future work.

2. Related Work

The literature on tool condition monitoring in machining processes has evolved significantly, driven by advances in smart sensors and machine learning techniques. Recent reviews indicate increasing adoption of data-driven approaches and integration of multiple information sources for fault diagnosis and prognosis [8, 5].

Historically, tool monitoring has been performed using intrusive sensors such as dynamometers, accelerometers, and current sensors, which measure physical quantities of the machining process. Although effective, these approaches present practical limitations, including high cost, installation complexity, and interference with the production process, motivating the search for non-invasive alternatives. In this context, the use of acoustic signals, particularly acoustic emission, has gained prominence. These signals capture information about tool-material contact and are sensitive to high-frequency events associated with wear. Recent studies show that acoustic emission can be used for real-time diagnosis, enabling identification of patterns associated with tool degradation in machining processes [13]. Industrial applications also demonstrate that these signals enable wear detection under real operating conditions, reinforcing their practical viability [11].

Machine learning has become one of the main tools for failure prediction and wear estimation in machining processes. Traditional feature engineering approaches explore descriptors in time, frequency, and time-frequency domains, serving as input for models such as Random Forest and Support Vector Machines. However, with advances in deep learning, there has been a transition toward models capable of learning directly from raw data or representations such as spectrograms. Convolutional neural networks have been widely applied to tool monitoring, especially when signals are transformed into two-dimensional representations. Recent works show that acoustic signal spectrograms can be used as input to CNNs, enabling high-performance classification of wear states [6, 1]. In addition, hybrid deep learning models applied to acoustic emission signals

have demonstrated high accuracy in fault detection in machining machines, combining automatic feature extraction with optimization techniques [13].

Despite these advances, the literature still presents important limitations. Many studies are conducted in controlled environments with limited datasets, compromising model generalization. Moreover, direct modeling of RUL can be sensitive to noise and uncertainty, making application in real industrial scenarios difficult. Another relevant challenge concerns variability in acoustic signals, which can be influenced by operating conditions and external noise. These challenges have been highlighted in recent reviews [8], emphasizing the need for more rigorous validation protocols and greater diversity of experimental data.

Given these limitations, this work investigates an approach based on ultrasonic acoustic signals collected non-invasively in real machining experiments, directly addressing the gap identified in the literature regarding the scarcity of studies using representative real-world data. The formulation of the problem as classification of future failure across multiple horizons provides greater robustness to data variability and better alignment with practical predictive maintenance applications. Additionally, the work compares classical and deep approaches, aligning with recent trends and contributing to understanding the trade-off between interpretability, stability, and predictive capability. Table 1 summarizes representative related works and highlights the distinctions of the proposed approach in terms of signal type, modeling strategy, datasets, evaluation metrics, and prediction objective.

Work	**Signal Type**	**Approach**	**Representation**	**Objective**	**Dataset**	**Validation / Metrics**
TANG (2019)	Acoustic emission	Machine Learning	Features (time/freq.)	Wear monitoring	Experimental	Acc.
KIM (2020)	Acoustic emission	Deep Learning	Spectrograms	Wear monitoring	Laboratory	Acc., F1
UMAR (2024)	Acoustic emission	Hybrid DL	Features + Deep	Fault diagnosis	Experimental	Acc., Rec.
FERRISI (2024)	Acoustic signals	CNN	Spectrograms	Condition monitoring	Experimental	Acc., AUC
BARBOSH (2024)	Acoustic waves	Deep Learning	Waveforms / spectra	Damage detection	Experimental	Acc.
This work	**Ultrasonic acoustic**	**RF + CNN**	**Features + Mel spectrogram**	**Failure prediction (multi-horizon)**	**Real CNC**	**Acc., Prec., Rec., F1, AUC**

Table 1. Comparison of representative related works and the proposed approach in terms of signal type, modeling strategy, datasets, and evaluation metrics. Acc. = Accuracy; Prec. = Precision; Rec. = Recall; F1 = F1-score; AUC = Area Under the Curve.

3. Theoretical Foundations of Tool Monitoring and Failure Prediction

Predictive maintenance aims to anticipate failures through analysis of operational data, enabling more efficient interventions than corrective or fixed-interval preventive strategies [7]. In machining processes, this approach is particularly relevant due to the direct impact of tool wear on product quality, equipment integrity, and production continuity. In this context, continuous monitoring of tool condition becomes central to decision-making, requiring models capable of capturing degradation patterns over time.

One of the most common formulations of this problem is RUL estimation, defined as the number of cycles or time remaining until failure [12]. However, in real machining scenarios, direct RUL modeling is hindered by variability between tools, uncontrolled operating conditions, and noise in the signals. These characteristics make continuous regression sensitive to uncertainty and reduce robustness. As an alternative, modeling based on classification of future failure in discrete horizons transforms the problem into binary decisions associated with specific time windows, favoring model stability and interpretability in operational terms.

Tool condition monitoring can be performed using different types of sensors, such as force, vibration, electric current, and acoustic emission. Among these, acoustic signals offer the advantage of being non-invasive and capturing phenomena directly related to tool-material interaction. Acoustic emission corresponds to elastic waves generated by deformation and fracture processes, being sensitive to events associated with tool wear and failure [13, 11]. In particular, ultrasonic signals enable identification of variations in the cutting regime and detection of subtle changes in process behavior, constituting a relevant source of information for diagnostic and prognostic tasks.

Analysis of these signals can be performed using representations in the time, frequency, and time–frequency domains, enabling the characterization of both global properties and local patterns. In this context, machine learning methods are employed to map these representations to tool condition. Feature-based models such as Random Forest can capture nonlinear relationships with good robustness and interpretability [2], making them suitable for scenarios with limited data and where model transparency is required. In contrast, deep learning approaches such as convolutional neural networks enable the direct learning of representations from data, particularly when applied to spectrograms [9], and are more appropriate in situations where complex patterns must be captured and larger datasets are available. The choice between these paradigms involves a trade-off between generalization capability, stability, and interpretability, which is particularly relevant in industrial applications.

4. Experimental Methodology

This section describes the experimental methodology adopted in this work, as illustrated in Figure 1. Initially, the dataset is presented, followed by procedures for acquisition and segmentation of acoustic signals into units corresponding to drilling operations. Next, feature extraction steps are detailed, covering representations in time, frequency, and time-frequency domains. Subsequently, the process of RUL definition and data labeling across different prediction horizons is described. Finally, the learning models and evaluation procedure used for performance analysis are presented.

4.1. Dataset

The dataset used in this study was obtained from real drilling experiments conducted in a machining environment instrumented for acoustic signal acquisition, as illustrated in Figure 2. The figure depicts the lifecycle of a single tool (drill_id), including the sequence of drilling operations, the evolution of RUL, and the generation of binary labels for failure prediction across different horizons (fail_in_1, fail_in_3, and fail_in_5), where colored regions indicate samples within the failure horizon. The experiments were carried

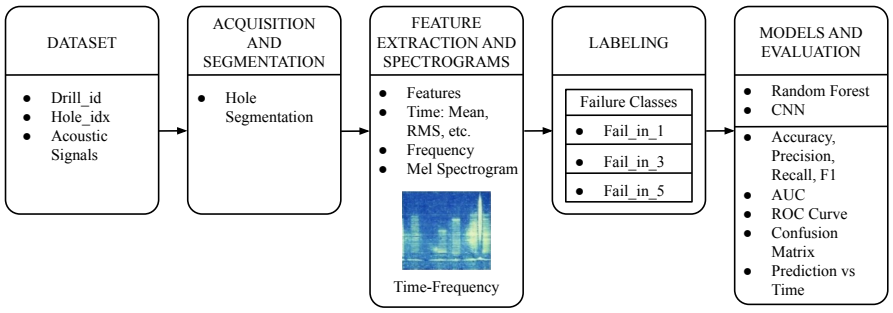

Figure 1. Flow of the proposed experimental methodology.

out on a CNC machine using drills with a nominal diameter of 4 mm. The instrumentation included conventional and ultrasonic microphones positioned in different regions of the system. In this work, exclusively the signals from ultrasonic microphones positioned inside the CNC machine, at a distance of 30 cm from the drill, were considered. Additionally, electric current signals were acquired through a datalogger system. The experimental procedure was designed to monitor the complete life cycle of the tools, from initial conditions to failure occurrence, characterized by jamming events.

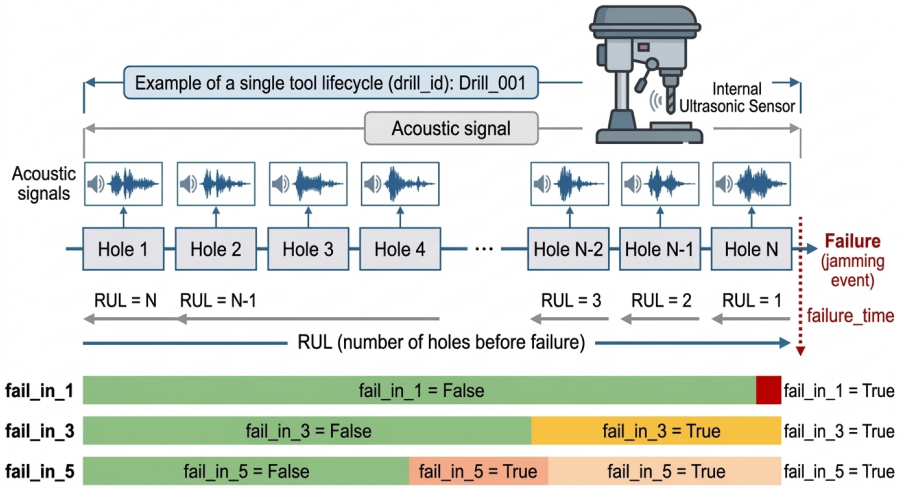

Figure 2. Dataset structure and labeling strategy for failure prediction in drilling operations.

Each drilling operation is treated as an independent sample, containing the corresponding acoustic signal and contextual information associated with the process. Tools are identified by a unique identifier (*drill_id*), while the variable *hole_idx* represents the sequential order of holes drilled by each tool. This structure allows explicit modeling of wear evolution, establishing a direct relationship between tool usage history and acoustic signal behavior over time. The final dataset consists of 20 tools and a total of 9470 drilling operations, corresponding to 9470 samples. Filtering criteria were applied to remove atypical behavior, aiming to increase consistency and robustness in modeling. Scripts and metadata used in this work are publicly available[1].

[1] `https://github.com/smugfungus/DrillCode.git`

4.2. Signal Acquisition and Segmentation

Acoustic signals were continuously acquired during drilling operations, covering multiple machining cycles throughout tool life. Acquisition was performed using sensors positioned within the machining environment, with ultrasonic internal microphones considered in this work due to their proximity to the cutting zone and higher sensitivity to wear-related variations. Additionally, ultrasonic microphones installed inside the CNC machine reduce the influence of external noise sources, since they operate in frequency ranges less affected by typical industrial noise, improving signal quality and robustness in real factory environments. This configuration enables the capture of subtle changes in the cutting regime, reflecting progressive changes in tool condition. The collected signals represent different degradation stages, including initial, intermediate, and near-failure conditions. Continuous acquisition preserves the temporal dynamics of the process, avoiding information loss associated with predefined windows and maintaining relevant characteristics for subsequent analysis and supervised modeling.

To enable supervised analysis, continuous recordings were segmented into units corresponding to individual drilling operations. Each audio segment was associated with a single hole, following the dataset structure in which each sample is identified by hole_idx. This process ensures correspondence between acoustic signals and modeling units, allowing direct association between each segment and tool condition at a specific moment in its life cycle. Segmentation was performed beforehand and provided through audio file paths (output_path). Signals are assumed to be synchronized during acquisition, allowing direct use without additional alignment.

4.3. Acoustic Feature Extraction

Feature extraction was performed to represent acoustic signals appropriately for machine learning models, considering distinct approaches for feature-based and deep learning models. For the Random Forest model, time-domain features were extracted, including mean, root mean square, standard deviation, maximum value, and kurtosis. These metrics summarize statistical signal behavior and are widely used in acoustic characterization of tools. Additionally, frequency-domain features were considered, obtained from spectral analysis, allowing identification of energy distribution across frequency bands, particularly relevant for ultrasonic signals.

For the CNN model, acoustic signals were transformed into time-frequency representations. Specifically, Mel spectrograms were used, followed by conversion to logarithmic scale, resulting in representations more suitable for pattern learning by deep models. This choice is aligned with recent studies demonstrating the effectiveness of Mel spectrograms for CNN-based acoustic monitoring tasks [1, 6]. Due to variability in signal length, padding or truncation was applied to ensure fixed input dimensions. A light normalization based on division by the maximum signal value was also applied to stabilize training while preserving relevant amplitude variations.

4.4. RUL Definition and Labeling

RUL was defined as the number of remaining holes until tool failure. For each drilling operation, the RUL value was calculated based on the distance in cycles to the first jamming event. In this work, jamming was adopted as the single criterion for terminal failure.

From this definition, supervised labels were generated based on discrete prediction horizons, converting the original regression problem into a binary classification task. Specifically, three scenarios were defined: fail_in_1, fail_in_3, and fail_in_5, indicating whether failure occurs within 1, 3, or 5 holes from the current operation.

This strategy captures different levels of failure anticipation, allowing the evaluation of model performance under increasingly permissive prediction conditions. Short horizons (e.g., fail_in_1) represent highly restrictive scenarios with minimal reaction time, where only imminent failures are considered. In contrast, longer horizons (e.g., fail_in_5) enable earlier detection, providing more time for maintenance planning and intervention. This formulation is particularly relevant in predictive maintenance, where the balance between early warning and false alarm control is critical. Additionally, framing the problem as a classification task across discrete horizons aligns with recent approaches that favor classification over continuous RUL regression in scenarios with limited data and high variability.

4.5. Models and Evaluation Procedure

Two supervised models were used for failure prediction: a Random Forest model and a convolutional neural network model. The Random Forest model was trained using acoustic features from time and frequency domains. To handle class imbalance, the parameter class_weight = "balanced" was used. Hyperparameter selection was performed using GridSearchCV with three-fold cross-validation.

The CNN model was trained using log-scale Mel spectrograms. The architecture consisted of three convolutional blocks, each followed by batch normalization, ReLU activation, and max pooling. Each convolutional block consists of convolutional layers with an increasing number of filters, allowing progressive extraction of higher-level features from the input representations. The use of pooling layers reduces spatial dimensionality while preserving relevant patterns, and batch normalization contributes to training stability. This design follows a hierarchical feature extraction strategy commonly adopted in acoustic signal analysis. The output was aggregated using adaptive average pooling and passed to a fully connected layer for binary classification. Training used Binary Cross-Entropy with Logits loss with class weighting and the Adam optimizer. Hyperparameters were optimized using a low-cost random search procedure to ensure comparability with the Random Forest model. Different combinations were explored, including learning rate ($\in 1e-3, 5e-4, 1e-4$), batch size ($\in 16, 32$), and dropout rate ($\in 0.2, 0.4$). Approximately five random configurations were evaluated using a short training phase (approximately 8 epochs), allowing rapid exploration of the search space. Model selection was based on the F1-score on the validation set. After selecting the best configuration, the model was retrained from scratch using a larger number of epochs (approximately 20–25), ensuring proper convergence. The results reported in this work correspond to this final training stage.

Data splitting was performed at the tool level, ensuring no overlap between training and test tools. Five tools were randomly selected for testing, while the remaining were used for training. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC. Class imbalance was addressed through class weighting strategies in both models (class_weight for Random Forest and pos_weight for CNN). However, additional

evaluation metrics such as Precision-Recall AUC (PR-AUC), which are particularly suitable for imbalanced datasets, were not explored and may provide further insights in future analyses. Confusion matrices, ROC curves, and prediction evolution over time were also analyzed.

5. Results and Discussion

The quantitative results obtained for the different prediction horizons are presented in Table 2, allowing comparison between the Random Forest and CNN models. In the most restrictive prediction scenario (fail_in_1), Random Forest achieved higher accuracy (0.756) and AUC (0.689), indicating better global discrimination capability between classes. However, a very low recall (0.087) is observed, indicating limited sensitivity in detecting imminent failures. The model exhibits a conservative behavior, with very high precision (1.000) and reduced ability to identify positive cases. The CNN, in turn, presented higher recall (0.2028) but lower precision (0.2456), resulting in an F1-score of 0.2222. In addition, the model achieved a lower accuracy (0.6201) compared to Random Forest, indicating a reduced overall classification performance. This behavior suggests that, although the model detects a larger number of failures, it struggles to control false positives under this restrictive prediction horizon, leading to a less reliable global performance.

Table 2. Model Performance Results.

Model	Target	Accuracy	Precision	Recall	F1-score	AUC
RF	FAIL_IN_1	**0.7558**	**1.0**	0.087	0.16	**0.6898**
CNN	FAIL_IN_1	0.6201	0.2456	**0.2028**	**0.2222**	0.4503
RF	FAIL_IN_3	**0.6434**	**0.8438**	0.2368	0.3699	**0.6033**
CNN	FAIL_IN_3	0.4418	0.4395	**0.9561**	**0.6022**	0.5435
RF	FAIL_IN_5	0.5271	**0.6667**	0.3056	0.419	**0.5733**
CNN	FAIL_IN_5	**0.5581**	0.5581	**1.0000**	**0.7164**	0.3773

For the intermediate horizon (fail_in_3), the RF achieved an F1-score of 0.392, with precision (0.844) and recall (0.238), reinforcing its conservative behavior, prioritizing correct positive predictions at the cost of missing failures. The CNN presented recall equal to 0.956 and F1-score of 0.602, demonstrating strong ability to detect failure events, albeit with lower precision. In the longer prediction horizon (fail_in_5), the CNN achieved the highest F1-score (0.716), with recall equal to 1.000, indicating strong capability for early failure detection. The Random Forest, however, presented precision of 0.887 and recall of 0.305, maintaining its conservative profile, with fewer false positives but reduced sensitivity. The results indicate a consistent pattern in which the CNN prioritizes sensitivity (high recall), while the Random Forest prioritizes precision and robustness. This distinction is particularly relevant in predictive maintenance applications, where the trade-off between missed failures and false alarms must be carefully considered depending on the operational context.

The ROC curves for the different prediction horizons are shown in Figure 3. It can be observed that model performance varies depending on the horizon considered. In the most restrictive scenario (fail_in_1), Random Forest shows higher discriminative capability, reflected in higher AUC values. In the intermediate horizon (fail_in_3), both

models present similar performance. For the longest horizon (fail_in_5), Random Forest maintains better discriminative capability, while CNN performance approaches that of a random classifier. These results indicate that Random Forest is more robust for detecting imminent failures, whereas the CNN is more effective for anticipating failures in longer horizons, albeit with lower global precision.

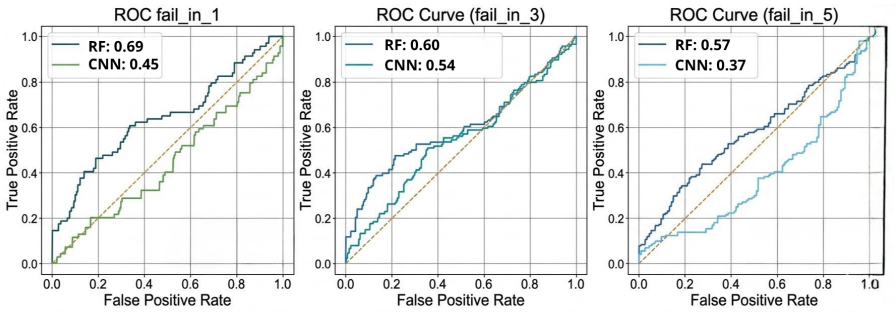

Figure 3. Side-by-Side Comparison of ROC Curves.

Figure 4 presents an example of a confusion matrix for the fail_in_5 scenario, consistent with the quantitative metrics reported in Table 2. It is observed that the CNN achieves maximum recall, correctly identifying all failure instances. However, this occurs at the cost of classifying nearly all samples as failures, resulting in a high false positive rate. This behavior indicates a bias toward the positive class, explaining the observed precision and AUC values and highlighting the trade-off between sensitivity and discriminative capability.

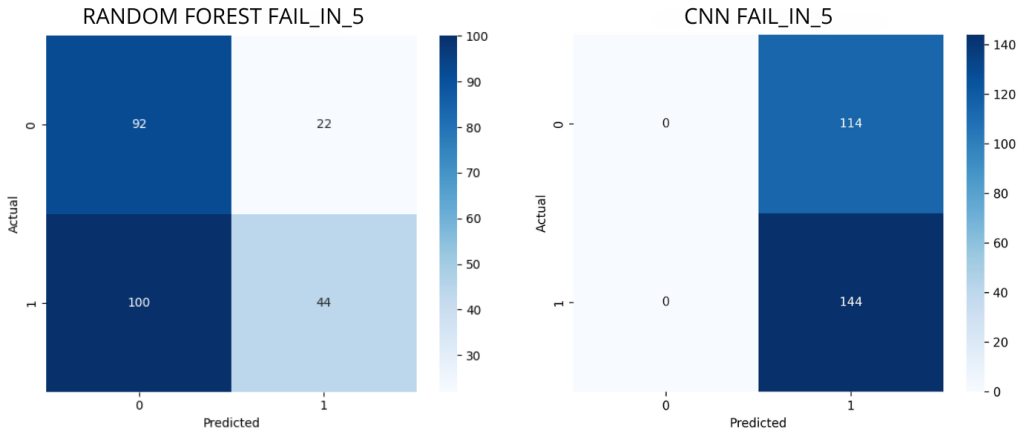

Figure 4. Confusion Matrices (fail_in_5).

The analysis of prediction behavior over time, also illustrated in Figure 5, shows that the Random Forest model exhibits a gradual increase in failure probability as the tool approaches the end of its useful life, reflecting greater robustness and alignment with progressive wear. In contrast, the CNN demonstrates higher sensitivity to failure-related patterns but with less stable behavior, presenting fluctuations in predicted probabilities over time and elevated values even in earlier stages. This pattern highlights its ability to detect early degradation signals at the cost of a higher incidence of false positives.

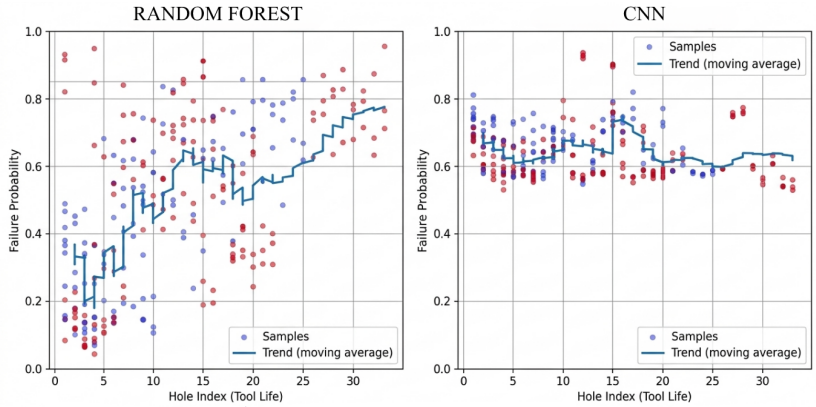

Figure 5. Failure probability evolution fail_in_5 scenario.

The obtained results are consistent with recent literature, which highlights the effectiveness of deep learning models in automatic feature extraction from acoustic signals [6, 1]. On the other hand, the greater robustness observed in Random Forest corroborates studies emphasizing the effectiveness of feature-based models in scenarios with limited datasets [2]. Despite promising results, CNN performance instability is observed, particularly reflected in lower AUC values in some scenarios, which may be associated with the limited dataset size, a well-known challenge in training deep learning models [8].

Finally, it is important to highlight that the exclusive use of internal microphones contributed to greater consistency in the results by reducing the influence of external noise. However, this choice limits comparative analysis across different sensor configurations, representing an opportunity for future work. The choice between CNN and Random Forest depends on the application context: in critical scenarios, it may be preferable to detect all failures even at the cost of false alarms, whereas in environments with high costs associated with unnecessary maintenance, the balanced behavior of Random Forest may be more suitable.

6. Conclusions

This work investigated the prediction of imminent drill failures during drilling processes through the analysis of ultrasonic acoustic signals and the application of supervised machine learning techniques. The problem was formulated as a binary classification task across multiple prediction horizons (fail_in_1, fail_in_3, and fail_in_5), allowing for the evaluation of the models' ability to anticipate failure events at different stages. The results demonstrate that acoustic signals contain relevant discriminative information for wear monitoring, making it possible to identify patterns associated with tool degradation over time. A distinct behavior was observed between the evaluated models: Random Forest showed a more conservative behavior, with higher precision and reduced sensitivity, while the CNN demonstrated higher sensitivity in failure detection at the cost of a higher false-positive rate.

The analysis of predictions over time indicated that the models capture relevant aspects of wear evolution, with a general tendency of increasing failure probability as the tool approaches the end of its useful life. Random Forest exhibited a more conservative and consistent behavior, whereas the CNN showed a tendency to anticipate failures ag-

gressively. Key limitations include the restricted dataset size, the inherent variability of the machining process, and the dependence on a specific acquisition setup, which impact the models' generalization capability. These results indicate that model selection must consider the application context, especially the trade-off between missed failures and false alarms in predictive maintenance systems.

For future work, it is suggested to expand the dataset with a greater diversity of operating conditions, as well as to investigate hybrid approaches combining manually extracted features with representations learned by neural networks. Furthermore, incorporating multiple data sources (e.g., vibration and electrical current) could contribute to the development of more robust monitoring systems. Finally, advanced validation strategies should be explored, including repeated experiments with different data splits, statistical significance tests, and confidence interval estimation, in order to better assess model stability, variability, and generalization.

Acknowledgments

The authors acknowledge the support of the National Council for Scientific and Technological Development (CNPq), Brazil, through project no. 401147/2025-8.

References

Barbosh, M., Ge, L., and Sadhu, A. (2024). Automated crack identification in structures using acoustic waveforms and deep learning. *Journal of Infrastructure Preservation and Resilience*, 5(1):10.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Byrne, G., Dornfeld, D., and Denkena, B. (2003). Advancing cutting technology. *CIRP Annals*, 52(2):483–507.

Carden, E. P. and Fanning, P. (2004). Vibration based condition monitoring: a review. *Structural health monitoring*, 3(4):355–377.

Chevtchenko, S. F., Rocha, E. D. S., Dos Santos, M. C. M., Mota, R. L., Vieira, D. M., De Andrade, E. C., and De Araújo, D. R. B. (2023). Anomaly detection in industrial machinery using iot devices and machine learning: A systematic mapping. *IEEe Access*, 11:128288–128305.

Ferrisi, S., Zangara, G., Izquierdo, D. R., Lofaro, D., Guido, R., Conforti, D., and Ambrogio, G. (2024). Tool condition monitoring for milling process using convolutional neural networks. *Procedia Computer Science*, 232:1607–1616.

Jardine, A. K., Lin, D., and Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510.

Lara de Leon, M. A., Kolarik, J., Byrtus, R., Koziorek, J., Zmij, P., and Martinek, R. (2024). Tool condition monitoring methods applicable in the metalworking process: Mal de leon et al. *Archives of computational methods in engineering*, 31(1):221–242.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., and Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to rul prediction. *Mechanical systems and signal processing*, 104:799–834.

Piorkowski, P., Roszkowski, A., and Szabla, Z. (2025). Diagnostics of milling head using acoustic emission. *Manufacturing Technology Journal*, 25(2):222–229.

Si, X.-S., Wang, W., Hu, C.-H., and Zhou, D.-H. (2011). Remaining useful life estimation–a review on the statistical data driven approaches. *European journal of operational research*, 213(1):1–14.

Umar, M., Siddique, M. F., Ullah, N., and Kim, J.-M. (2024). Milling machine fault diagnosis using acoustic emission and hybrid deep learning with feature optimization. *Applied Sciences*, 14(22):10404.