

Journal of Intelligent Digital Systems and Innovation

available at http: www.econjournals.com

Journal of Intelligent Digital Systems and Innovation, 2026, 2(3), 1-10.

Alarm Thresholds in Intelligent Systems for Acoustic Machine Monitoring

Massila Kamalrudin[1]*, Mustafa Musa[2]

[1]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malacca, Malaysia, [2]Center of Research and Innovation Management, Universiti Teknikal Malaysia Melaka, Malacca, Malaysia. *Email: massila@utem.edu.my

Received: 01 July 2026 **Accepted:** 20 September 2026 **DOI:** https://doi.org/10.32479/jidsi.25436

ABSTRACT

Nine acoustic monitoring systems tied for the highest mean F1 in a machine-sound benchmark. All nine reported the precision and recall of a rule that sends every recording for inspection. Across 35 teams and eight machine types in DCASE 2025, the official winner ranked twelfth when judged by the mean F1 of its submitted alarm decisions, and the Spearman correlation between official score and mean F1 was 0.425. How well did individual machines fare? Of 280 team–machine results, 20 exceeded the F1 of the alarm-on-every-clip rule and 21 reached at least 60% on both precision and recall. A separate May 2026 study supplied 129 chainsaw recordings. Across four thresholds on those recordings, detections fell from nine to five without eliminating an observed false alarm. Under independent sampling, zero false alarms in 120 normal recordings gives a one-sided 95% upper false-positive rate of 2.47%. For a plant choosing an acoustic monitor, the useful evidence is a machine-specific alarm decision at a declared threshold, accompanied by the counts of normal and faulty recordings on which it was tested.

Keywords: Intelligent Systems, Acoustic Anomaly Detection, Machine Condition Monitoring, Alarm Thresholds, Predictive Maintenance
JEL Classifications: C44, C88, O33

1. INTRODUCTION

A maintenance engineer can use an acoustic detector in two quite different ways. One is to review recordings in order, starting with the sound least like that of a healthy machine. The other is to let the detector decide which machines need inspection. The score can serve the first purpose even if the engineer knows very little about where to set the cutoff for the second, yet that cutoff will determine how many healthy machines join the queue for inspection.

A microphone can pick up the operating sound of a moving part without touching it. It also picks up neighbouring machines. If operating speed or background noise changes between training and inspection, a perfectly healthy run can sound unfamiliar, while the brief sound of a defect may account for only a small part of the clip (Qurthobi et al., 2022). Setting the threshold after the test labels are known gives the analyst an advantage the engineer will not have. The 2025 detection and classification of acoustic scenes and events challenge (DCASE) make the distinction open to empirical study. Teams supplied scores for ranking the recordings and decisions for flagging individual clips. Only the scores determined the official ranking. The organisers also published precision, recall and F1 for the submitted decisions, so both sides of the comparison can be examined on the same machines without fitting another model or moving a threshold in hindsight.

We ask how often a team with a high official score also produced useful alarms on the individual machines, taking a rule that flags every recording as the simplest reference against which to judge its F1. A second question requires a different design. Given one model and a fixed set of recordings, what does changing the threshold buy in observed false alarms, and what does it cost in missed faults? Four published chainsaw inspection thresholds provide that comparison.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

2. RELATED WORK

2.1. Learning an Acoustic Anomaly Score

Di Fiore et al. (2022) incorporated machine identity into an acoustic model for predictive maintenance. Ota and Unoki (2023) used short-time features and timbral descriptors to distinguish abnormal sounds. A reconstruction model may score an unfamiliar sound highly because it cannot reproduce it, while a classifier can give the same sound a high score because its timbre resembles examples of a fault rather than because the sound is hard to reconstruct. Both scores still need a cutoff.

Transfer between operating conditions poses a particular difficulty. Wilkinghoff and Kurth (2024) examined angular margin losses in semi-supervised acoustic anomaly detection. Lin et al. (2026) used prototype learning across domains, and Chen et al. (2025) designed a dual-decoder autoencoder for machine sounds under domain shift. A simple shift in score scale illustrates the remaining problem: adding a constant to every score preserves every pairwise ordering, yet a machine whose sound was just below the original cutoff can suddenly be flagged for inspection with no change in its condition.

The available training examples constrain how a cutoff can be checked. Kuroyanagi et al. (2025) selected pseudo-anomalous recordings and used pseudo-labels when operating states were scarcely labelled. Kuroyanagi et al. (2025) considered what could be learned from a small set of actual anomalies. With only normal sounds, an engineer can estimate how often healthy recordings exceed the proposed cutoff, but even an accurate estimate of the upper tail of the normal scores cannot tell that engineer how many genuine faults will fall below the proposed cutoff.

2.2. Evaluating a Decision

Machine type enters acoustic models in several ways. Zhang et al. (2025) selected features according to the sound characteristics of the machine, Zhou et al. (2026) combined local and longer-range spectral information using attention, and Wu et al. (2026) examined a lightweight temporal pyramid. If a plant adopts one of these representations and the distribution of scores changes, its tolerance for false alarms may stay exactly as it was, while the score at which an inspector should be called may have to change.

Gaudreault and Branco (2024) examined how imbalance and noise affect classifier assessment. Sørbø and Ruocco (2024) traced differences among anomaly metrics to the tasks for which they are used. In this study the decision unit is the clip. Ten alarms from adjacent recordings of one machine can lead to a single inspection, whereas alarms on ten separate machines can require ten visits to different parts of the plant, even if the clip-level precision and recall in the two cases are identical. The benchmark supplies no maintenance or downtime records.

3. MATERIALS AND METHODS

3.1. Study Design and Data Sources

The two records available by May 31, 2026 offer different comparisons. DCASE 2025 Task 2 places the team submissions on a common set of eight machine types. Ashourpour and Azizpour (2026) instead hold one chainsaw detector and its test recordings fixed while changing the threshold 4 times. The first record compares teams and the second compares thresholds, so pooling them would sacrifice the distinction between variation attributable to different submitted systems and variation that can be observed with the same detector, test labels and recordings (Table 1).

We joined the organisers' "Teams ranking" table to the table of binary decisions by submission identifier and machine type. We used the one row out of 36 that belongs to the organisers' Mahalanobis baseline to check the extraction, then removed it before calculating any team statistic so that the comparison included entrants on the same footing. All 35 remaining identifiers matched on every evaluation machine. The seven development machines were excluded.

We retain the organisers' names for all eight machine types so that a reader comparing a result in the tables with the public record can locate it directly: AutoTrash, BandSealer, CoffeeGrinder, HomeCamera, Polisher, ScrewFeeder, ToyPet and ToyRCCar. Within each source or target condition the test design supplies 50 normal and 50 anomalous clips, giving 200 clips for each machine when the two conditions are combined. The anomalous fraction is therefore one half, a feature that enters the precision calculation even when the detector is held fixed.

Although the DCASE 2026 schedule placed additional training data in May, the evaluation data were due on 01 June and the results on 30 June, leaving no paired scores and alarm decisions from the 2026 competition within the cutoff. The chainsaw results were already available: the four held-out operating points were published on May 21, 2026.

Primary records: DCASE 2025 Task 2 results (https://dcase.community/challenge2025/task-first-shot-unsupervised-anomalous-sound-detection-for-machine-condition-monitoring-results) and the archived evaluation dataset (https://doi.org/10.5281/zenodo.15519362). Industrial counts are from Table 2 of Ashourpour and Azizpour (2026).

3.2. Discrimination and a Fixed Decision Threshold

For a recording x, write s(x) for the anomaly score and t for the system's cutoff. An alarm occurs when s(x) exceeds t. AUC

Table 1: Samples and units of analysis

Source or subset	Systems	Machines	Clips or cells	Unit
DCASE team panel	35	8	280	Team–machine cell
Organiser baseline	1	8	8	Check only
DCASE normal clips	—	8	800	Clip
DCASE anomalous clips	—	8	800	Clip
Chainsaw inspection	1	1	129	Test recording
Chainsaw normal	1	1	120	Test recording
Chainsaw anomalous	1	1	9	Test recording
Chainsaw cutoffs	1	1	4	Operating point

The same 1,600 DCASE test clips underlie all 35 team profiles. The four chainsaw cutoffs reuse the same 129 recordings

in equation (1) can be interpreted by drawing one anomalous recording and one normal recording from the evaluation population and asking how often the anomalous recording receives the higher score, with half credit when the two scores tie.

$$a(x) = 1\{s(x) > t\},\ AUC = Pr(s_1 > s_0) + 1/2\ Pr(s_1 = s_0) \quad (1)$$

Rescaling every score by the same strictly increasing function leaves all comparisons in equation (1) intact. A cutoff fixed on the original scale need no longer select the same clips, and an engineer seeking to recover the former decisions would have to know how the scale had changed in order to apply the corresponding transformation to the cutoff. The curve in Figure 1 is illustrative.

DCASE combines source- and target-domain AUC with partial AUC (pAUC) in the region up to a false-positive rate of 0.10. The official score is the organisers' harmonic mean. We retain it as published. The machine AUC and pAUC columns alongside that score describe useful features of a submission, but substituting their arithmetic mean would change the weight given to each component of performance and hence the quantity that determined the official ordering of the teams. We use mean machine pAUC only in a separate comparison.

TP, FP and FN count true positives, false positives and false negatives. We checked the column alignment by reconstructing F1 from displayed precision and recall using equation (2). Because the check yielded a maximum absolute difference below 0.012% point and the displayed precision and recall have already been rounded, we retain the published F1 whenever deciding whether a machine result clears a boundary in the analysis.

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F_1 = \frac{2PR}{P+R}. \quad (2)$$

3.3. Reference Decisions and Rank Comparisons

A service that alarms on every clip catches every fault. If the anomalous fraction of the test set is π, that same fraction of its alarms is genuine, so equation (3) awards the rule an F1 of 66.67% on the balanced DCASE test even though the engineer must inspect every healthy recording too.

$$R_{all} = 1, P_{all} = \pi, F_{1,all} = \frac{2\pi}{1+\pi}. \quad (3)$$

We reserve "reference pattern" for the three displayed values together: 50.00% precision, 100.00% recall and 66.67% F1. A deliberate threshold, a default decision and a submission error could all yield this pattern, so we infer nothing about the entrant's intent from these values. A zero triple means that all three displayed metrics are zero. Zero recall alone does not identify the false-alarm count.

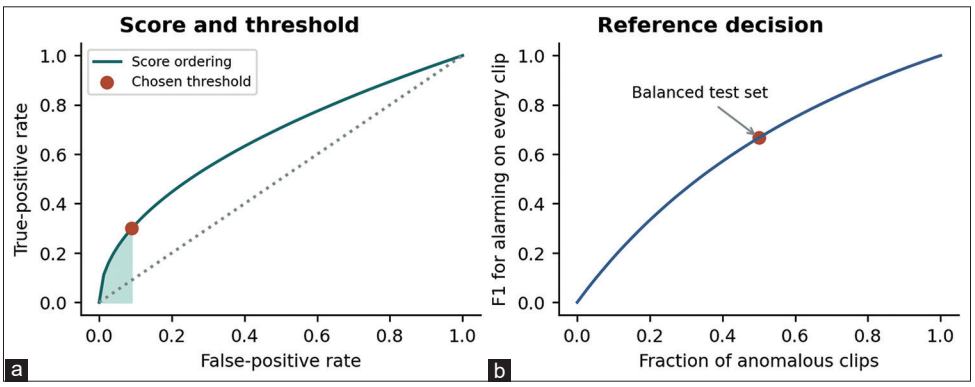
Each team contributes one observation to the system-level correlation. We first average its F1 across the eight machine types with equal weights, then compare the resulting ranks of the 35 representatives with their official-score ranks using Spearman correlation. Ties receive their average rank. To distinguish ties from disagreements, we also calculate Kendall's τ-b and count

Table 2: Ten leading representatives by official score

Representative	Official rank	Score	Mean pAUC	Mean F1	F1 rank	Passes
Wang_MYPS/3	1	61.63	59.34	61.61	12.00	3
Saengthong_SCITOK/2	2	61.57	59.05	53.89	18.00	1
Yang_NBU/1	3	61.20	60.05	49.16	23.00	0
Fujimura_NU/1	7	59.99	58.67	48.10	26.00	1
Jiang_THUEE/2	10	59.79	58.39	60.82	13.00	3
Zheng_SJTU-AITHU/2	11	59.50	57.99	59.59	14.00	3
Guan_HEU/1	22	58.25	57.57	40.84	27.00	0
Ozeki_MELCO/1	23	58.23	56.16	66.67	5.00	0
Zhou_XAUAT/1	24	58.19	56.35	62.29	11.00	1
Huang_XJU/1	25	58.14	55.43	66.67	5.00	0

Percentage units except ranks and passes. Official ranks refer to all submitted configurations; F1 ranks use 35 representatives. Passes count machines attaining 60% on both precision and recall. The suffix after the slash is the submitted configuration

Figure 1: (a and b) A detector score and a reference decision. Panel A is a schematic ROC curve. The dot marks one operating point, and the shaded region illustrates the low-FPR range. Panel B gives the analytic F1 of an alarm on every clip



the pairs of submissions for which selecting on the published official score gives the opposite order from selecting on the alarm decisions of the same two teams.

$$\bar{F}_i = \frac{1}{8}\sum_{m=1}^{8} F_{im}, \rho_S = \mathrm{corr}\left(\mathrm{rank}\left(\Omega_i\right), \mathrm{rank}\left(\bar{F}_i\right)\right). \quad (4)$$

The coverage count in equation (5) preserves the machine identity. A machine counts for team i only when the submitted decision satisfies both a recall floor r and a precision floor p. We vary the two floors from 50% to 80%. The submissions stay fixed. When a stricter floor removes a machine from the count it is the operator's requirement that has changed, since neither the detector's numerical cutoff nor any of the alarms it issued has been altered for this comparison.

$$C_i\left(r,p\right) = \sum_{m=1}^{8} 1\{R_{im} \geq r, P_{im} \geq p\} \quad (5)$$

We resampled 35 team identifiers with replacement 5,000 times (seed 20260531). Each draw carries the entire eight-machine profile of the selected team, so the central 95% resampling range describes how Spearman correlation varies with the mix of entrants while preserving the dependence among results from the same submission. We make no population claim for industrial detectors. Separate analyses omit each machine in turn or exclude teams displaying specified decision patterns.

3.4. Industrial Thresholds and Uncertainty

The chainsaw test contained 120 normal and nine anomalous recordings. We used the counts in Table 2 of Ashourpour and Azizpour (2026) for four training-derived thresholds: the 97th percentile, the adaptive rule, the mean plus one standard deviation and the mean plus three standard deviations. From those counts we calculated recall, F1 and exact equal-tailed 95% binomial intervals for recall. All four thresholds face the same nine anomalies.

If the false-positive rate is f, then the probability of zero false alarms in n independent normal trials is (1-f) raised to the nth power. Setting that probability equal to 0.05 and solving for f gives the one-sided 95% upper limit in equation (6), whereas solving the same expression for the integer n gives the minimum number of normal trials needed for a specified limit following zero false alarms.

$$f_U = 1 - 0.05^{1/n}, n_{\min} = \frac{\log\left(0.05\right)}{\log\left(1-f_0\right)}. \quad (6)$$

The binomial interpretation depends on independent, comparable trials and a cutoff that was fixed before the test. Several clips from one run could share a noise source or a mechanical condition. We report limits for the chainsaw counts conditional on this assumption, but applying the same calculation to our DCASE panel would require run identifiers or individual clip decisions that neither of the two published summary tables supplies (Table 3).

Equation (7) expresses precision in terms of service prevalence π, recall r and the false-positive rate f. Its denominator is the fraction of clips that trigger an inspection. Multiplying each component by 1,000 gives expected true and false alarms per thousand clips. Prevalence is a scenario input.

Table 3: Endpoints and analysis denominators

Analysis	Teams	Machines	Observations	Weight or unit
Official score versus mean F1	35	8	35	One per team
Machine associations	35	1	35	Per machine
Reference comparison	35	8	280	One per cell
Joint floors	35	8	280	One per cell
One machine omitted	35	7	245	35 team means
Team resampling	35	8	5,000 draws	Team profile
Industrial recall	—	1	9	Anomalous clips
Industrial false alarms	—	1	120	Normal clips

A denominator of 280 describes the number of published operating points. Dependence among cells belonging to the same team is retained in the resampling analysis

$$P\left(\pi, r, f\right) = \frac{\pi r}{\pi r + \left(1-\pi\right) f}. \quad (7)$$

Because the four chainsaw thresholds have the same denominators of normal and faulty recordings, we can count directly whether a change in cutoff that missed more faults also spared the inspector any of the false alarms observed at the lower cutoff. No comparison of the models or operating noise between the chainsaw and DCASE studies is required.

4. RESULTS

4.1. Official Rank and the Quality of Submitted Alarms

The Spearman correlation between official score and mean F1 was 0.425 across the 35 systems. Kendall's τ-b was 0.303. Of the 595 system pairs, 367 retained their order, 192 reversed and 36 tied on mean F1, so an operator choosing between two teams with a definite ordering on both criteria would face a reversal in 34.35% of the 559 such comparisons. The central 95% team-resampling range for Spearman correlation ran from 0.090 to 0.673.

Wang_MYPS_task2_3 won the official competition with 61.63%. Its mean F1 of 61.61% placed it twelfth among the team representatives. The next two official systems placed eighteenth and twenty-third on mean F1. Both rankings use the same recordings. The difference arises because the official competition evaluates scores over possible thresholds while the F1 comparison asks what happened at the particular cutoff that each team used when it actually assigned a decision to every recording (Table 2).

At the top of the F1 ordering, nine teams tied on 66.67%. Every one of them reported the reference pattern on every machine. Figure 2 shows the nine as a horizontal band spanning several official scores: each would outrank the official winner if mean F1 alone determined the choice, although its reported precision on every one of the eight machines was just 50%. Changing the ranking metric would thus change which deficiency escaped notice.

4.2. Machine-Level Differences

Table 4 compares the same 35 teams on every machine. No change of entrant explains a difference between rows. We take the median separately for each metric, so the entrant at median precision need not be the entrant at median recall. Substituting these two medians into equation (2) need not give median F1 (Figure 3).

AutoTrash and CoffeeGrinder differed by 17.86 points in median F1. For AutoTrash the medians were 76.28% AUC and 66.67% F1; for CoffeeGrinder they were 51.36% and 48.81%. The ordering changes if the operator looks instead at median pAUC: ScrewFeeder has the highest value at 62.79%, although its median F1 falls just below that of AutoTrash when the comparison uses the alarms actually submitted for those machines. The ordering is specific to these recordings.

On AutoTrash, the across-team Spearman correlation between AUC and F1 was 0.693. On ToyRCCar it was 0.023. CoffeeGrinder gave a correlation of −0.133 between pAUC and F1, but with 35 teams sharing the same clips and several returning tied alarm metrics, that value offers little evidence that improving score ordering would generally worsen a coffee grinder's alarms.

4.3. Alarm Patterns and the Operating Point of the Official Winner

Figure 4 contains 280 operating points. Of these, 80 have the reference pattern (28.57%), and another 24 show zero precision, recall and F1 (8.57%). These concentrations are easy to underestimate visually. Teams with different anomaly scores occupy precisely the same point whenever their displayed precision and recall coincide.

Only 20 results exceeded 66.67% F1. AutoTrash supplied eleven, ScrewFeeder six, and BandSealer, HomeCamera and ToyPet one apiece. CoffeeGrinder, Polisher and ToyRCCar supplied none. An operator facing costly inspections might still choose a decision with lower F1, because the reference rule calls for inspection of every normal recording and the metric assigns no explicit cost to that work (Table 5).

The official winner's F1 ranged from 44.00% on ToyRCCar to 83.02% on ScrewFeeder, so the same submission produced a difference of 39.02 points between its best and worst machines without any change of team or benchmark protocol. On ScrewFeeder the winning system reached 83.05% precision and 82.99% recall, whereas on ToyRCCar both fell below 45%, a distinction that would matter more than its standing in the overall competition to an engineer who needed to monitor only one of the two machines. A supervisor choosing the system for one of these machines would need that machine's result. The summary scores do not identify whether a threshold change would close the gap.

4.4. Joint Precision and Recall Requirements

At floors of 60% precision and 60% recall, 21 of 280 team–machine pairs qualified (7.50%). Thirteen teams qualified at least once, and the most any team covered was three of the eight machines.

Figure 2: (a and b) Official score and submitted alarm performance. Each point represents one team. Colour indicates the number of machines displaying the alarm-on-every-clip reference pattern. The diagonal in panel B indicates equal ranks; tied teams share a midrank

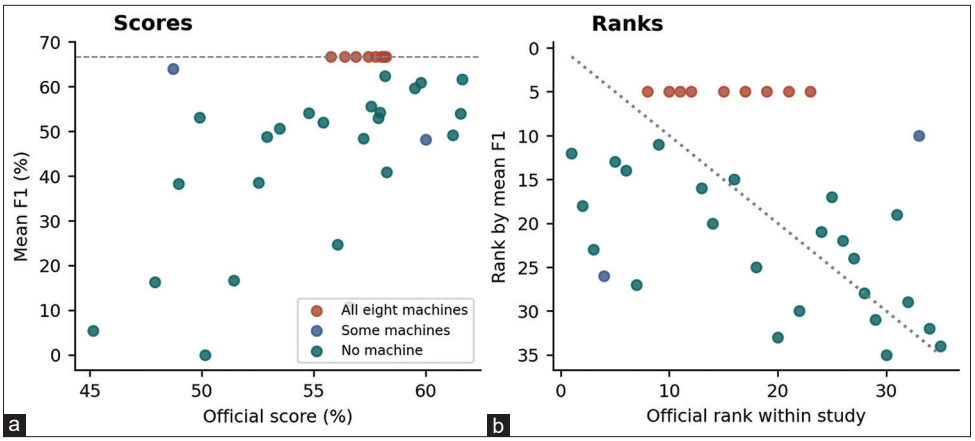


Table 4: Machine-specific reference comparisons and rank correlations

Machine	References	Zero	F1 above references	Pass both	ρ (AUC, F1)	ρ (pAUC, F1)
AutoTrash	9	4	11	11	0.69	0.72
BandSealer	9	2	1	3	0.42	0.42
CoffeeGrinder	10	4	0	0	0.11	−0.13
HomeCamera	10	1	1	1	0.12	0.22
Polisher	10	2	0	1	0.40	0.35
ScrewFeeder	11	3	6	5	0.40	0.36
ToyPet	10	4	1	0	0.37	0.37
ToyRCCar	11	4	0	0	0.02	0.23

Counts are out of 35 teams per row. "Ref." is the displayed 50/100/66.67 precision/recall/F1 pattern. Joint floors are 60%. Correlations use midranks

Figure 3: Median scores by machine type. Each cell is the median percentage across 35 systems. AUC and pAUC describe discrimination; precision, recall and F1 use the submitted decisions

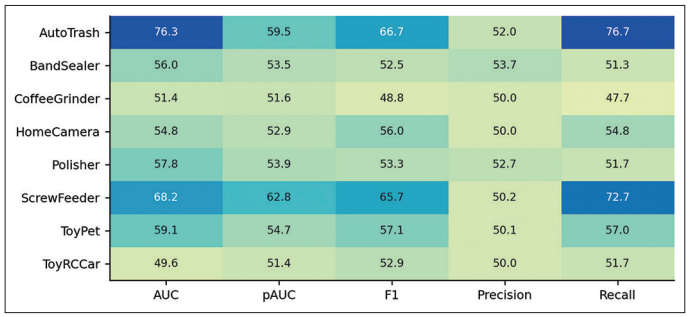
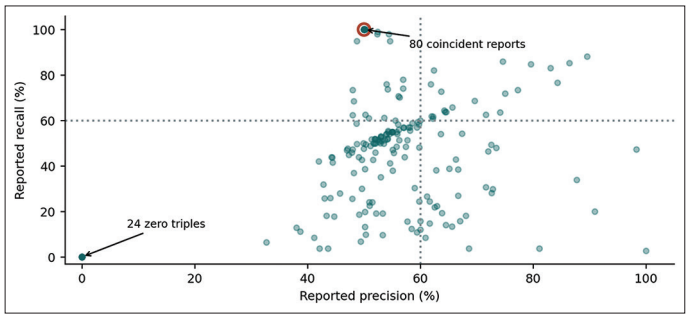


Figure 4: Submitted precision and recall on eight machine types. The 280 team–machine results include repeated coordinates. The dashed lines mark 60% precision and recall. The upper-right region meets both conditions.



AutoTrash supplied 11 qualifying pairs and ScrewFeeder supplied five, whereas CoffeeGrinder, ToyPet and ToyRCCar contributed no qualifying pair even though each of those machines was tested against the same group of 35 teams that had submitted results for the other machines (Table 6).

Holding the recall floor at 60% and raising the precision floor from 50% to 55% reduces the qualifying count from 117 to 26. Of the 91 lost results, 80 had the reference pattern and the remaining eleven fell outside the 50.00% precision tie. Because so many teams land on the same coordinate in the precision–recall plane, Figure 5 shows a sharp change in coverage even when the two neighbouring precision floors differ by just 5% points and the underlying set of decisions is identical.

Requiring 80% on both measures leaves three results, each from a different team and a single machine. As with the 60% screen, these floors illustrate a requirement someone could specify before selecting a detector. They do not establish a universal standard. A claim of adequate performance on eight machine types requires eight qualifying decisions regardless of how high a team's average score is.

4.5. Sensitivity to Team and Machine Inclusion

Removing the nine teams with the reference pattern on all machines changed the official-score versus mean-F1 correlation from 0.425 to 0.476. Excluding any team with that pattern left 24 and yielded 0.679. Removing teams with any zero triple had the opposite effect: 28 teams remained and the correlation fell to

Figure 5: Coverage under joint precision and recall requirements. Each cell counts qualifying team–machine pairs out of 280 at the indicated floors. Results are recomputed from published operating points without retuning any detector

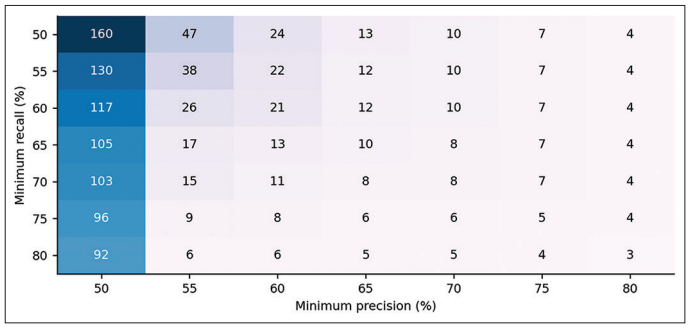


Table 5: Machine decisions of the highest official-scoring representative

Machine	AUC	pAUC	Precision	Recall	F1
AutoTrash	80.61	77.05	75.04	71.89	73.43
BandSealer	64.22	51.63	59.52	59.40	59.46
CoffeeGrinder	57.94	52.16	52.54	51.47	52.00
HomeCamera	62.45	53.79	60.01	59.93	59.97
Polisher	68.76	54.05	62.22	61.74	61.98
ScrewFeeder	90.24	79.16	83.05	82.99	83.02
ToyPet	62.64	54.05	59.73	58.33	59.02
ToyRCCar	44.35	52.84	44.36	43.64	44.00

All entries are percentages for Wang_MYPS_task2_3. The lowest and highest F1 occur on ToyRCCar and ScrewFeeder, respectively

Table 6: Number of machines and teams meeting paired floors

Recall floor	Precision floor	Cells	Percentage of 280	Teams	Max. machines
50	50	160	57.14	31	8
50	60	24	8.57	15	4
55	55	38	13.57	15	6
60	50	117	41.79	26	8
60	55	26	9.29	13	3
60	60	21	7.50	13	3
60	70	10	3.57	7	2
65	65	10	3.57	8	2
70	70	8	2.86	7	2
75	75	5	1.79	5	1
80	80	3	1.07	3	1

Floors use percentage units. Teams have at least one qualifying machine. The final column is the largest number of qualifying machines attained by one team

0.234. The highest coefficient thus describes the subgroup obtained by excluding every team with the reference pattern, whereas an engineer choosing from the entire field would have to contend with all the reported decisions, including those that produced the ties.

When one machine was omitted from mean F1, its correlation with the published official score ranged from 0.395 to 0.456. That score still includes all eight machines. To remove the mismatch in machine coverage between the two sides of the comparison, we calculated both mean pAUC and mean F1 on the same seven retained machines and obtained correlations ranging from 0.324 to 0.483. Depending on the omission, the official winner placed between tenth and fourteenth on seven-machine mean F1 (Figure 6).

The exclusions in Table 7 answer a sensitivity question. Selecting the group with the highest correlation after seeing the results would hide information that a prospective purchaser needs, including the number of teams whose published decisions tell the inspector almost nothing about which recordings can be set aside. The relevant screen is whether a proposed detector passed on the machine the plant intends to monitor.

4.6. Four Thresholds on a Chainsaw Inspection Test

Table 8 holds the chainsaw recordings and the detector fixed. The two lower thresholds detected all nine anomalies and produced zero false positives among 120 normal recordings. At 2.189, seven anomalies still triggered alarms. At 2.918, only five did. These are changes in the decisions of one detector as its cutoff moves (Figure 7).

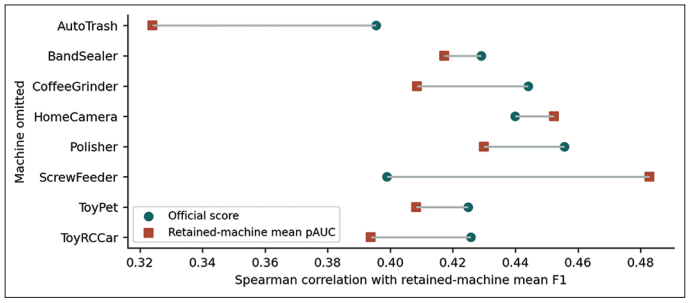
All four rules had the same observed specificity. Recall differed by as much as 44.44% points. Raising a threshold on a fixed score can remove alarms but cannot create new ones, so the two faults missed at the third threshold are necessarily included among the four missed at the fourth when the same detector scores the same nine faulty clips. Neither increase spared an observed false alarm.

Table 7: Rank associations by submitted decision pattern

Cohort	Teams	Official ρ	Official τ-b	pAUC ρ
All	35	0.425	0.303	0.411
Exclude all-signature	26	0.476	0.378	0.528
Exclude any signature	24	0.679	0.522	0.703
Exclude any zero	28	0.234	0.172	0.133
No signature or zero	18	0.558	0.438	0.472

The outcome is team mean F1. “Signature” denotes the reference pattern; zero denotes the displayed zero triple. Categories overlap and are separate sensitivity analyses

Figure 6: Rank concordance after omitting one machine. The circle retains the official eight-machine score; the square uses mean pAUC over the retained seven. Each comparator is correlated with mean F1 on those seven machines



Nine detections out of nine give 100% recall. Under the binomial model its exact two-sided 95% interval extends from 66.37% to 100%. With five detections the interval is 21.20-86.30%. Although separate training data supplied each of the four cutoffs, all four decisions on a particular faulty recording are based on the same underlying sound, so the experiment still has nine independent fault opportunities at most and cannot yield thirty-six.

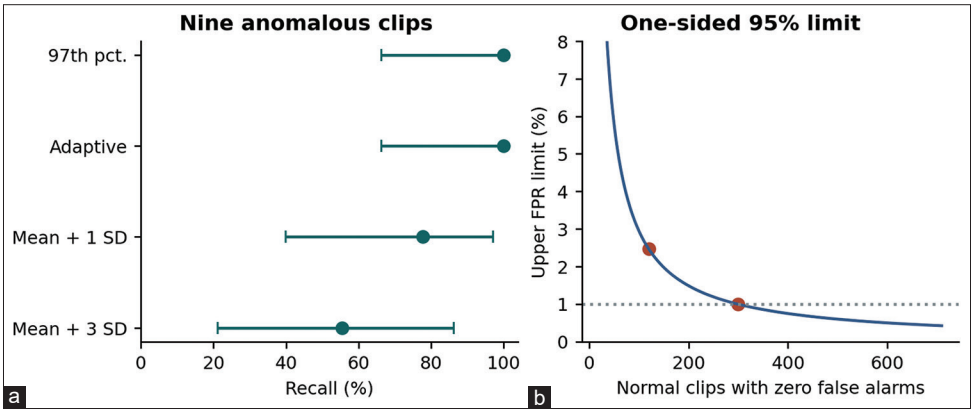
4.7. What Zero Observed False Alarms can Establish

In equation (6) the denominator is the 120 normal clips, although the full test contains 129. With zero false alarms the one-sided 95% upper limit is 2.466%. The nine anomalous clips provide no opportunity for a false positive and therefore cannot be added to that denominator, whereas obtaining a limit of 1% at the same confidence level would require at least 299 independent normal trials with zero false alarms.

At 1% prevalence and 100% recall, a service would correctly flag 10 of every thousand clips in expectation. If the false-positive rate were 0.5%, another 4.95 clips would generate false alarms and precision would be 66.89%. Keeping the same service prevalence and recall while using the candidate false-positive rate of 2.466% instead raises expected false alarms to 24.41 and lowers precision to 29.06%, with both quantities referring to the same hypothetical workload of a thousand clips (Figure 8).

At 0.1% prevalence the same candidate false-positive rate of 2.466% yields 24.63 expected false alarms against one true alarm per thousand clips. Precision falls to 3.90%. Table 9 holds recall fixed to show the effect of the uncertainty in normal recordings. Its projected precision is therefore conditional on recall, rather than a joint confidence limit for field performance.

Figure 7: (a and b) Uncertainty in the chainsaw inspection test. Panel A gives exact two-sided 95% intervals for recall. Panel B shows the one-sided 95% upper false-positive limit when all normal trials have zero alarms. Dots mark 120 and 299 independent trials



5. DISCUSSION

5.1. Choosing a Detector from a Shortlist

Ranking a review queue and deciding which recordings to discard are distinct tasks. A team can submit a score that sorts the queue usefully while assigning almost every clip the same alarm decision. DCASE exposes the difference because both outputs were recorded before official ranks were published. Had the organisers published only a table of AUC values, an operator could still compare systems on how well they sort a review queue but would have no way to identify which entrants had effectively sent every recording for inspection.

Suppose an engineer forms a shortlist in official-score order, as in Table 10. Its highest mean F1 can only increase or stay fixed as further teams enter. Once the first team with the reference pattern enters the shortlist, the maximum reaches 66.67% and ceases to distinguish the remaining additions, whereas counting the machines that meet both requirements can still identify a reason for keeping another team.

The official winner met the 60% precision and recall requirements on three machines. Evidence for a fourth machine must come from that machine. Retaining both percentages next to its identifier also distinguishes a failed precision requirement, which bears on unnecessary inspections, from a failed recall requirement, which bears on faults that pass unnoticed. An F1 mean merges those consequences.

Timbral measurements (Ota and Unoki, 2023), machine-dependent feature selection (Zhang et al., 2025) and attention to local spectral detail (Zhou et al., 2026) each suggest a model worth testing on a particular machine. None identifies the cause of a rank reversal among these 35 representatives. To discover whether a reversal could be attributed to an architectural choice, an evaluator would have to retain the individual clip scores and set each model's threshold by a common procedure on recordings independent of the test set.

5.2. Planning a Validation Run

The two lowest chainsaw thresholds occupy different positions on the score scale but yield the same observed decision counts: each detected all nine faults and generated zero false positives among 120 normal recordings. The test therefore supplies no empirical reason to prefer one of those thresholds over the other on those counts. What matters next is the distribution of new healthy recordings close to each cutoff.

Table 11 gives a planning calculation for a validation run that returns zero false alarms. At a target false-positive rate of 0.5%, 120 normal trials fall well short of what is needed for a one-sided 95% limit. Independence matters. Ten clips from adjacent moments in the same operating run may give a plant less information about the false alarms it will encounter in service than ten clips collected across the tools, production runs and noise conditions it expects to monitor.

For a follow-up study, the plant could fix the cutoff before recording new runs, retain tool and run identifiers, and record what the technician finds after each alarm. The resulting sequence would distinguish repeated clip alarms from separate inspections. With

Figure 8: Precision under hypothetical service prevalences. Curves use recall of 100%. The dashed line is the 2.466% upper false-positive limit from 120 independent normal trials. The horizontal axis is logarithmic

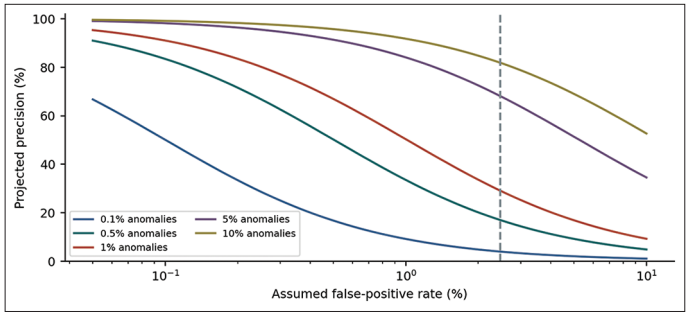


Table 8: Reconstructed chainsaw inspection decisions

Rule	Cutoff	TP	FP	TN	FN	F1	Recall
97th percentile	1.549	9	0	120	0	100.00	100.00
Adaptive	1.703	9	0	120	0	100.00	100.00
Mean+1 SD	2.189	7	0	120	2	87.50	77.78
Mean+3 SD	2.918	5	0	120	4	71.43	55.56

Source counts: Ashourpour and Azizpour (2026), Table 2. TP, FP, TN and FN are counts. F1 and recall are percentages. SD denotes standard deviation

Table 9: Projected precision by anomaly prevalence

FPR (%)	0.1%	0.5%	1%	5%	10%	FP/1,000[1]
0.100	50.03	83.40	90.99	98.14	99.11	0.99
0.500	16.68	50.13	66.89	91.32	95.69	4.95
1.000	9.10	33.44	50.25	84.03	91.74	9.90
2.000	4.77	20.08	33.56	72.46	84.75	19.80
2.466	3.90	16.93	29.06	68.10	81.84	24.41
5.000	1.96	9.13	16.81	51.28	68.97	49.50

Column headings give assumed prevalence. Entries are precision (%) at recall of 100%. [1]False alarms per 1,000 clips at 1% prevalence. Rates are scenarios

Table 10: Alarm evidence available in official-score shortlists

Teams kept	Best F1	All 8 ref.	Passing cells	Teams passing	Best coverage
1	61.61	0	3	1	3
2	61.61	0	4	2	3
3	61.61	0	4	2	3
5	61.61	0	8	4	3
8	66.67	1	11	5	3
10	66.67	2	12	6	3
12	66.67	4	12	6	3
15	66.67	5	14	7	3
20	66.67	7	17	9	3
25	66.67	9	19	11	3
30	66.67	9	20	12	3
35	66.67	9	21	13	3

Each shortlist is ordered by official score within the 35-team panel. Passing requires at least 60% recall and precision. Best coverage is the number of passing machines for the best-covered team

Table 11: Normal clips required after zero observed false alarms

Target FPR (%)	90% limit	95% limit	99% limit	Extra versus 120[1]
10.00	22	29	44	0
5.00	45	59	90	0
2.50	91	119	182	0
2.00	114	149	228	29
1.00	230	299	459	179
0.50	460	598	919	478
0.25	920	1197	1840	1077
0.10	2302	2995	4603	2875

One-sided exact limits under independent trials. [1]Additional normal clips beyond 120 at 95% confidence, assuming the full sequence produces no false alarms

timestamps and run identifiers, the plant could also test whether an apparently high clip-level alarm count represents many separate inspections or one prolonged fault that triggers the same decision repeatedly, a distinction the published summaries cannot resolve.

Small fault samples are a recurring problem in intelligent prognostics (Li et al., 2024). Methods using pseudo-labels or a few actual anomalous training examples recognise that constraint explicitly (Kuroyanagi et al., 2025; Kuroyanagi et al., 2025). The chainsaw example makes clear why the validation sample needs two denominators. Even if more normal recordings become available and narrow the upper false-positive limit, they cannot tell the engineer how many faults would escape detection, because uncertainty about recall depends on the separate count of truly anomalous recordings.

5.3. Prevalence and the Limits of Transfer

A balanced benchmark gives each class equal representation. An inspection line with rare faults gives the normal class much more weight, so a threshold that produced acceptable benchmark precision could generate mostly false alarms in service. Rearranging equation (7) gives the maximum false-positive rate f compatible with a target precision p_0 once prevalence π and recall r have been specified. No official rank enters this calculation.

$$f_{\max} = \frac{\pi r (1 - p_0)}{(1 - \pi) p_0}. \tag{8}$$

In equation (8), with 1% prevalence and 80% recall, a supervisor who requires half the alarms to be genuine can allow a false-positive rate of at most 0.808%. If the same supervisor requires 80% precision, the permissible rate falls to 0.202%, leaving both false-positive allowances below the 2.466% upper limit supported by the 120 normal clips that the chainsaw study actually examined. This comparison does not establish a breach of either requirement. The sample leaves both unresolved (Table 12).

A change of prevalence alone leaves the conditional rates in equation (7) intact. Once a change in prevalence is accompanied by a change in background noise, the type of equipment or the severity of faults that enter the stream, both conditional rates in equation (7) may also change and must be measured again. Domain-shift autoencoders and prototype learning address transfer at the representation stage (Chen et al., 2025; Lin et al., 2026). Establishing an acceptable

Table 12: Maximum allowable false-positive rate (%)

Prevalence (%)	60/50	80/50	100/50	60/80	80/80	100/80
0.1	0.060	0.080	0.100	0.015	0.020	0.025
0.2	0.120	0.160	0.200	0.030	0.040	0.050
0.5	0.302	0.402	0.503	0.075	0.101	0.126
1.0	0.606	0.808	1.010	0.152	0.202	0.253
2.0	1.224	1.633	2.041	0.306	0.408	0.510
5.0	3.158	4.211	5.263	0.789	1.053	1.316
10.0	6.667	8.889	11.111	1.667	2.222	2.778

Headers report assumed recall/required precision, both in percent. Each entry solves equation (8); it is a maximum allowable rate rather than a measured rate

alarm rate on the new line still requires observing its normal recordings at the proposed cutoff.

To establish performance on a second production line, a validation sample would have to cover the operating conditions expected for healthy machines on that line and include fault labels assigned independently of the detector, with agreement checked between the technicians who supplied them. Neither empirical record here tells us whether an alarm led to a maintenance intervention.

5.4. Study Limitations

DCASE selected each team's representative using its official score. This prevents a team with several entries from receiving several votes in the analysis, but it also means the representative need not be the team's best binary classifier. Our estimates describe the 35 selected submissions. A team might have submitted another configuration that made better use of its threshold while attaining a lower official score, so the results cannot describe all configurations that entered the challenge or establish a ranking for commercial monitors in general.

Published summaries were sufficient to pair each team's score and decision metrics by machine. They were insufficient to identify the individual clips on which the models disagreed, the threshold-selection procedure, or the reason for a repeated decision pattern. To assess whether an alarm rule could be repaired, an evaluator would need the original clip scores and a calibration set separate from the test recordings. Choosing a new cutoff on those test labels would no longer assess the submitted rule.

Equal weighting assigns each of the eight machine types one vote. That answers a cross-machine benchmark question, but it may be a poor representation of a plant with many screw feeders and only one polisher. The machine-specific results allow this weighting choice to be inspected. Omitting one observed machine and recalculating the mean shows the contribution of that machine to the ordering of the 35 systems, but gives no observation of what those same systems would do on a ninth machine they have never encountered.

The chainsaw record describes four thresholds on one test set. It provides no independent test of the DCASE team correlations. Recall rests on nine anomalous clips, and recordings grouped within runs could make the binomial intervals too narrow. The projections vary prevalence alone. If the sound of healthy equipment changes between the inspection study and the

production line, or if the line encounters milder faults than the test set contained, the engineer must estimate recall and the false-positive rate in the new population.

6. CONCLUSION

The DCASE winner placed twelfth on mean alarm F1. Nine teams shared first place on that metric with decisions consistent with flagging every clip. Once the requirement was stated separately for precision and recall, only 21 of the 280 team–machine results reached 60% on both. An operator deciding whether to use a submission on one particular machine can therefore obtain a more relevant answer from the precision and recall reported for that machine than from either the official rank or the mean F1 across machines.

In the chainsaw test, raising the cutoff reduced detections from nine to five without sparing the inspector a single observed false alarm, a distinction that was visible only because the two thresholds were compared against the same set of recordings. Observing zero false alarms in 120 normal clips still permits a one-sided 95% upper false-positive rate of 2.47% under independent sampling. For an acoustic monitor, a defensible acceptance claim therefore names the machine, fixes the threshold, specifies the required alarm performance and gives the independent normal and faulty counts behind it.

REFERENCES

Ashourpour, M., Azizpour, G. (2026), Dual-stream autoencoder architecture for audio-based anomaly detection in industrial quality control. IEEE Access, 14, 77879-77905.

Chen, S., Sun, Y., Wang, J., Wan, M., Liu, M., Li, X. (2025), A multi-scale dual-decoder autoencoder model for domain-shift machine sound anomaly detection. Digital Signal Processing, 156, 104813.

Di Fiore, E., Ferraro, A., Galli, A., Moscato, V., Sperlì, G. (2022), An anomalous sound detection methodology for predictive maintenance. Expert Systems with Applications, 209, 118324.

Gaudreault, J.G., Branco, P. (2024), Empirical analysis of performance assessment for imbalanced classification. Machine Learning, 113, 5533-5575.

Kuroyanagi, I., Fujimura, T., Takeda, K., Toda, T. (2025), Improving anomalous sound detection through pseudo-anomalous set selection and pseudo-label utilization under unlabeled conditions. APSIPA Transactions on Signal and Information Processing, 14(1), e13.

Kuroyanagi, I., Hayashi, T., Takeda, K., Toda, T. (2025), Serial-OE: Anomalous sound detection based on serial method with outlier exposure capable of using small amounts of anomalous data for training. APSIPA Transactions on Signal and Information Processing, 14(1), e1.

Li, C., Li, S., Feng, Y., Gryllias, K., Gu, F., Pecht, M. (2024), Small data challenges for intelligent prognostics and health management: A review. Artificial Intelligence Review, 57, 214.

Lin, Y., Li, Y., Wu, W., Sun, S., Shi, F., Shao, J. (2026), An improved prototype learning method for domain generalized industrial machine anomaly sound detection. IEEE Access, 14, 43694-43708.

Ota, Y., Unoki, M. (2023), Anomalous sound detection for industrial machines using acoustical features related to timbral metrics. IEEE Access, 11, 70884-70897.

Qurthobi, A., Maskeliūnas, R., Damaševičius, R. (2022), Detection of mechanical failures in industrial machines using overlapping acoustic anomalies: A systematic literature review. Sensors, 22(10), 3888.

Sørbø, S., Ruocco, M. (2024), Navigating the metric maze: A taxonomy of evaluation metrics for anomaly detection in time series. Data Mining and Knowledge Discovery, 38, 1027-1068.

Wilkinghoff, K., Kurth, F. (2024), Why do angular margin losses work well for semi-supervised anomalous sound detection? IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 608-622.

Wu, Y., Ye, X., Fu, Q., Zhang, W. (2026), Machine anomalous sound detection method based on lightweight temporal pyramid and ECA-MobileFaceNet. Sensors, 26(10), 3214.

Zhang, Y., Feng, J., Zhang, Q., Hu, J., Zhang, W. (2025), An unsupervised anomalous sound detection method based on similarity-driven automatic feature selection. Digital Signal Processing, 161, 105123.

Zhou, H., Zhou, Y., Liu, Y., Liu, H. (2026), MSANet: Multi-stage attention network for anomalous sound detection in machine condition monitoring. Digital Signal Processing, 168, 105626.