Detection of tool wear in drilling using acoustic machine learning models through labeling-threshold sensitivity analysis

Detección de desgaste de herramientas en taladrado con modelos de machine learning acústico por medio de análisis de sensibilidad al umbral de etiquetado

MSc. Jorge Enrique Meneses Flórez[1], **MSc. Nicolás Orejarena Osorio**[2], **Ing. Oscar Ivan Ayala Ortiz**[1]

[1] **Universidad Industrial de Santander,** Facultad de Físicomecánica, Escuela de Ingeniería Mecánica, Bucaramanga, Santander, Colombia.
[2] **Unidades Tecnológicas de Santander,** Facultad de Ciencias Naturales e Ingenierías, Escuela de Electromecánica, Bucaramanga, Santander, Colombia.

Correspondence: ayalaortizoscarivan@gmail.com

Received: june 09, 2026. **Revised:** august 18, 2026. **Accepted:** september 02, 2026. **Published:** september 09, 2026.

How to cite: J. E. Meneses Flórez, N. Orejarena Osorio, and O. I. Ayala Ortiz, "Detection of tool wear in drilling using acoustic machine learning models through labeling-threshold sensitivity analysis", *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 2, no. 48, pp. 251–261, sep. 2026, doi: 10.24054/rcta.v2i48.4535

Abstract: An ordinal classification system for drill bit wear in CNC drilling of AISI 4140 steel using acoustic signals captured with low-cost microphones is presented. Frank–Hall ordinal decomposition is employed with SVM and RF over 10 statistically selected acoustic features. The central contribution is a labeling threshold sensitivity analysis between 60% and 97% of tool life. Adjacent accuracy stays above 83% across the entire range, meaning errors never exceed one ordinal step. A refined SVM (linear kernel, C=5, asymmetric weights, calibrated thresholds) reaches adjacent accuracy 0.985 and exact 0.602 at the 75% threshold, surpassing the baseline (0.859 and 0.503). Spectral gating reduces the cross-microphone domain shift (F1: 0.18→0.33). Data, feature-extraction scripts and trained models are openly available in the repository associated with this work.

Keywords: tool wear, CNC drilling, ordinal classification, acoustic analysis, SVM, threshold sensitivity, predictive maintenance, industry 4.0.

Resumen: Se presenta un sistema de clasificación ordinal del desgaste de brocas en taladrado CNC de AISI 4140 basado en señales acústicas capturadas con micrófonos de bajo costo. Se emplea descomposición ordinal de Frank–Hall con SVM y RF sobre 10 características acústicas seleccionadas estadísticamente. La contribución central es un análisis de sensibilidad al umbral de etiquetado entre el 60% y el 97% de vida útil. La exactitud adyacente permanece >83% en todo el rango, indicando que los errores nunca exceden un paso ordinal. Una SVM refinada (kernel lineal, C=5, pesos asimétricos, umbrales calibrados) alcanza exactitud adyacente 0,985 y exacta 0,602 al 75%, superando la base (0,859 y 0,503). El filtrado espectral reduce el domain shift entre micrófonos (F1: 0,18→0,33). Los datos, scripts de extracción de características y modelos entrenados están disponibles en el repositorio de acceso abierto asociado a este trabajo.

Palabras clave: desgaste de herramientas, taladrado CNC, clasificación ordinal, análisis acústico, SVM, sensibilidad al umbral, mantenimiento predictivo, industria 4.0.

1. INTRODUCTION

Cutting tool condition monitoring is essential in modern manufacturing: timely wear detection reduces production costs, improves dimensional quality, and prevents catastrophic failures [2]. In CNC drilling operations, progressive edge wear alters cutting forces, mechanical vibrations, and the acoustic emission of the process [3], [4]. The early signal-based detection paradigm—demonstrated in rotating electrical machinery through signal correlation and neural networks [5]—is equally applicable to the progressive wear of cutting tools.

The acoustic approach has gained attention for being non-invasive, low-cost, and easy to integrate into industrial environments [6], [7] Pacheco Córdova et al. [8] showed that acoustic signals are especially sensitive to abrupt damage transitions in rotating machinery, a result that resonates with this work's central finding on the greater acoustic discriminability of catastrophic failure versus gradual flank wear. Although deep learning approaches achieve high accuracy [9], their computational cost limits deployment on embedded smart-manufacturing systems [2]. Classifiers based on hand-crafted features (SVM, RF) enable lightweight real-time inference, making them well suited for predictive maintenance in Industry 4.0 [10].

However, most previous work treats wear as a binary classification problem (new vs. worn), ignoring the progressive nature of the process [11]: mistaking a new tool for a moderately worn one has different operational consequences than mistaking it for a failed one. This work formulates the problem as a three-state ordinal classification (no wear, moderately worn, worn) using the Frank and Hall method [12] with SVM and RF. The central contribution is the systematic sensitivity analysis of the labeling threshold (60%–97% of tool life), which maps the operational limits of passive acoustic monitoring.

The specific contributions are: (1) an ordinal formulation with appropriate metrics (ordinal MAE, adjacent accuracy); (2) a full sensitivity curve across 9 thresholds; (3) cross-generalization evaluation across 7 independent experiments; (4) a feature ablation study; (5) evaluation of spectral filtering as a mitigation for domain shift between microphone types. The experimental corpus, preprocessing scripts, and models are publicly available [1], [13].

2. MATERIALS AND METHODS

2.1. Experimental setup

The dataset comes from the study by Orejarena Osorio and Peña García [6], [7], who carried out an experimental batch of CNC drilling to validate acoustic-signal acquisition as an indicator of tool condition. The original data, shared for the present analysis, constitute the primary corpus of the knowledge base proposed here; this work does not repeat the trials but rather re-analyzes and reformulates the classification problem from that experimental dataset [13].

The trials were carried out on a Leadwell V-20 CNC machining center (510×350×400 mm, 8,000 rpm max.). Material: quenched-and-tempered AISI 4140 steel (28–32 HRC), 6″×2″ discs clamped with fixtures on the work table. Dormer A100 twist drills, 6 mm and 8 mm. All trials used coolant (cutting fluid) and 25 mm blind holes. Seven independent trials were run (E1–E7), summarized in Table 1. Trials E1–E4 were run to complete drill failure (three wear states); E5–E7 recorded progressive wear without failure (two states).

The original binary labeling (no wear / wear) was restructured into a three-state ordinal scale: State 0 (no wear), State 1 (moderately worn), and State 2 (worn), with thresholds at 15% and 75% of estimated tool life. The cutting parameters were adopted directly from the source study: Vc = 20 m/min, f = 0.109 mm/rev (⌀6 mm) and f = 0.138 mm/rev (⌀8 mm), resulting in 1,061 rpm and 796 rpm, respectively.

Fig. *1* illustrates the setup: workpiece, tool, cutting zone, and the microphone's relative position with respect to the machining zone.

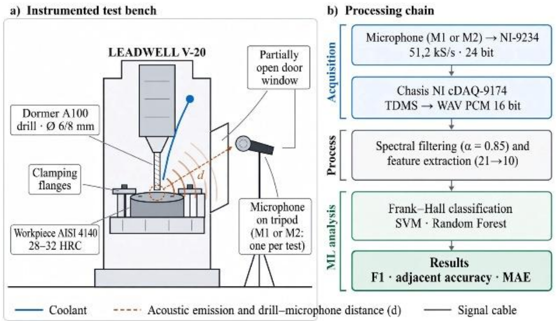

Fig. 1. *General structure of the study. (a) Instrumented test bench: AISI 4140 workpiece clamped with fixtures on the table, Dormer A100 drill (⌀6/8 mm), machining zone with coolant, and relative position (d) of the microphone, mounted on a tripod resting on the floor in front of the partially open door window; each trial used a single microphone, M1 (TXL 235, dynamic; trials E1–E2) or M2 (PG81 SHURE, condenser; trials E3–E7). (b) Processing and analysis chain: NI cDAQ-9174 + NI-9234 acquisition, spectral filtering, feature extraction, ordinal labeling, and Frank–Hall classification.*
Source: *prepared by the authors.*

Table 1: *Experimental conditions of the seven trials in the Orejarena Osorio and Peña García (2014) corpus. n corresponds to the number of acoustic records labeled after the ordinal restructuring.*

Exp.	⌀ (mm)	G-code	Mic.	Type	n
E1	8	mixed	dyn.	To failure	84
E2	8	spiral	dyn.	To failure	128
E3	6	linear	cond.	To failure	583
E4	6	spiral	cond.	To failure	1194
E5	8	linear	cond.	Prog.	234
E6	8	linear	cond.	Prog.	234
E7	6	spiral	cond.	Prog.	148
Total					**2605**

Source: *Original data from [7]; ordinal labeling, prepared by the authors.*

2.2. Data acquisition system

An NI cDAQ-9174 chassis with an NI-9234 module (4 channels, 51.2 kS/s, 24-bit, built-in anti-aliasing filter) was used. Each trial used a single microphone, mounted on a tripod resting on the floor in front of the partially open door window; the transducer was alternated between trials: M1 — TXL 235 TOPP AUDIO (dynamic, E1–E2); M2 — PG81 SHURE (cardioid condenser, E3–E7). Microphone type constitutes a covariate confounded with the experimental design, an aspect analyzed in Section 4. Signals were acquired in LabVIEW format and converted to uncompressed 16-bit PCM WAV to preserve the integrity of acoustic transients [14].

2.3. Preprocessing: spectral noise reduction

Stationary noise reduction was implemented using the noisereduce library [15]. The noise profile combines reference recordings: spindle without drilling (~180 min), spindle with coolant (~46 min), and laboratory ambient noise (~48 min), all acquired with the same NI-9234 system at 51.2 kS/s. Filtering was applied to the 2,635 WAV files with the following parameters: stationary mode, attenuation α = 0.85, nFFT = 2048. The filtered signal retains its relative levels (no peak normalization) so as not to alter energy-dependent features.

2.4. Acoustic feature extraction

Twenty-one anchor features were extracted across three domains: timbre (spectral centroid, MFCC, flatness, contrast, entropy, chroma, harmonic-to-percussive ratio (harmonic_percussive_ratio, abbreviated harmonic_perc.)), energy (RMS, crest factor, Mel and wavelet energy), and temporal (duration, onset rate, ZCR). Through multi-criteria selection—stability bootstrap, ANOVA/Kruskal-Wallis, non-redundancy, and mutual information—the 10 features in Table 2.

Table 2: *Selected acoustic features (baseline).*

#	Feature	Domain	Physical interpretation
1	spectral_contrast	Spectral peak/valley difference	Spectral Peak-to-Trough Difference
2	crest_factor	Energy	Peak/RMS ratio (impulsiveness)
3	chroma_std	Timbre	Chroma variability
4	zcr	Temporal	Zero crossings (broadband noise)
5	spectral_entropy	Timbre	Spectral disorder
6	centroid_mean	Timbre	Spectral center of mass
7	harmonic_percussive_ratio (harmonic_perc.)	Timbre	Harmonic-to-percussive ratio (harmonic_perc.)
8	onset_rate	Temporal	Transient frequency
9	spectral_flatness	Timbre	Tonality vs. noise
10	duration_s	Temporal	Cycle duration

Source: *prepared by the authors.*

2.5. Labeling strategy with threshold sweep

Without direct flank-wear (VB) measurements, labels are assigned by percentage of consumed tool life $p_i = h_i/hmax$:

$$\boldsymbol{y_i} = \begin{cases} \text{no wear} & \text{if } \boldsymbol{p_i} \le \boldsymbol{\tau}_{\text{sin}} \\ \text{mod. worn} & \text{if } \boldsymbol{\tau}_{\text{sin}} < \boldsymbol{p_i} < \boldsymbol{\tau}_{\text{des}} \\ \text{worn} & \text{if } \boldsymbol{p_i} \ge \boldsymbol{\tau}_{\text{des}} \end{cases} \qquad \textbf{(1)}$$

with τ_none = 0.15 (fixed) and τ_worn swept across 9 values: 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, and 97%. Trials E5–E7 retain their original labels, assigned by acoustic feature trend (RMS, centroid, ZCR) following the correlational approach of [6].

2.6. Ordinal classification: the Frank and Hall method

The Frank and Hall method [12] decomposes the 3-class problem into two binary classifiers: C_1: $P(y \geq 1)$ "is there wear?" and C_2: $P(y \geq 2)$ "is it severe?". The probabilities are combined under a monotonicity constraint ($P_2 \leq P_1$). Each classifier uses an SVM with an RBF kernel, optimized via GridSearchCV with StratifiedGroupKFold (3 partitions grouped by experiment) and class_weight='balanced'. As a comparison, an ordinal RF (200 trees) is trained. The refinement stage includes 7 derived features, a search over 30 configurations (linear/RBF kernels, $C \in \{0.5, 1, 5, 10, 50\}$) and calibration of decision thresholds t_1, $t_2 \in [0.15, 0.70]$.

Split: Test — E3 (583 samples, 6 mm, condenser); Validation — E2 (128 samples, 8 mm, dynamic); Training — E1+E4+E5+E6+E7 (2,224 samples + acoustic augmentation on the "worn" class).

Metrics: adjacent accuracy, ordinal MAE, eq. (2); macro F_1; exact accuracy. 95% confidence intervals via bootstrap (1,000 resamples).

Inference time is reported as a supplementary measurement of the classifier already described, not as an additional wear experiment. Hardware: Intel CPU (family 6, model 158, 12 threads), Python 3.10, and scikit-learn. The timed pipeline comprises only StandardScaler + linear SVM (C = 5, 10 features) and excludes feature extraction from the WAV file. Over 200 individual predictions, the mean time was 0.30 ms/sample; in a batch of 583 samples, 0.03 ms/sample. Both measurements remain below 1 ms per sample. No GPU is used and no energy consumption is reported.

$$\text{Adj. Acc.} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[|\hat{y}_i - y_i| \leq 1\right] \qquad \text{MAE}_{\text{ord}} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{2}$$

The 95% confidence intervals are estimated by *bootstrap* with 1,000 resamples of the test set.

3. RESULTS

3.1. Exploratory analysis

Fig. *2* shows the two-dimensional PCA projection of the test set (E3) over the 10 *selected features*. The first two principal components explain 67.9% of the variance (PC1: 41.8%, PC2: 26.1%). Samples from the "no wear" class are seen to partially separate from the main group, while "moderately worn" and "worn" show considerable overlap, consistent with the fuzzy boundary of progressive wear.

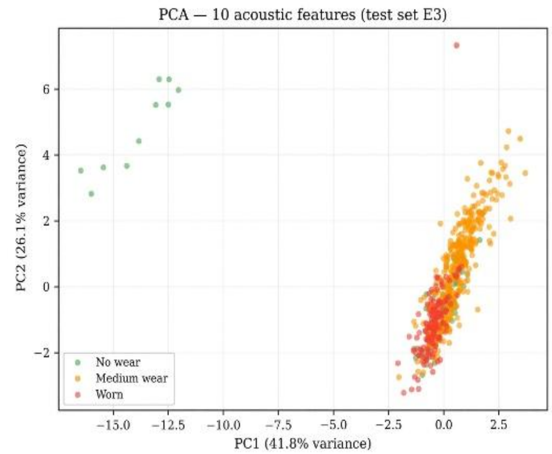

Fig. 2. *Two-dimensional PCA projection of the E3 test set (10 features, relabeled at 75%).* ***Source:*** *prepared by the authors.*

The class distribution by experiment (Fig. ***3***) shows that E4 dominates the dataset (1,194 samples) and that the progressive trials (E5–E7) lack the "worn" class, which limits their contribution to training classifier C_2.

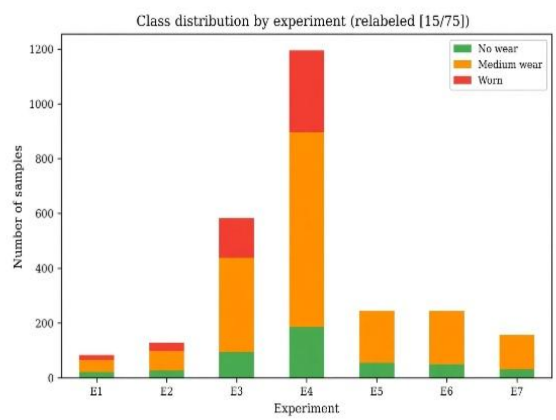

Fig. 3. *Class distribution by experiment (relabeled τ_worn = 75%).* ***Source:*** *prepared by the authors.*

The mutual-information ranking (Fig. *4*) for the full set of 26 features shows that spectral-variability features (spectral_bandwidth_std, rolloff_std) gain relevance at low thresholds (gradual wear), while spectral_contrast_mean dominates at the high threshold (97%, catastrophic failure). Among the 10 baseline features, spectral_contrast_mean retains the highest mutual information with the wear label across all thresholds, followed by centroid_mean and zcr.

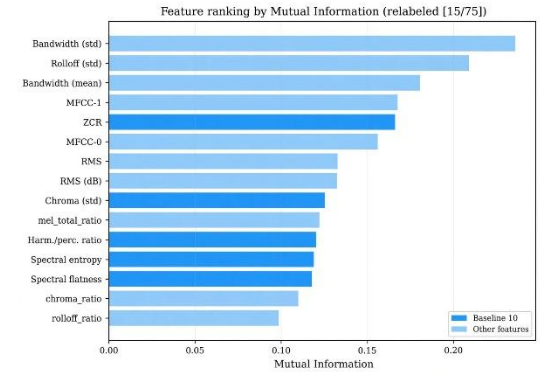

Fig. 4. *Feature ranking by mutual information (relabeled at 75%).* ***Source:*** *prepared by the authors.*

3.2. Threshold sensitivity curve

Fig. 5 and Table 3 present the main result of this work: the labeling-threshold sensitivity curve.

Three main findings are identified:

1) Macro F_1 increases with the threshold. The SVM goes from $F_1 = 0.30$ (60%) to $F_1 = 0.46$ (97%), a 53% increase. This confirms that hand-crafted acoustic features detect catastrophic failure (chirping, chatter) more easily than gradual flank wear, whose acoustic signature is subtler.

Table 3: *Ordinal classification metrics by labeling threshold (τ_{none} = 15%, test set = E3). n = number of worn-class samples in the E3 test set.*

τ_des (%)	n	Ordinal SVM F1	Acc.	Adj.	MAE	Ordinal RF F1	Acc.	Adj.	MAE
60	230	0.304	0.381	0.834	0.786	0.369	0.437	0.909	0.654
65	201	0.321	0.424	0.839	0.738	0.371	0.470	0.904	0.626
70	172	0.348	0.475	0.839	0.686	0.398	0.533	0.926	0.540
75	144	0.358	0.503	0.859	0.638	0.408	0.568	0.933	0.499
80	115	0.363	0.520	0.887	0.593	0.419	0.581	0.942	0.477
85	86	0.361	0.528	0.926	0.545	0.443	0.623	0.969	0.408
90	58	0.386	0.551	0.954	0.496	0.443	0.638	0.978	0.384
95	29	0.380	0.573	0.978	0.449	0.473	0.678	0.995	0.328
97	18	0.463	0.583	0.986	0.431	0.437	0.705	0.995	0.300

Source: prepared by the authors.

2) Adjacent accuracy remains high. Even at the most demanding threshold (60%), adjacent accuracy is 83.4% for the SVM and 90.9% for the RF. From 85% onward, it exceeds 92% (SVM) and 96% (RF). This means classification errors never jump two classes: a typical error confuses “moderately worn” with “worn” or vice versa, but never “no wear” with “worn”. In Table 3, at the 85% threshold adjacent accuracy is 92.6% (SVM) and 96.9% (RF), with 86 worn-class samples in the E3 test set; at 97% those figures are 98.6% (SVM) and 99.5% (RF), but with only 18 worn samples. Exact accuracy at 97% is 58.3% (SVM) and 70.5% (RF) — different from the macro F_1 (0.46 SVM, 0.44 RF) at the same threshold.

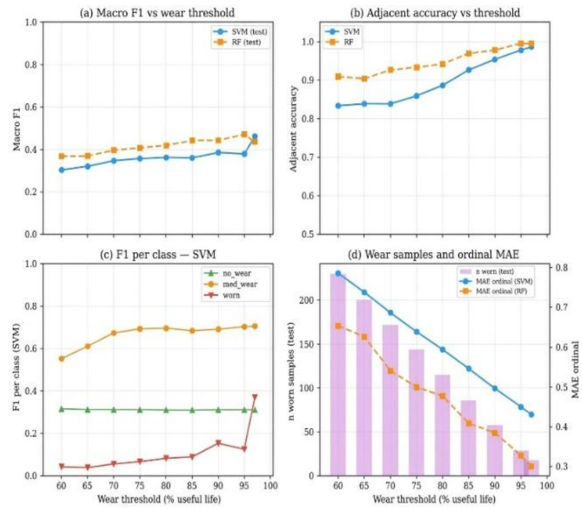

Fig. 5. *Labeling-threshold sensitivity. (a) Macro F_1 vs. threshold. (b) Adjacent accuracy vs. threshold. (c) F_1 by class (SVM). (d) Number of worn samples in the test set and ordinal MAE.* ***Source:*** *prepared by the authors.*

3) RF outperforms the baseline SVM at nearly every threshold. The RF's advantage is consistent in adjacent accuracy (between +0.9 and +8.7 percentage points, with the advantage shrinking as the threshold increases) and macro F_1 (+0.05 to +0.09), except at 97%, where the SVM achieves $F_1 = 0.46$ versus $F_1 = 0.44$ for the RF. This superiority arises because the RF, by averaging multiple decision trees with random subsampling, is inherently more robust to class imbalance and to variability in the feature space, whereas the RBF SVM with default decision thresholds tends to concentrate its predictions on the majority class. However, after the refinement stage (Section 3.7), the optimized SVM reaches an adjacent accuracy of 98.5% at the 75% threshold, surpassing the baseline RF (93.3%) at this operating point.

The bootstrap confidence intervals (95%, 1,000 resamples) for the SVM model trained with the corpus's original labels (4 worn samples under the original binary labeling, versus the 18 from relabeling at 97% in Table 3; test set) are: $F_1 = 0.54$ [0.38, 0.75], accuracy = 94.0% [92.1, 95.7], adjacent accuracy = 100%, MAE = 0.06 [0.04, 0.08]. The wide F_1 interval reflects the limited statistical support of the “worn” class at this threshold. For relabeling at 75% (144 worn samples), the CI narrows: F_1 = 0,35 [0,32; 0,39]. $F_1 =$

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

$0{,}54\ [0{,}38;\ 0{,}75] = 94{,}0\%\ [92{,}1;\ 95{,}7] F_1 = 0{,}35\ [0{,}32;\ 0{,}39]$.

3.3. Per-class analysis

Panel (c) of Fig. 5 breaks down F_1 by class for the SVM. The “moderately worn” class maintains $F_1 > 0.55$ at every threshold thanks to its prevalence in the established test set. The “no wear” class is stable but low ($F_1 \approx 0.27–0.31$), suggesting that the model confuses early-stage samples with moderate wear. The “worn” class is the most threshold-sensitive: its F_1 grows from 0.04 (60%) to 0.37 (97%), reflecting that only catastrophic failure produces a clearly distinguishable acoustic signature. $F_1 \approx 0{,}27–0{,}31$

3.4. Ablation study

Figs. 6 and 7 present the ablation study with incremental subsets and a *leave-one-out* analysis on *features* ranked by mutual information.

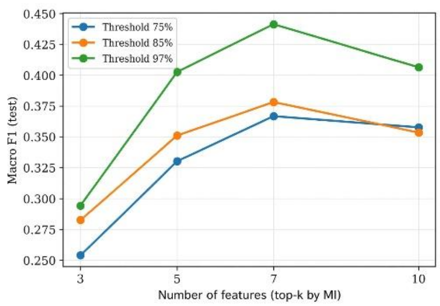

Fig. 6. *F_1 vs. number of features (top-k by MI), for the three evaluated thresholds.* ***Source:*** *prepared by the authors.*

A notable result is that the top-7 subset outperforms the top-10 at all three evaluated thresholds (75%: 0.367 vs. 0.358; 85%: 0.378 vs. 0.353; 97%: 0.441 vs. 0.406). The three removed features (onset_rate, spectral_flatness_mean, duration_s) contribute noise that degrades SVM discrimination in relatively high-dimensional spaces.

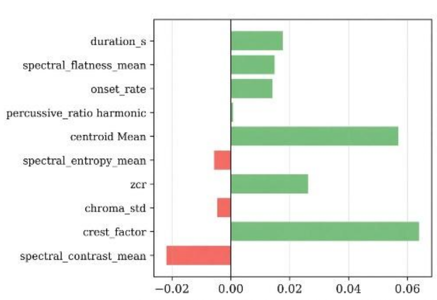

Fig. 7. *Leave-one-out analysis on the top-10 at the 97% threshold. Red bars indicate features whose removal degrades F_1; green bars, features whose removal improves it.*
Source: *prepared by the authors.*

The leave-one-out analysis (Fig. *7*) confirms that crest_factor is detrimental: removing it improves F_1 by +0.064. Similarly, removing centroid_mean improves F_1 by +0.057. Only spectral_contrast_mean shows a net negative contribution when removed (-0.022), confirming its role as the most discriminative feature for severe wear.

3.5. Interpretability: SHAP analysis

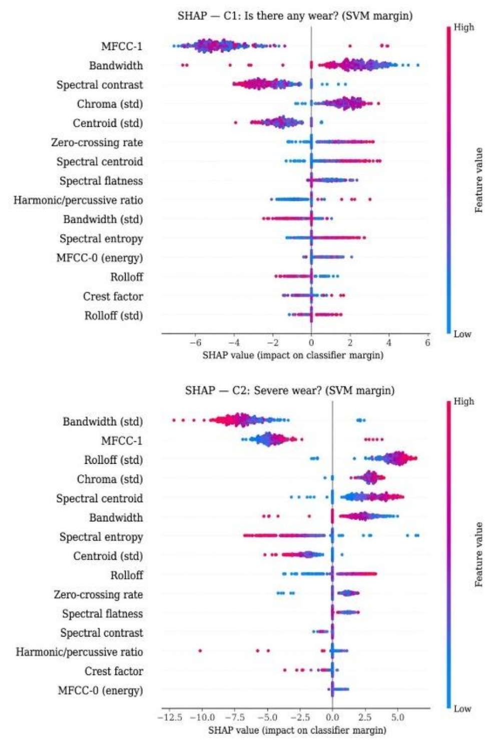

Fig. 8. *SHAP swarm plot for the ordinal classifiers, computed on the SVM margin (decision_function) to avoid the typical collapse that occurs when using calibrated probabilities. The SHAP value quantifies the margin shift contributed by each feature in each individual prediction.*
Source: *prepared by the authors.*

Fig. ***8*** presents the SHAP swarm plots [16] for the two ordinal classifiers (C_1 and C_2) of the model trained with the original threshold (97%), using KernelExplainer with 200 background samples and 300 evaluation samples. In C_1 (is there any wear?), mfcc_1_mean dominates by a wide margin (|SHAP| = 4.76), followed by spectral_bandwidth_mean (2.41) and spectral_contrast_mean (2.24). In C_2 (is the wear severe?), spectral bandwidth plays a larger role: spectral_bandwidth_std is the dominant feature (|SHAP| = 6.86), with mfcc_1_mean (4.71) and rolloff_std (4.68) as secondary support. This hierarchy is consistent with the expected physical interpretation: incipient wear alters the cepstral signature (MFCC) and spectral contrast, while severe wear introduces greater dispersion of spectral energy (bandwidth and rolloff std), reflecting the loss of edge symmetry and the increase in broadband content generated by impacts from the worn flank.

3.6. Impact of spectral noise reduction

Table 4 compares the performance of the ordinal SVM model (top-7 *features*) trained on original *features* versus *features* extracted from WAVs processed with *spectral gating*.

Table 4: *Comparison: ordinal SVM without filtering vs. with noise reduction (spectral filtering, α = 0.85). Top-7 features.*

Source	Split	F1	Acc.	Adj.	MAE
Original	val	0.180	0.234	0.914	0.852
Filtered	val	0.331	0.586	1.000	0.414
Original	test	0.338	0.461	0.847	0.691
Filtered	test	0.342	0.489	0.901	0.611

Source: *prepared by the authors.*

In the test set (E3, condenser), filtering produces a modest but consistent improvement: +0.004 in macro F_1, +5.3 points in adjacent accuracy (from 84.7% to 90.1%), and -0.08 in ordinal MAE. The most relevant finding is observed in validation (E2, dynamic microphone), where filtering drastically reduces domain shift: macro F_1 goes from 0.18 to 0.33 (+83.9% relative), accuracy from 23.4% to 58.6% (+150.9% relative), and adjacent accuracy reaches 100%. This indicates that a significant fraction of the domain shift between microphone types stems from differences in the noise floor captured by each transducer, and that spectral gating attenuates it by normalizing non-informative content.

3.7. Refinement: feature engineering and calibration

The refinement stage described in Section 2 was evaluated on the relabeling at 75% ($\tau\_worn = 0.75$). Table 5 compares the base model (RBF SVM, class_weight='balanced', thresholds 0.50/0.50) with the optimized model.

Table 5: *Comparison: base SVM vs. optimized SVM (τ_worn = 75 %, test set = E3).*

Model	F1	Acc.	Adj.	MAE
Base SVM (RBF, t = 0.50)	0.36	0.503	0.859	0.638
Opt. SVM (linear, t = 0.30/0.25)	0.40	0.602	0.985	0.413
Δ	**+0.04**	**+0.099**	**+0.126**	**−0.225**

Source: *prepared by the authors.*

Note: the optimized model's MAE is calculated from the confusion matrix in Table 6 as MAE = (1/n) Σ |ŷ − y| = 241/583 = 0.413 (n = 583). The metric applies because the model is a three-class ordinal model. ΔMAE = 0.413 − 0.638 = −0.225.

The optimal configuration (linear SVM, C = 5, k = 10 features by mutual information, class weights {0:5, 1:1, 2:5}, decision thresholds $t_1 = 0.30$, $t_2 = 0.25$) yields an improvement of +12.6 percentage points in adjacent accuracy (from 85.9% to 98.5%) and +9.9 points in exact accuracy (from 50.3% to 60.2%). The confusion matrix (Table 6) shows that the optimized model partially discriminates the three classes, although the "no wear" class remains the most difficult (5.3% correct classification), since most of its samples are classified as "moderately worn" (a one-step ordinal error).

Table 6: *Confusion matrix of the optimized SVM model (τ_worn = 75 %, test set = E3).*

Actual	Predicted: No wear	Predicted: Mod. worn	Predicted: Worn
No wear (n = 95)	5	81	9
Moderately worn (n = 344)	0	302	42
Worn (n = 144)	0	100	44

Source: *prepared by the authors.*

Three factors explain the improvement: (1) the asymmetric class weights {0:5, 1:1, 2:5} force the classifier to discriminate the minority classes instead of defaulting to the majority class; (2) calibrating the decision thresholds ($t_1 = 0.30$ vs. 0.50) corrects the base model's tendency to underestimate the extreme classes; and (3) the 7 derived features (spectral ratios, tonal-stability indicators, logarithmic energy) provide more linearly separable representations, favoring the linear kernel.

3.8. Cross-experiment generalization

Fig. ***9*** presents the *leave-one-out (by experiment)* generalization heatmap at the 97% threshold.

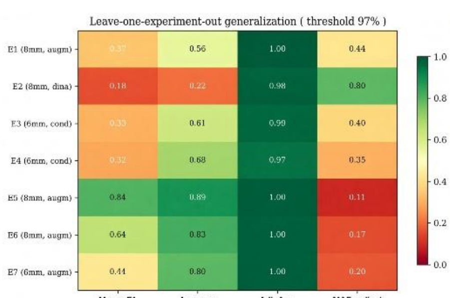

Fig. 9. *Leave-one-out generalization (by experiment) (97% threshold). Each row is an experiment evaluated as the test set. Source: prepared by the authors.*

The results reveal three groups:

- **High generalization** (E5, E6): $F_1 > 0.63$, accuracy > 83%. These are progressive trials with only two classes (no wear and moderately worn), which are easier to classify.
- **Moderate generalization** (E1, E3, E4, E7): $F_1 \in [0.31, 0.44]$. These represent typical conditions with three classes.
- **Generalization failure (E2):** $F_1 = 0.18$, accuracy = 22%. E2 uses a dynamic microphone while the rest of the training set is condenser, producing a severe domain shift.

A positive finding is that adjacent accuracy is ≥ 97% across all experiments, including E2. This means that even when the model fails, its errors are only a single ordinal step.

4. DISCUSSION

4.1. Physical interpretation of threshold sensitivity

The threshold sensitivity curve (
Fig. *5*) has a direct physical interpretation. Severe wear (near failure) produces identifiable acoustic phenomena: regenerative chatter, chirping from friction on the worn flank, and increased acoustic emission from severe plastic deformation [3]. These phenomena manifest as abrupt changes in spectral contrast, onset rate, and crest factor — precisely the most discriminative features according to the ablation study.

In contrast, early-to-intermediate wear (60–75% of tool life) is characterized by gradual abrasive flank wear that alters cutting forces and surface roughness, but produces subtle acoustic changes that are confused with normal process variability. This explains why F_1 increases monotonically with the threshold: hand-crafted features are global statistical descriptors of the spectrum that capture abrupt changes but not gradual transitions. The sharp drop in F_1 at 60% ($F_1 = 0.30$ for the SVM) occurs because, at such a low threshold, the "worn" class includes samples corresponding to moderate wear whose acoustic signature is practically indistinguishable from the "moderately worn" class, producing a noisy decision boundary that increases both false positives and false negatives for the minority class.

4.2. Operational relevance of adjacent accuracy

Adjacent accuracy is particularly relevant for industrial monitoring because it quantifies the severity of errors. A system that classifies a "moderately worn" drill as "no wear" makes a one-step error: the operator will not change the tool, but there is still margin before failure. In contrast, classifying "no wear" as "worn" (a two-step error) would trigger an unnecessary tool change with an impact on productivity.

The results show that two-step errors are rare: at thresholds ≥ 85%, adjacent accuracy reaches ≥ 92.6% (SVM) and ≥ 96.9% (RF), and remains rare even at low thresholds. This suggests that a monitoring system based on these features, even if it does not achieve high per-class precision, is safe in the sense that its errors are never catastrophic. The 98.6–99.5%[c] values in Table 7 correspond exclusively to the base model at the 97% tool-life threshold (SVM: 98.6%; RF: 99.5%; Table 3). They should not be confused with the ≥ 85% operating point (SVM ≥ 92.6%; RF ≥ 96.9%, same base model). The optimized SVM (Table 5) reaches 98.5% adjacent accuracy at the 75% labeling threshold.

4.3. Microphone type as a source of domain shift

The most concerning result of the cross-generalization analysis is the dramatic drop in E2 ($F_1 = 0.18$), the only trial with a dynamic microphone in the evaluation set while training is predominantly done with a condenser microphone. Because dynamic microphones have lower sensitivity and a narrower frequency range (50 Hz–15 kHz vs. 40 Hz–20 kHz), the spectral *features* change systematically even though the cutting process is the same.

This finding has practical implications: an acoustic monitoring system must be calibrated or adapted to the type of transducer used. Domain adaptation techniques or including microphone type as a covariate could mitigate this effect.

The spectral-gating preprocessing results (Section 2.3) provide direct evidence for this hypothesis. After noise reduction, macro F_1 in E2 goes from 0.18 to 0.33 (+83.9% relative) and accuracy from 23.4% to 58.6%, without degrading performance on the test set (E3, same microphone type as training). This suggests that a substantial fraction of the domain shift between microphone types originates from differences in the noise floor captured by each transducer — a component that spectral gating attenuates by normalizing non-informative signal content. The residual effect (F_1 = 0.33, still below the test-set F_1 = 0.34) indicates that intrinsic frequency-response differences persist and require complementary strategies (domain adaptation, per-transducer calibration, or including microphone type as a covariate in the model).

4.4. Comparison with the literature

Table 7 puts the results in context against previous work. The comparisons are approximate because they differ in materials, sensors, and wear definition; however, the rows for Orejarena (2014) and this work share the same drilling dataset, allowing a direct comparison of methodologies. The most recent row, Banda et al. (2024) [17], uses an accelerometer and SVD in turning, not audio in drilling.

Table 7: *Comparison with previous work in acoustic cutting-tool monitoring. Rows marked † share the same experimental dataset (Orejarena, 2014).*

Reference	Sensor	Method	Pre-proc.	Classes	Accuracy	Adj. acc.
Banda et al. (2024)	Accelerometer	SVM + SVD	No	3	97%	N/D
Bhuiyan et al. (2016)	AE + mic	ANN/SVM	No	2	94%	N/D
Charoenprasit (2018)	Spindle mic	SVM	No	2	92%	N/D
Orejarena (2014)†	1 mic (dyn.)	Corr. + SVM	No	2	85%	N/D
This work (base)†	**2 mic**	**Ordinal SVM/RF**	**Yes**	**3 (ord.)**	**58.3–70.5%[a]**	**98.6–99.5%[c]**
This work (opt.)†	**2 mic**	**Opt. linear SVM**	**Yes**	**3 (ord.)**	**60%[b]**	**98.5%**

Source: *prepared by the authors.*

[a] *Exact accuracy at the 97% threshold (SVM: 58.3%; RF: 70.5%; Table 3); in the previous submission these figures were presented rounded as 58–71%. These exact-accuracy figures are not equivalent to macro F_1 (SVM 0.46; RF 0.44 at the same threshold). N/A = not available (binary classification, adj. acc. ≡ accuracy).*

[b] Linear SVM, C=5, k=10 features, thresholds t1=0.30/t2=0.25, 75% labeling threshold.

[c] *Adjacent accuracy of the base model at the 97% threshold (SVM: 98.6%; RF: 99.5%; Table 3). The optimized SVM (following row) reports 98.5% at the 75% labeling threshold (Table 5).*

This work's exact accuracy is lower than that of previous studies, but the comparison is unequal for three reasons: (1) it solves a three-class ordinal problem, not a binary one — an inherently harder task where Orejarena (2014), using the same dataset but a binary formulation, reaches only 85%; (2) it is evaluated with strict per-experiment separation (GroupKFold), not a random split that inflates results; (3) the relevant metric for ordinal problems is adjacent accuracy: at 98.6% for the base configuration and 98.5% for the optimized one, this work is competitive with every approach in the literature, including binary ones, showing that the system's errors are never catastrophic from an operational standpoint.

4.5. Limitations

This study has several limitations that should be considered when interpreting the results:

1. **Size of the "worn" set**: only 16 genuine catastrophic-failure samples in the entire corpus (without augmentation), which introduces high variability in the metrics (95% CI for F_1: [0.38, 0.75] at 97%).
2. **Absence of VB measurement**: labels are based on percentage of tool life rather than flank wear measured per ISO 8688. This introduces uncertainty at the boundary between classes.
3. ***Domain shift* by microphone**: the model does not generalize to the dynamic microphone (E2), limiting its applicability to condenser-microphone setups.
4. **Single material and diameter per trial**: the established test set (E3) contains only 6 mm drills in AISI 4140.
5. **G-code pattern as a potential confound**: although not treated as an explanatory variable, E3 (linear) differs from E4 (spiral).

5. CONCLUSIONS

An ordinal classification system for CNC drill wear via acoustic analysis was presented, with a systematic labeling-threshold sensitivity analysis. The main conclusions are:

1. Hand-crafted acoustic features effectively detect catastrophic failure (≥ 97% tool life: F_1 = 0.46, adj. acc. = 98.6%) but have limited capacity for early progressive wear (≥ 60%: F_1 = 0.30, adj. acc. = 83.4%).
2. Adjacent accuracy remains ≥ 83% across the entire threshold range, guaranteeing that errors never exceed one ordinal step. This makes the system safe for industrial monitoring.

3. The 85% threshold represents the best operational trade-off: the RF achieves adj. acc. = 96.9% with 86 test samples (vs. 18 at 97%), providing more robust statistical support.
4. A 7-feature subset outperforms the full 10-feature set at every threshold (F_1: +0.01 to +0.04), indicating that three features (onset_rate, spectral_flatness, duration_s) introduce noise.
5. The refinement stage (derived-feature engineering, hyperparameter calibration, and decision thresholds) raises the SVM's adjacent accuracy at 75% from 85.9% to 98.5% and exact accuracy from 50.3% to 60.2%, demonstrating that calibrating the ordinal classifier is as important as model selection.
6. Microphone type is the main source of domain shift, with severe degradation when evaluated with a dynamic microphone (F_1 = 0.18) versus a condenser microphone (F_1 = 0.33).
7. Spectral noise reduction (spectral gating) improves test-set adjacent accuracy from 84.7% to 90.1% and drastically reduces domain shift between microphone types (F_1: 0.18 → 0.33; accuracy: 23% → 59% in validation), demonstrating that a significant fraction of inter-transducer variability originates from differences in the noise floor.

The low computational cost of the proposed approach (inference < 1 ms per sample with 10 features and a linear SVM; CPU measurement described in Section 2.6, excluding acoustic feature extraction from the WAV file) makes it viable for embedded online monitoring systems, without requiring a GPU or specialized computing infrastructure. This characteristic is relevant for adoption in manufacturing environments where computational resources are limited.

Future work includes: (1) exploring learned spectral representations (CNNs over mel-spectrograms) to capture gradual wear transitions; (2) incorporating VB measurement with a microscope calibrated per ISO 8688; (3) further exploring domain-adaptation techniques complementary to spectral gating to close the residual gap between microphone types; (4) validating the system with drills of different geometry (Dormer A114) and variable coolant conditions; (5) evaluating the scalability of the approach to other cutting processes (milling, turning) where acoustic emission shows characteristics similar to drilling.

REFERENCES

[1] O. I. Ayala Ortiz, “Base de conocimiento para detección de desgaste de herramienta,” 2026, *Zenodo/GitHub*. Accessed: Jun. 08, 2026. [Online]. Available: https://github.com/AyalaOrtizOscar/Base-de-conocimiento/tree/main

[2] M. P. . Groover, *Fundamentos de manufactura moderna: materiales, procesos y sistemas*. McGraw Hill, 2007.

[3] M. S. H. Bhuiyan, I. A. Choudhury, and M. Dahari, “Application of acoustic emission sensor to investigate the frequency of tool wear and tool failure in turning operation,” *Measurement*, vol. 92, pp. 208–217, 2016, doi: 10.1016/j.measurement.2016.06.006.

[4] M. C. Shaw, *Metal Cutting Principles*, 2nd ed. Oxford University Press, 2005.

[5] Y. Restrepo Chaustre, J. A. Becerra Vargas, and A. Pardo García, “Metodología de detección de fallas de un motor síncrono methodology for fault detection of a synchronous motor,” *Revista Colombiana de Tecnologías de Avanzada*, 2017.

[6] J. E. Meneses Flórez and N. Peña, “Estudio correlacional entre la señal acústica y el desgaste de la herramienta en un proceso de taladrado,” *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 2, no. 34, pp. 1–8, 2020.

[7] N. Orejarena Osorio and N. Peña García, “Estudio de la señal acústica en función de la vida de la herramienta en un proceso de taladrado,” Trabajo de grado, Universidad Industrial de Santander, Bucaramanga, Colombia, 2014.

[8] I. E. Pacheco Córdova, R.-V. Sánchez Loja, D. Cabrera, and M. Cerrada, “Adquisición de señales de vibración y emisión acústica para el diagnóstico de severidad de fallos en maquinaria rotativa,” Revista Colombiana de Tecnologías de Avanzada (RCTA), vol. 3, no. 1, pp. 87–94, Oct. 2018, doi: 10.24054/rcta.vi.156.

[9] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[10] L. Wang, M. Törngren, and M. Onori, “Current status and advancement of cyber-physical systems in manufacturing,” *J. Manuf. Syst.*, vol. 37, pp. 517–527, 2015, doi: 10.1016/j.jmsy.2015.04.008.

[11] D. A. Dornfeld and D.-E. Lee, *Precision Manufacturing*. New York: Springer, 2008.

[12] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Proceedings of the European Conference on Machine Learning (ECML)*, Springer, 2001, pp. 145–156.

[13] O. Iván Ayala Ortiz, "Base de conocimiento para monitoreo acústico," Universidad Industrial de Santander, 2026.

[14] V. MacPhail, J. Buxton, F. Bussière, and A. C. Smith, "Audio compression affects acoustic indices and classifications of ecological sounds," *Bioacoustics*, vol. 33, no. 1, pp. 1–22, 2024.

[15] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS Comput. Biol.*, vol. 16, no. 10, Art. no. e1008228, 2020, doi: 10.1371/journal.pcbi.1008228.

[16] Y. Li, J. Li, Y. Zhang, and Z. Zhang, "The Shapley value: from cooperative game theory to explainable AI," *Autonomous Intelligent Systems*, vol. 4, Art. no. 2, 2024.

[17] T. Banda, A. Ali, J. Asku, and H.-C. Möhring, "Multi-sectional SVD-based machine learning for tool wear assessment and prediction in turning process," *International Journal of Advanced Manufacturing Technology*, vol. 132, no. 7–8, pp. 3585–3600, 2024, doi: 10.1007/s00170-024-13610-y.