Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Grouped cross-validation and inter-sensor agreement analysis of expert labels in an open milling anomaly dataset

**Adam Park**

apark1@imsa.edu

Illinois Mathematics and Science Academy https://orcid.org/0009-0005-7400-4383

**Luke Yin**

Princeton University

**Cooper Lee**

Illinois Mathematics and Science Academy

**Andrew Wong**

University of Michigan

**Yoo Choi**

Cornell Tech

**Research Article**

**Keywords:** Milling monitoring, Anomaly detection, Benchmark labels, Microphone clipping, Grouped cross-validation

**Posted Date:** October 8th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11145072/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

# Grouped cross-validation and inter-sensor agreement analysis of expert labels in an open milling anomaly dataset

Adam Park[1*], Luke Yin[2], Cooper Lee[1], Andrew Wong[3], Yoo Choi[4]

[1*]Illinois Mathematics and Science Academy, Aurora, Illinois, USA.
[2]Princeton University, Princeton, New Jersey, USA.
[3]University of Michigan, Ann Arbor, Michigan, USA.
[4]Cornell Tech, New York, New York, USA.

*Corresponding author(s). E-mail(s): apark1@imsa.edu;

**Abstract**

Open datasets let detectors of machining anomalies be compared on the same labels. We analysed the expert path labels of the multi-sensor and MTConnect (MSM) milling dataset on its laboratory machine with accelerometers: 1,498 labelled cutting paths from 34 runs, 325 of them abnormal. Of three predictions specified before any of their statistics was computed, one held and two failed. For logistic models on the root-mean-square (RMS) amplitudes of the accelerometer and sound channels, the area under the receiver operating characteristic curve (AUC) fell by 0.002 (95 percent interval −0.001 to 0.007) from unseen runs to unseen tool families, with their materials and spindle speeds; for models on the spindle signals streamed through MTConnect it fell by 0.033 (0.015 to 0.060). Both falls stayed below the specified threshold of 0.05; a post hoc jackknife over refitted models leaves the size of the MTConnect fall unresolved. Among abnormal paths, the spindle accelerometer and a stethoscope sensor on the base column showed a dominant spectral line off the spindle harmonics on the same paths (Cohen's kappa 0.71, against 0.004 under independence within spindle-speed classes; threshold 0.41), and no robustness variant contradicted this result. Seen before the predictions were specified, every abnormal path and 38 percent of normal paths had more than 1 percent of one microphone's samples at full scale; in exploratory analyses that share alone ranked the labels with an AUC of 0.985 on unseen runs. Detector comparisons on this dataset should split by run and report a clipping baseline.

**Keywords:** Milling monitoring, Anomaly detection, Benchmark labels, Microphone clipping, Grouped cross-validation

## 1 Introduction

Chatter harms surface quality and tool life in milling and limits the material removal rate, and its detection from sensor signals has been studied for decades [1]. A recent review covers the sensing set-ups, signal processing and chatter indicators reported for milling and discusses data-driven classification of the machining condition [1]. Microphones were found early to give a chatter signal competitive with, and in many instances better than, those of dynamometers, displacement probes and accelerometers [2]. Data-driven detectors need labelled cuts, and open datasets

let different detectors be compared on the same labels.

What such a comparison means depends on what the labels encode and on how the evaluation separates training data from test data. When observations are grouped, as cutting paths are within runs, random cross-validation underestimates the predictive error [3]. Leakage between training and test data has produced over-optimistic conclusions in many fields that adopted machine learning [4], and a guide to machine learning in computer numerical control (CNC) machining advises attention to it [5]. A detector can also score well by learning a shortcut, a decision rule that works on the benchmark and fails under other conditions [6]. In machinery diagnostics, the Case Western Reserve University bearing data became a standard reference for testing diagnostic algorithms, and a benchmark study examined the whole data set and recommended how it should be used [7]. A later article on the evaluation of machine learning for condition-based maintenance describes incorrect separation of training and test examples as a fairly common problem in studies of bearing fault classification, one that leads to optimistic conclusions [8].

The multi-sensor and MTConnect (MSM) dataset is an open benchmark for anomaly detection in milling [9, 10]. It holds synchronised sound and MTConnect controller data from laboratory and production milling machines, with acceleration on the laboratory machine imi_vm20i, and every cutting path carries an expert label: normal (label 0), abnormal (label 1) or tool defect (label 2). The descriptor defines the abnormal label as cutting that should be stopped to protect the tool, workpiece or machine, lists excessive tool load, setup errors, fixture instability, tool wear, chatter and forced vibration among its causes, and states that the label does not denote a single vibration mechanism [9]. Cutting intervals were refined by examining the synchronised signals, in particular the sound recordings [9]. In earlier work, two of the dataset authors labelled chatter events from machining sound with expert knowledge and trained a convolutional network on those labels [11].

Two published studies have used the dataset. Parmar and Raman classified whole runs from MTConnect summaries, with labels taken from the run remarks rather than the path labels, and cross-validated over runs [12]. Karkadakattil built interval-wise MTConnect features over the labelled cutting paths, merged the abnormal and tool-defect labels, and split the samples by stratified sampling without stating whether runs were kept apart [13]. A public code repository for Bayesian monitoring of milling, with an unpublished manuscript (https://github.com/Guldek1987/bayesian-milling-monitoring-hbaf, accessed 22 September 2026), uses experiment-disjoint splits on the MTConnect data. A public notebook on the dataset's Kaggle page (https://www.kaggle.com/code/stephenhoanganh/experiment-anomaly-detection, accessed 23 September 2026) trains tree ensembles on short windows of sound and MTConnect data from imi_vm20i with folds that keep the windows of each run together, and its attributions rank Sound0 first among the sensors. None of these works measures what the path labels encode.

This paper measures it on imi_vm20i. It asks how much of the detectors' ranking performance survives when whole tool families are held out, and whether a spindle accelerometer and a sensor on the base column mark the same abnormal paths with a spectral line that the spindle harmonics do not explain. The predictions (H1 to H3) and their machine-checkable decision rules were specified and frozen before any of their statistics was computed. Loudness, clipping and tool identity had been examined before the predictions were specified and are reported as seen, and exploratory analyses describe how the evaluation scheme changes the scores. The results lead to recommendations for users of the dataset.

## 2 Dataset and analysis populations

### 2.1 Machine, sensors and labels

All confirmatory analyses use imi_vm20i, a 3-axis CNC mill (Hurco VM20i) in the laboratory part of the MSM dataset [9]. It is the only machine with high-rate accelerometers: Acc0 on the spindle and Acc1 on the base column, sampled at 51.2 kHz. The sound channels were recorded at 48,000 Hz with 16-bit resolution: Sound0, a microphone inside the machine enclosure, and Sound1 and Sound2, internal sound sensors made of a stethoscope and the same microphone, placed near the

spindle and on the base column [9]. MTConnect carried the controller data together with a spindle-motor vibration sensor and a power meter, whose observations the adapter updated about every 1 s [9]. The laboratory cuts were made wet, and some cutting parameters were chosen to be aggressive so that anomalies such as chatter or tool failure would occur [9].

A run is one folder of the dataset and generally corresponds to one workpiece machined without a tool change [9]; on imi_vm20i each run used a single tool family and a single workpiece material. The file of each run lists the start, end and label of every cutting path. Tools come in families: a tool and its defective twin, marked with an asterisk (A and A*), form one family, and the descriptor labels a path cut with a defective tool abnormal when the cutting itself was anomalous and tool defect otherwise [9].

## 2.2 Populations

imi_vm20i holds 37 runs with 1,585 labelled paths: 1,173 label 0, 325 label 1 and 87 label 2. The analysis population is every label 0 and label 1 path, 1,498 paths in 34 runs; label 2 marks a defective tool, a different question, and stays out. Label 1 occurs in 19 runs. Table 1 gives the counts by tool family and workpiece material, and Fig. 1 the label 1 share of each combination.

The label is concentrated. Family A holds 253 of the 325 label 1 paths and family E 70; families D and F1 hold 2 between them, and families B and C1 hold none. Family A was cut almost only in AA 6061 (405 of its paths, 61 percent of them label 1) and mostly at high spindle speeds (87 percent of its label 1 paths at or above 9,000 r/min, against 53 percent in family E), so tool family, workpiece material and spindle speed move together in this design. Counted by runs, the label 1 paths correspond to 8.6 effective runs (the squared total count divided by the sum of the squared counts of the runs), and those outside family A to 3.7. The summary workbook records an early stop for seven runs of imi_vm20i, two of them for excessive vibration.

The second laboratory machine, imi_vmx30ui, a 5-axis mill, has no accelerometers and no Sound1 [9], and its label 1 paths sit in six runs; it enters exploratory analyses only. The production machine tmf_vf10 holds seven label 1 paths and is not analysed.

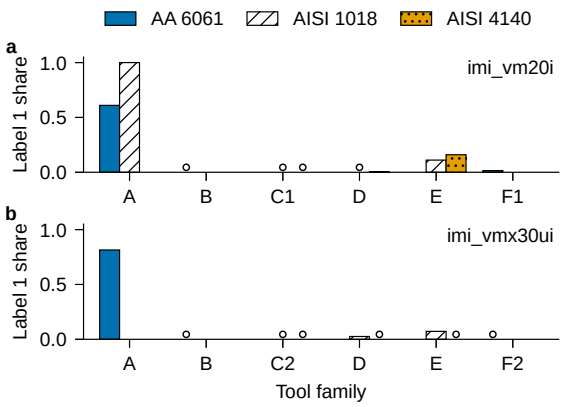



**Fig. 1** Share of label 1 paths by tool family and workpiece material on imi_vm20i (a) and imi_vmx30ui (b). Bar fill and hatching mark the material, open circles mark family and material combinations with paths but no label 1 path, and empty positions mark combinations without paths. Table 1 gives the counts behind every bar and circle

# 3 Methods

## 3.1 Features and models

Each labelled cutting path is one observation. The high-rate feature set holds the logarithms of the root-mean-square (RMS) amplitudes of Acc0, Acc1, Sound0, Sound1 and Sound2 over the path. The MTConnect feature set holds the logarithms of the spindle-motor vibration RMS, peak, crest factor and velocity RMS and of the true power, with the spindle load as recorded, since it is not strictly positive. Both sets are those of the exploratory screen that preceded the predictions.

The detector is a logistic regression on standardised features with inverse regularisation strength 1.0 and balanced class weights. Balancing matters here because the label 1 share of the family-held-out training sets ranges from 7 percent to 28 percent; it removes the part of each fitted intercept that follows the training base rate, and it leaves the shifts that unseen feature levels cause, which are part of what H1 and H2 measure. A simple model suits the question, which is what amplitude detectors learn from these labels rather than how good a detector can be made.

**Table 1** Analysis populations on the laboratory machines by tool family and workpiece material: runs and labelled cutting paths with label 0 (normal) and label 1 (abnormal). Label 2 paths are excluded

| | | imi_vm20i | | | imi_vmx30ui | | |
|---|---|---|---|---|---|---|---|
| Tool family[a] | Material | Runs | Label 0 | Label 1 | Runs | Label 0 | Label 1 |
| A | AA 6061 | 8 | 158 | 247 | 3 | 26 | 114 |
| A | AISI 1018 | 3 | 0 | 6 | | not used[b] | |
| B | AA 6061 | 1 | 64 | 0 | 1 | 64 | 0 |
| C1 | AISI 1018 | 1 | 64 | 0 | | not used | |
| C1 | AISI 4140 | 1 | 64 | 0 | | not used | |
| C2 | AISI 1018 | | not used | | 1 | 64 | 0 |
| C2 | AISI 4140 | | not used | | 1 | 64 | 0 |
| D | AISI 1018 | 1 | 64 | 0 | 2 | 76 | 2 |
| D | AISI 4140 | 5 | 256 | 1 | 1 | 64 | 0 |
| E | AISI 1018 | 7 | 208 | 26 | 5 | 218 | 17 |
| E | AISI 4140 | 5 | 231 | 44 | 1 | 64 | 0 |
| F1 | AA 6061 | 2 | 64 | 1 | | not used | |
| F2 | AA 6061 | | not used | | 1 | 64 | 0 |
| All | | 34 | 1,173 | 325 | 16 | 704 | 133 |

[a]A family joins a tool and its defective twin (A and A*). Each run used one tool family and one workpiece material

[b]Not used: the family was not used with that material on that machine

## 3.2 Evaluation schemes and metric

Run-held-out evaluation assigns whole runs to five folds and repeats the assignment twenty times. Tool-family-held-out evaluation uses a fold per family, six folds on imi_vm20i; holding families out rather than tools keeps a defective twin from leaking its sound tool's geometry into training. Both schemes keep every path of a run on one side of each split. The exploratory analyses add random path-level folds, stratified by label and repeated in the same way, and a fold per workpiece material.

The metric is the pooled out-of-fold area under the receiver operating characteristic curve (AUC) over all paths of the population: the probability that a random label 1 path scores above a random label 0 path, ties counting half [14]. For the run scheme it is the mean over repeats of each repeat's pooled AUC, so that in both schemes every score comes from a single fitted model. H1 and H2 test the fall from the run-held-out to the family-held-out AUC.

## 3.3 Uncertainty

Labels cluster within runs, so runs are the unit of resampling. Every interval and p-value comes from 4,000 bootstrap resamples in which the runs of each tool family are drawn with replacement within the family [15], with the predictions held fixed and one matrix of run weights shared by the schemes, so that differences between schemes are paired. Intervals are percentile intervals at 95 percent. A one-sided p-value is the share of resamples at or below the null value, with the plus-one correction. Holding the predictions fixed leaves out the variability of refitting, so these intervals are narrower than a full refit would give. As a post hoc gauge of how much narrower, the standard error of the delete-one-run jackknife over the refits of the leave-one-run-out check (Section 3.6) is set beside the standard deviation of the bootstrap draws.

## 3.4 Spectral indicator

The spectra are Welch power spectra [16] over each whole path, with a Hann window of 8,192 samples and 50 percent overlap, which gives bins of 6.25 Hz for the accelerometers and 5.86 Hz for the sound channels. Periodic cutting forces repeat with the spindle rotation, so their lines fall on the harmonics of the spindle rotation frequency. A path's indicator is on when the largest line between 200 Hz and 10.0 kHz lies more than two bins from every harmonic of the commanded spindle frequency, and off otherwise (Fig. 2). It has no tuned threshold and treats every sensor alike. It is defined from 3,000 r/min upwards, where the harmonic comb leaves clean gaps, which removes 112 paths, all of them label 0. The MTConnect spindle speed equals the commanded speed on 99.8 percent of paths and cannot correct the comb.

The dataset holds no force, displacement or frequency response measurement, so the indicator marks a spectral pattern and cannot by itself identify chatter.

### 3.5 Agreement between sensors

H3 uses the label 1 paths with both the Acc0 and the Sound2 indicator defined. Their agreement is measured by Cohen's kappa [17],

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

where $p_o$ is the share of paths on which the indicators agree and $p_e = ab + (1-a)(1-b)$ follows from their overall shares $a$ and $b$. Both indicators depend on spindle speed through the comb geometry: the share of the band that lies off the comb grows from 38 percent at 3,000 r/min to 79 percent at 9,000 r/min for Acc0. The null is therefore independence within spindle-speed classes, below and at or above 9,000 r/min: with class weights $w_c$ and class shares $a_c$ and $b_c$, the agreement expected under the null is $p_o^0 = \sum_c w_c[a_c b_c + (1-a_c)(1-b_c)]$, which gives the null kappa $\kappa_0$. The largest kappa the class shares allow, $\kappa_{\max}$, uses $p_o = \sum_c w_c(1 - |a_c - b_c|)$. The threshold lies 50 percent of the way from $\kappa_0$ to $\kappa_{\max}$ and is computed from the data; the test compares $\kappa - \kappa_0$, with $\kappa_0$ recomputed in every resample, with zero. Each speed class holds label 1 paths of more than one tool family.

### 3.6 Pre-specified predictions

The plan fixed the predictions, their decision rules, testability conditions, robustness variants and failure taxonomy before any of their statistics was computed. The plan and the estimator code were frozen with SHA-256 digests at 00:53 UTC on 23 September 2026, and the confirmatory script wrote its results at 01:22 UTC on 23 September 2026. Both times come from the authors' own files, and no third party holds a time-stamped copy of the plan, so a reader can verify what was frozen (Online Resource 2) but not when. The decision rules are

- H1: on the high-rate set, the fall in AUC is at least 0.05 and its Holm-adjusted p-value is below 0.05;
- H2: the same rule on the MTConnect set;
- H3: kappa is at least its threshold and the Holm-adjusted p-value of $\kappa - \kappa_0$ is below 0.05.

A fall of 0.05 from the run-held-out AUC of the exploratory screen would multiply the share of misordered label 1 and label 0 pairs by 7.3 for the high-rate set and by 2.8 for the MTConnect set. H1 and H2 are tested only if every family-held-out training set holds at least 30 label 1 and 30 label 0 paths and every run-held-out training fold at least 30 label 1 paths. H3 is tested only if each speed class holds at least 30 label 1 paths from at least 5 runs. Each prediction also needs label 1 paths from at least 5 effective runs, and at execution it is dropped if more than 1 percent of the resamples leave a statistic undefined or, for H3, if the attainable kappa does not exceed the null. The Holm step-down procedure [18] controls the family-wise error rate at 0.05 over the tested predictions.

A prediction is confirmed when its rule holds and no robustness variant contradicts it, and confirmed but not robust when a variant does. It fails in direction when the estimate lies on the wrong side of the null, in magnitude when it lies short of the threshold, and in precision when it reaches the threshold without a Holm-adjusted p-value below 0.05; a prediction whose conditions fail is not tested. A check script classifies the predictions and writes the tally.

Each robustness variant changes one nuisance choice. V1 and V2 apply to H1, H2 and H3: V1 leaves out the defective-tool twins and V2 the early-stopped runs. V3 and V5 apply to H1 and H2: V3 replaces the logistic model by a random forest [19] of 500 trees with leaves of at least 5 paths and balanced weights, and V5 clips every feature at the training-fold quantiles 1 percent from each end. The others apply to H3: V6a and V6b narrow and widen the comb to one and three bins, V6c widens each harmonic's zone to 0.5 percent of its frequency wherever that exceeds two bins, V7 ends the searched band at 5.0 kHz, and V9 computes the null and the maximum within strata of tool family and speed class. A variant contradicts its prediction only when its whole 95 percent interval lies below the threshold, since the variants run on smaller or noisier versions of the data and a shortfall of their point estimate can come from lost power alone; a variant can change only the class of a prediction whose rule holds.

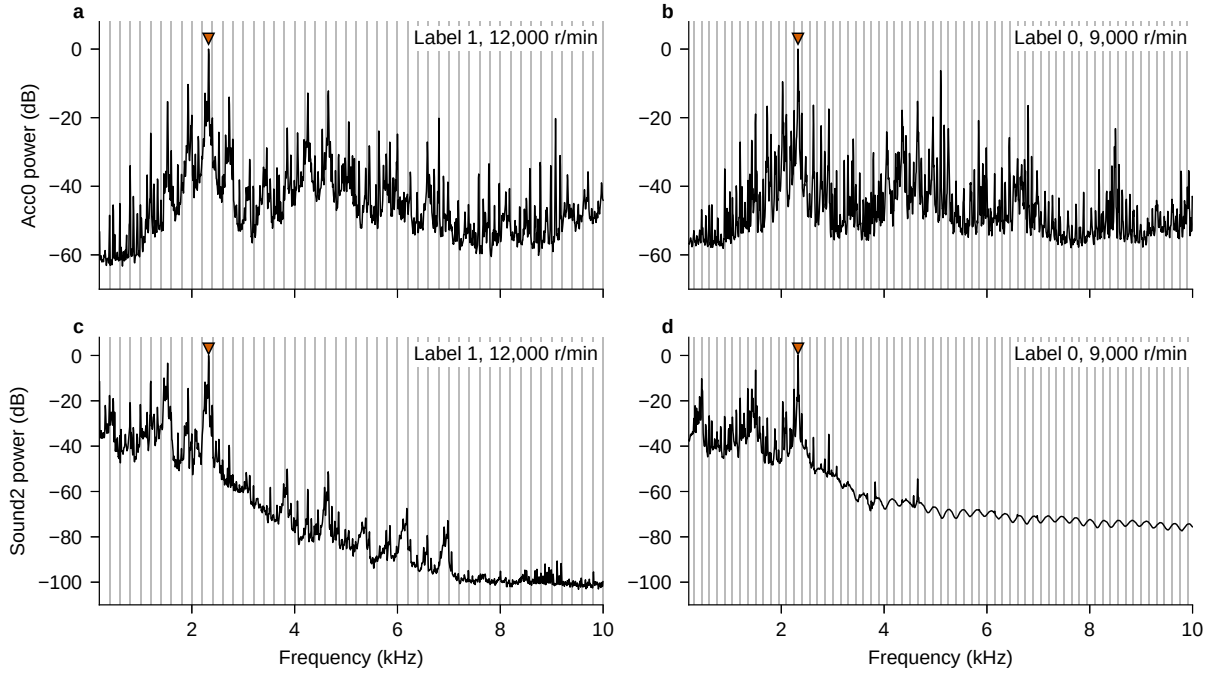



**Fig. 2** Power spectra of Acc0 (a, b) and Sound2 (c, d) for a label 1 path (a, c) and a label 0 path (b, d) of run 25, chosen by the fixed rule of the plan: the run with the most label 1 paths among runs holding both labels, and in it the path whose Acc0 RMS lies closest to the median of its label in the run. Power is relative to the largest line between 200 Hz and 10.0 kHz. Grey bands mark the harmonic comb, within two bins of a spindle harmonic, and the triangle marks the dominant line, which lies off the comb in all panels. In both paths the dominant line of both sensors lies near 2.3 kHz (post hoc frequencies), so the label 0 example carries both lines, as only 2.7 percent of label 0 paths do

Assumption checks accompany the predictions without entering their verdicts: every statistic is recomputed with each run left out in turn, and H3 is repeated on paths of at least 3 s, where entry and exit weigh less, and on the paths with at most 1 percent of their Sound2 samples at full scale.

## 3.7 Exploratory and post hoc analyses

The exploratory analyses of the plan report estimates with run-bootstrap intervals and no p-values:

- E1: full-scale clipping by machine, sound channel and label, and how the Sound0 level flattens as clipping grows;
- E2: the high-rate set without Sound0, and a score made of the Sound0 full-scale share alone;
- E3: models on the design variables, that is spindle speed, depths of cut, feed per tooth, cutting direction and workpiece, with and without tool family;
- E4: the random, run and family schemes for every feature set;
- E5: the indicator shares under the comb and band variants;
- E6: label 0 paths with both lines and label 1 paths with neither;
- E7: a fold per workpiece material;
- E8: imi_vmx30ui, including the MTConnect model of imi_vm20i applied there;
- E9: whether the Acc0 and Sound2 lines share a frequency;
- E10: the pooled AUC split into label 1 and label 0 pairs within one family and pairs across families;
- E11: the Acc0 line share of families A and E within label 1.

Analyses added after the first draft was reviewed are marked post hoc wherever they appear:

- the AUC of the Sound0 full-scale share itself, with no model and no split, on both laboratory machines;

- the jackknife gauge of Section 3.3;
- H3 against independence within runs, with the null, the maximum and the threshold rule computed as for H3 but with runs as strata;
- the frequencies of the dominant off-harmonic lines, as quartiles by label and as their rank correlation with the commanded spindle speed.

The deviations log in Online Resource 2 records each of them with its reason.

### 3.8 Software

The analysis ran in Python 3.11.15 with scikit-learn 1.8.0 [20], NumPy 2.4.4 [21] and pandas 3.0.2 [22]; the figures were drawn with Matplotlib 3.10.9 [23].

## 4 Results

### 4.1 What was seen before the predictions were specified

The results in this subsection come from the exploratory screen that preceded the predictions. They were known when the predictions were written and are not confirmatory.

The Sound0 microphone of imi_vm20i clips. Every label 1 path and 38 percent of label 0 paths have more than 1 percent of their Sound0 samples at full scale (Fig. 3). The Sound0 RMS alone separates the labels with an AUC of 0.99 over all paths and 0.98 within strata of run, tool and depths of cut that hold both labels; the Acc0 RMS gives 0.95 and 0.97. Whether a path was cut with a family A tool gives an AUC of 0.82 on its own. The off-harmonic line is the dominant line in 68 percent of label 1 and 5 percent of label 0 paths for Acc0, and in 60 percent and 4 percent for Sound2.

Exploratory analysis E1 adds how the level behaves as clipping grows. As the share of Sound0 samples at full scale grows, the Sound0 RMS approaches the full-scale level: its median is 30 percent of full scale among paths with at most 1 percent of samples at full scale and 87 percent among paths with more than 30 percent, all of which are label 1. Even among label 0 paths, 95 percent have at least one Sound0 sample at full scale. Sound2 has more than 1 percent of its samples at full scale on 26 percent of label 1 and 0.9 percent of label 0 paths, and Sound1 on 18 percent and 18 percent (Online Resource 4). A loudness label and a saturated sensor are hard to tell apart on this machine.

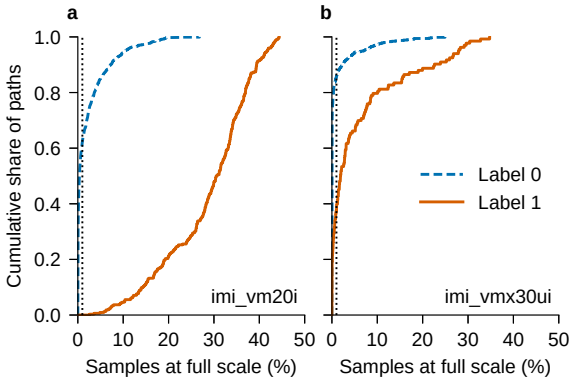



**Fig. 3** Empirical cumulative distributions of the share of Sound0 samples at full scale in each path on imi_vm20i (a) and imi_vmx30ui (b), by label. Dashed line: label 0; solid line: label 1; dotted vertical line: 1 percent of samples at full scale

### 4.2 Pre-specified predictions

All three predictions passed their testability conditions, and no resample left a statistic undefined. Table 2 gives the outcomes: one confirmed, zero confirmed but not robust, two failed (two on magnitude) and zero not tested.

**H1 failed on magnitude.** The high-rate detector ranked the labels almost as well on unseen tool families as on unseen runs: the AUC was 0.992 (0.982 to 0.999) with runs held out and 0.991 (0.977 to 0.999) with families held out (Fig. 4). The fall of 0.002 (−0.001 to 0.007) lies on the predicted side of zero and below the threshold of 0.05, with a Holm-adjusted p-value of 0.22. Leaving out any single run changed the fall by at most 0.014. The post hoc jackknife over those refits gives a standard error of 0.014, 6.2 times the standard deviation of the bootstrap draws, and on that scale the estimate lies 3.5 standard errors below the threshold.

**H2 failed on magnitude.** The MTConnect detector lost more: 0.973 (0.943 to 0.994) with runs held out against 0.940 (0.885 to 0.976) with families held out. The fall of 0.033 (0.015 to 0.060) exceeds zero under the specified bootstrap, with a Holm-adjusted p-value of 0.00075, the smallest that 4,000 resamples allow after adjustment, and it stays below the threshold, which its interval includes. Leaving out single runs gave falls from

0.012 (run 29 left out) to 0.057 (run 16 left out), a range as wide as the bootstrap interval, and leaving out run 16 or run 27 alone would make the rule hold. The post hoc jackknife gives a standard error of 0.045, 3.8 times the standard deviation of the bootstrap draws, and on that scale the fall lies 0.7 standard errors above zero and 0.4 below the threshold. Once refitting is counted, the fall is therefore resolved neither against zero nor against the threshold. The specified class stands. Table 3 sets both detectors beside the other feature sets and evaluation schemes.

**H3 was confirmed.** Among the 325 label 1 paths, the Acc0 and Sound2 indicators agreed with a kappa of 0.71 (0.59 to 0.81). Independence within the speed classes would give 0.004, and the class shares allow at most 0.82, which put the threshold at 0.41. The excess over the null was 0.71 (0.59 to 0.81), with a Holm-adjusted p-value of 0.00075. In both speed classes the paths with both lines outnumber those with one line (Fig. 5a). Leaving out any single run kept kappa above its threshold by at least 0.28. On the 255 paths of at least 3 s, kappa was 0.65 (0.49 to 0.75; threshold 0.39), and on the 241 paths with at most 1 percent of their Sound2 samples at full scale it was 0.66 (0.48 to 0.77; threshold 0.40). Post hoc, with runs as strata, independence would give a kappa of 0.12 and the run shares allow at most 0.81, which puts the threshold rule at 0.46, below the observed kappa; kappa exceeded that null by 0.60 (0.47 to 0.73). For kappa, which involves no fitting, the jackknife standard error of 0.065 is close to the standard deviation of the bootstrap draws, 0.055.

## 4.3 Robustness

Table 4 gives the estimate and interval of every variant. For H3, the variant kappas range from 0.70 to 0.71 against thresholds of 0.40 to 0.46, and no variant's interval lies below its threshold, so H3 is confirmed and robust. The relative comb (V6c) and the band ending at 5.0 kHz (V7) each changed the Acc0 indicator of three paths, none of them label 1, and no Sound2 indicator, so they left H3 as it was. H1 and H2 failed their rules, so their variants could not change their class. For H1 the intervals of V1, V2 and V5 lie wholly below the threshold, while the random forest (V3) lost

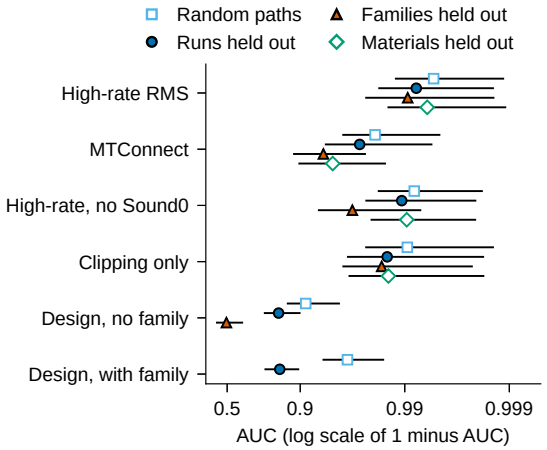



**Fig. 4** Pooled out-of-fold AUC on imi_vm20i by feature set and evaluation scheme, with 95 percent run-bootstrap intervals. Squares: random paths; circles: runs held out; triangles: tool families held out; diamonds: workpiece materials held out. The axis is logarithmic in one minus the AUC, so equal distances mark equal ratios of misordered pairs. Rows as in Table 3; missing markers are cells that are not defined there

0.029 (0.009 to 0.053); for H2 every variant's interval reaches the threshold, and the estimate of the random forest, 0.061 (0.021 to 0.127), exceeds it.

## 4.4 Exploratory and post hoc analyses

**Evaluation schemes (E2 to E4, E7).** A score made of the Sound0 full-scale share alone ranked the labels with an AUC of 0.985 (0.965 to 0.998) on unseen runs and 0.983 (0.961 to 0.998) on unseen families, close to the full high-rate set (Table 3). Without Sound0, the high-rate set scored 0.989 and 0.968, a fall of 0.021 (0.004 to 0.047) between the schemes. Random path-level folds overstated the amplitude models little (0.995 against 0.992 for the high-rate set) but the design-variable models a great deal: without tool family, those variables reached 0.911 on random folds, 0.839 on unseen runs and 0.492 (0.374 to 0.640) on unseen families, an interval consistent with chance ranking. Holding out each of the three workpiece materials gave 0.994 for the high-rate set and 0.951 for the MTConnect set.

**Pairs within and across families (E10).** Of all label 1 and label 0 pairs, 81 percent join paths from different tool families. For the high-rate set, pairs within a family ranked at 0.976 with

**Table 2** Pre-specified predictions on imi_vm20i and their outcomes. Estimates with 95 percent intervals from 4,000 run resamples drawn within tool families, with the fitted predictions held fixed

| | Prediction | Threshold | Estimate (interval) | Holm-adjusted p | Outcome |
|---|---|---|---|---|---|
| H1 | Fall in AUC from runs to families held out, high-rate set | 0.05 | 0.002 (−0.001 to 0.007) | 0.22 | Failed: magnitude |
| H2 | Fall in AUC from runs to families held out, MTConnect set | 0.05 | 0.033 (0.015 to 0.060) | 0.00075 | Failed: magnitude |
| H3 | Kappa of the Acc0 and Sound2 indicators, label 1 paths | 0.41[a] | 0.71 (0.59 to 0.81) | 0.00075 | Confirmed |

[a]Computed at execution by the specified rule, 50 percent of the way from the kappa expected under independence within speed classes (0.004) to the largest attainable kappa (0.82)

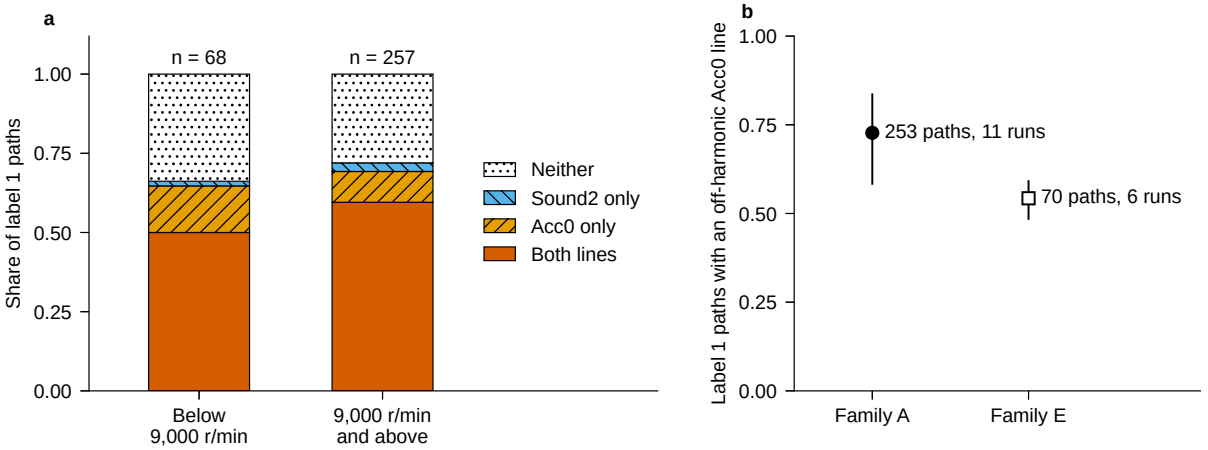



**Fig. 5** Off-harmonic lines among label 1 paths of imi_vm20i. (a) Combinations of the Acc0 and Sound2 indicators below and at or above 9,000 r/min, with the number of paths above each bar (H3). (b) Share of label 1 paths with an off-harmonic Acc0 line in tool families A and E, with 95 percent run-bootstrap intervals (exploratory analysis E11)

runs held out and 0.982 with families held out, and pairs across families at 0.996 and 0.993. For the MTConnect set, the within-family AUC held (0.959 and 0.961) while the across-family AUC fell from 0.976 (0.947 to 0.996) to 0.935 (0.875 to 0.975). Held out, family A kept a within-family AUC of 0.961 and family E 0.960 for MTConnect, and 0.970 and 0.997 for the high-rate set.

**Spectral lines (E5, E6, E9, E11).** The Acc0 share in label 1 ranged from 62 percent with the wide comb to 75 percent with the narrow one, and the Sound2 share from 54 percent to 67 percent; in label 0 neither exceeded 6.2 percent, while Acc1 reached 21 percent of label 0 paths with the narrow comb (Online Resource 4 gives every share). Of the 1,061 label 0 paths, 29 (2.7 percent) carried both lines, 24 of them in family A and 12 in run 6. Of the 325 label 1 paths, 95 (29 percent, 21 to 39 percent) carried neither, spread over 16 runs, 61 of them in family A and 32 in family E. Where both lines were present (187 label 1 paths), they lay within 12.5 Hz of each other on 64 percent (52 to 81 percent) of the paths, against 14 percent (11 to 31 percent) when the Sound2 line of another path of the same run was paired instead. Within label 1, the Acc0 line was dominant on 73 percent (58 to 84 percent) of family A paths and 54 percent (48 to 59 percent) of family E paths (Fig. 5b), a difference of 18 percentage points (4 to 31 percentage points). The family E interval rests on six runs, 3.5 effective, and is likely too narrow. Family A's label 1 paths also ran faster (87 percent at or above 9,000 r/min against 53 percent), so the difference cannot be assigned to the tool.

**Line frequencies (post hoc).** Among label 1 paths with an off-harmonic Acc0 line, the line frequency had a median of 2.0 kHz and quartiles of 1.4 kHz and 2.3 kHz; among label 0 paths with

**Table 3** Pooled out-of-fold AUC on imi_vm20i by feature set and evaluation scheme, logistic models with balanced class weights. Each cell gives the estimate and its 95 percent run-bootstrap interval. The run and family columns of the high-rate and MTConnect rows are the estimates of H1 and H2; every other cell is exploratory

| Feature set[a] | Random paths | Runs held out | Families held out | Materials held out |
|---|---|---|---|---|
| High-rate RMS | 0.995 (0.988 to 0.999) | 0.992 (0.982 to 0.999) | 0.991 (0.977 to 0.999) | 0.994 (0.986 to 0.999) |
| MTConnect | 0.981 (0.961 to 0.995) | 0.973 (0.943 to 0.994) | 0.940 (0.885 to 0.976) | 0.951 (0.898 to 0.985) |
| High-rate, no Sound0 | 0.992 (0.982 to 0.998) | 0.989 (0.977 to 0.998) | 0.968 (0.934 to 0.993) | 0.990 (0.979 to 0.998) |
| Clipping only | 0.991 (0.977 to 0.999) | 0.985 (0.965 to 0.998) | 0.983 (0.961 to 0.998) | 0.986 (0.966 to 0.998) |
| Design, no family | 0.911 (0.868 to 0.957) | 0.839 (0.782 to 0.899) | 0.492 (0.374 to 0.640) | not defined[b] |
| Design, with family | 0.965 (0.940 to 0.984) | 0.843 (0.785 to 0.895) | not defined | not defined |

[a]High-rate RMS: logarithms of the RMS of Acc0, Acc1, Sound0, Sound1 and Sound2. MTConnect: the spindle features of Section 3.1. Clipping only: the share of Sound0 samples at full scale as the only feature. Design: logarithms of spindle speed, axial and radial depth of cut and feed per tooth, cutting direction and workpiece material, with or without tool family

[b]Not defined: the held-out family or material would enter the model only through a term it never saw in training

**Table 4** Robustness variants of the pre-specified predictions on imi_vm20i: estimate and 95 percent run-bootstrap interval of every variant. The threshold is 0.05 for H1 and H2; for H3 it is the value the specified rule gives for the variant

| Variant | Change | H1: fall in AUC | H2: fall in AUC | H3: kappa | H3: threshold |
|---|---|---|---|---|---|
| V1 | No defective-tool twins | −0.006 (−0.013 to 0.005)* | 0.031 (0.013 to 0.060) | 0.71 (0.56 to 0.82) | 0.41 |
| V2 | No early-stopped runs | 0.003 (−0.001 to 0.010)* | 0.038 (0.018 to 0.070) | 0.71 (0.59 to 0.80) | 0.40 |
| V3 | Random forest | 0.029 (0.009 to 0.053) | 0.061 (0.021 to 0.127) | not defined[a] | |
| V5 | Winsorised features | 0.002 (−0.001 to 0.007)* | 0.029 (0.012 to 0.055) | not defined | |
| V6a | Narrow comb | not defined | | 0.70 (0.58 to 0.81) | 0.41 |
| V6b | Wide comb | not defined | | 0.71 (0.58 to 0.81) | 0.46 |
| V6c | Relative comb | not defined | | 0.71 (0.59 to 0.81) | 0.41 |
| V7 | Band ending at 5.0 kHz | not defined | | 0.71 (0.59 to 0.81) | 0.41 |
| V9 | Null within family and speed class | not defined | | 0.71 (0.59 to 0.81) | 0.42 |

The primary estimates are 0.002 for H1, 0.033 for H2 and 0.71 for H3 (Table 2). A variant can change only the class of a prediction whose decision rule holds; H1 and H2 failed theirs, and their variants are shown for completeness

[a]Not defined: the variant changes a setting that the prediction does not use

*The whole interval lies below the threshold

such a line the quartiles were 1.4 kHz and 2.3 kHz. For Sound2 the quartiles were 1.4 kHz and 2.3 kHz in label 1 and 1.4 kHz and 1.6 kHz in label 0. Within label 1 the line frequency rose with the commanded spindle speed, with Spearman rank correlations of 0.53 for Acc0 and 0.47 for Sound2 over 16 and 15 spindle speeds. Online Resource 3 gives the frequency of every line.

**The second laboratory machine (E8).** On imi_vmx30ui, the MTConnect model fitted to every path of imi_vm20i ranked the labels with an AUC of 0.946 (0.876 to 0.988) over 837 paths, 133 of them label 1 in six runs. Its Sound0 microphone clips on fewer paths: 22 percent of its label 0 and label 1 paths have more than 1 percent of their samples at full scale, against 51 percent

on imi_vm20i; by label, 62 percent of its label 1 and 14 percent of its label 0 paths do (Fig. 3b). Post hoc, the full-scale share itself, with no model, ranked the labels with an AUC of 0.860 (0.786 to 0.921) on imi_vmx30ui and 0.991 (0.977 to 0.999) on imi_vm20i. Seen before the predictions were specified, the Sound2 line of imi_vmx30ui is dominant on 59 percent of label 1 but also on 27 percent of label 0 paths, against 60 percent and 4 percent on imi_vm20i, and whether a path was cut with a family A tool ranks the labels of imi_vmx30ui with an AUC of 0.91, against 0.82 on imi_vm20i.

# 5 Discussion

## 5.1 What the labels encode

On imi_vm20i, the abnormal label goes with a saturated microphone. In results seen before the predictions were specified, every label 1 path has more than 1 percent of its Sound0 samples at full scale; in exploratory analyses the Sound0 level approaches full scale as clipping grows, and the full-scale share alone ranks the labels about as well as the full set of sensor amplitudes, on unseen runs, tool families and materials alike. The data cannot show how the annotators judged the paths; the descriptor says that the cutting intervals were refined with the sound recordings [9]. Whatever the route, a detector trained on these labels can reach a high AUC by recognising a loud, clipped microphone. If the aim is to detect unstable cutting in general, that is a shortcut in the sense of Geirhos et al. [6], because saturation depends on the gain and placement of one microphone. On imi_vmx30ui the relation is weaker but present: only 62 percent of label 1 paths pass the 1 percent criterion there, and the full-scale share ranks the labels with an AUC of 0.860 (post hoc). A clipping threshold set on one machine therefore need not carry over to another, even where the ranking does.

The pre-specified expectation was that amplitude models would lose ranking performance on tool families they had not seen, with the materials and speeds that come with them, because label 1 sits mostly with family A. For the high-rate logistic model the fall was 0.002 (−0.001 to 0.007), an interval that lies wholly below the threshold, and in exploratory analyses the model also ranked well when whole workpiece materials were left out. That finding does not extend to other models or feature sets. The random forest lost 0.029 on the same features, the high-rate set without Sound0 lost 0.021 (exploratory), and the MTConnect fall of 0.033 is unresolved once refitting is counted. In the exploratory split of the pooled AUC, the MTConnect ranking within families held while the ranking across families fell, which is consistent with a loss in placing an unseen family's spindle load, power and vibration levels on a common scale; no interval was computed for the difference between the within-family and across-family parts. Trees do not extrapolate beyond the feature levels seen in training, which fits the larger losses of the random forest, so the size of the penalty for unseen families depends on the model class as well as on the data.

Among abnormal paths, an accelerometer on the spindle and a stethoscope sensor on the base column, transducers of different kinds at different places, carry a dominant line off the spindle harmonics on the same paths. The agreement exceeds what independence within the spindle-speed classes gives (H3) and, post hoc, what independence within runs gives, and in exploratory analyses the lines of both sensors usually share a frequency. An artefact of a single sensor, such as clipping, electrical interference or a mounting resonance, would not produce this pattern, which points to a vibration of the machine structure. The lines of label 0 and label 1 paths fall in the same frequency range, between about 1.4 kHz and 2.3 kHz for Acc0, and their frequency rises with spindle speed (post hoc). Chatter produces lines of this kind, and so can a structural mode that the cutting excites without self-excited vibration; the dataset holds no force, displacement or frequency response data that could distinguish these sources, and the descriptor states that the abnormal label covers several causes [9]. The label 1 paths without any line (29 percent) are consistent with those other causes. The label 0 paths with both lines cluster: 24 of the 29 are in family A and 12 in a single run. They are candidates for unstable cutting that was not labelled, or for lines that family A produces at high speed, and these data cannot tell these cases apart.

### 5.2 How to evaluate detectors on this dataset

The results lead to practices for users of the dataset. First, split by run. In exploratory analyses, random path-level folds changed little for the amplitude models but raised the design-variable models from 0.839 on unseen runs to 0.911, because paths of one run share their tool, workpiece and settings; a result on random folds cannot be compared with one on grouped folds [3, 4]. Second, report a clipping baseline. The share of Sound0 samples at full scale reached an AUC of 0.985 on unseen runs of imi_vm20i (exploratory), and 0.991 with no model at all (post hoc); a detector that does not beat it has not shown that it detects anything beyond microphone saturation. Third, a claim that a detector transfers to unseen tools needs whole tool families, with their materials and speeds, held out in its evaluation. This is a design rule for such claims rather than a finding of this study, whose confirmatory falls stayed below their threshold. When families differ in their base rates, the AUC within families should accompany the pooled AUC.

### 5.3 Limitations

The confirmatory tests rest on a single machine, on which tool family, workpiece material and spindle speed move together, and family A holds most label 1 paths; the label 1 paths correspond to 8.6 effective runs and those outside family A to 3.7. The small number of clusters limits every interval: family B is a single run and adds no variance to the bootstrap, and the family E interval of E11 rests on six runs. The bootstrap holds the fitted predictions fixed. The post hoc jackknife over refits puts the standard error of the fall at 6.2 times that of the bootstrap for H1 and 3.8 times for H2, against 1.2 times for kappa, which involves no fitting; runs 16, 27 and 29 dominate it, so the intervals of H1 and H2 are too narrow and the jackknife itself is only a rough gauge. The thresholds are judgments specified with their reasons, and no external convention was consulted. The plan was frozen with digests that the authors hold, and no third party holds a time-stamped copy. The spectra cover whole paths, entry and exit included, although H3 held on the longer paths. The loudness and clipping results were seen before the predictions were specified and are descriptive. Finally, the spectral indicator has no ground truth, so H3 shows agreement between sensors; it does not show that the label marks chatter.

## 6 Conclusions

On the laboratory machine of the MSM dataset that carries accelerometers, one of three prespecified predictions held and two failed. The high-rate logistic model lost little ranking performance on tool families it had not seen, with their materials and speeds (a fall of 0.002); the fall of the MTConnect model, 0.033, stayed below the threshold but is unresolved once refitting is counted. A spindle accelerometer and a stethoscope sensor on the base column marked the same abnormal paths with an off-harmonic spectral line, beyond what independence within spindle-speed classes gives and, post hoc, beyond independence within runs; the line may come from chatter or from a structural mode that the cutting excites. In results seen before the predictions were specified, every abnormal path had more than 1 percent of one microphone's samples at full scale, and in exploratory analyses that share alone ranked the labels almost as well as every sensor amplitude together, with an AUC of 0.985 on unseen runs. Detector comparisons on the dataset should split by run and report a clipping baseline, and claims of transfer to unseen tools need tool families held out.

**Supplementary information.** Online Resource 1 (ESM_1.csv): fold assignments of the 1,498 label 0 and label 1 paths of imi_vm20i in every evaluation scheme, with each path's run, tool family and workpiece material. Online Resource 2 (ESM_2.zip): the analysis code, the frozen plan with its SHA-256 digests, the values files, the deviations log and the check results. Online Resource 3 (ESM_3.csv): the off-harmonic indicators of the same paths for Acc0, Acc1, Sound1 and Sound2, with the frequency of the dominant off-harmonic line of Acc0 and Sound2. Online Resource 4 (ESM_4.csv): the shares of clipped paths by machine, sound channel and label (E1) and of paths with an off-harmonic dominant line by channel, indicator setting and label (E5).

## Statements and Declarations

**Funding.** No funding was received for conducting this study.

**Competing interests.** The authors have no relevant financial or non-financial interests to disclose.

**Ethics approval and consent to participate.** Not applicable: the study analyses a public dataset and involves no human participants, human data or animals.

**Consent for publication.** Not applicable.

**Data availability.** The MSM dataset analysed in this study is openly available on Kaggle under the CC BY 4.0 licence at https://doi.org/10.34740/kaggle/ds/8392825 [9, 10]. The data generated in this study, the split assignments, per-path indicators and summary tables, are Online Resources 1, 3 and 4.

**Materials availability.** Not applicable.

**Code availability.** The analysis code, the frozen plan and the values files are Online Resource 2.

**Author contributions.** Adam Park: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Project administration. Luke Yin: Writing – review & editing. Cooper Lee: Formal analysis, Data curation. Andrew Wong: Methodology. Yoo Choi: Supervision.

**ORCID identifiers.** Adam Park: 0009-0005-7400-4383; Luke Yin: 0009-0005-1681-2825; Cooper Lee: 0009-0006-3338-3580; Andrew Wong: 0009-0003-3480-2765; Yoo Choi: none.

## References

[1] Navarro-Devia JH, Chen Y, Dao DV, et al (2023) Chatter detection in milling processes: a review on signal processing and condition classification. Int J Adv Manuf Technol 125(9-10):3943–3980. https://doi.org/10.1007/s00170-023-10969-2

[2] Delio T, Tlusty J, Smith S (1992) Use of audio signals for chatter detection and control. J Eng Ind 114(2):146–157. https://doi.org/10.1115/1.2899767

[3] Roberts DR, Bahn V, Ciuti S, et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. https://doi.org/10.1111/ecog.02881

[4] Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4(9):100804. https://doi.org/10.1016/j.patter.2023.100804

[5] von Hahn T, Mechefske CK (2022) Machine learning in CNC machining: best practices. Machines 10(12):1233. https://doi.org/10.3390/machines10121233

[6] Geirhos R, Jacobsen JH, Michaelis C, et al (2020) Shortcut learning in deep neural networks. Nat Mach Intell 2(11):665–673. https://doi.org/10.1038/s42256-020-00257-z

[7] Smith WA, Randall RB (2015) Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study. Mech Syst Signal Process 64-65:100–131. https://doi.org/10.1016/j.ymssp.2015.04.021

[8] Matania O, Cohen R, Bechhoefer E, et al (2024) Test-training leakage in evaluation of machine learning algorithms for condition-based maintenance. PHM Soc Eur Conf 8(1):13. https://doi.org/10.36001/phme.2024.v8i1.4125

[9] Kim E, Sim Y, Li AS, et al (2026) Multisensor and MTConnect dataset of metal cutting anomaly in milling from laboratory and industry settings. Sci Data 13(1):945. https://doi.org/10.1038/s41597-026-07255-7

[10] Kim E, Sim Y, Li AS, et al (2025) Metal milling anomaly dataset. Kaggle, https://doi.org/10.34740/kaggle/ds/8392825, dataset, version 2

[11] Kim E, Bui T, Yuan J, et al (2024) Online real-time machining chatter sound detection using convolutional neural network by adopting expert knowledge. Manuf Lett 41:1386–1397. https://doi.org/10.1016/j.mfglet.2024.09.165

[12] Parmar HK, Raman S (2026) AI/ML-based interoperable anomaly detection in advanced manufacturing using MTconnect-derived multi-machine data. Int J Adv Manuf Technol 145(7-8):5361–5379. https://doi.org/10.1007/s00170-026-18611-7

[13] Karkadakattil A (2026) A physics-guided unified damage index framework for explainable milling anomaly detection using interval-wise MTConnect multi-sensor fusion. J Intell Manuf Spec Equip pp 1–33. https://doi.org/10.1108/jimse-05-2026-0020

[14] Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747

[15] Field CA, Welsh AH (2007) Bootstrapping clustered data. J R Stat Soc Ser B Stat Methodol 69(3):369–390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

[16] Welch P (1967) The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 15(2):70–73. https://doi.org/10.1109/TAU.1967.1161901

[17] Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104

[18] Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70. URL https://www.jstor.org/stable/4615733

[19] Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

[20] Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12(85):2825–2830. URL https://jmlr.org/papers/v12/pedregosa11a.html

[21] Harris CR, Millman KJ, van der Walt SJ, et al (2020) Array programming with NumPy. Nature 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2

[22] McKinney W (2010) Data structures for statistical computing in Python. In: Proceedings of the Python in Science Conference, pp 56–61, https://doi.org/10.25080/Majora-92bf1922-00a

[23] Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.csv
- ESM3.csv
- ESM4.csv
- a1leaveonerunout.csv
- check3result.json
- check4result.json
- d01accpass.py
- d02audiovideopass.py
- d03segmentfeatures.py
- deviations4.json
- figureinspection.json
- ijamtrecenttitles.json
- linefrequencies.csv
- lintallow4.json
- p00sources.py
- p01inventory.py
- p02profile.py
- p03stats.py
- p04screen.py
- p05ideas.py
- p06report.py
- p07lint.py
- p08check.py
- perpathindicators.csv
- phase3plan.md
- plan.json
- plantemplate.md
- q00env.py
- q01venue.py
- q02judgments.py
- q03population.py

- q04check.py
- q05synthetic.py
- q06freeze.py
- qcommon.py
- qconfirm.py
- qrender.py
- qrules.py
- r01confirm.py
- r02checks.py
- r03explore.py
- r04figures.py
- r05facts.py
- r06lint.py
- r07build.py
- r08tally.py
- r09esm.py
- r10check.py
- r11review.py
- r12package.py
- rfast.py
- rfreeze.py
- references.json
- reporttemplate.md
- segtable.py
- segments.parquet
- sources4.json
- vals.py
- values.json
- values3.json
- values4.json