Università di Genova
DIBRIS DIPARTIMENTO
DI INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

Learning Sounds For Machine Monitoring

Saba Belal

Master Thesis

Università di Genova, DIBRIS Via Opera Pia, 13 16145 Genova, Italy
https://www.dibris.unige.it/

Università
di Genova

MSc Computer Science
Data Science and Engineering Curriculum

Learning Sounds For Machine Monitoring

Saba Belal

Advisor: Stefano Rovetta

Examiner: Francesca Odone

July, 2026

Abstract

Acoustic condition monitoring allows industrial machines to be watched for incipient faults from sound alone, and the DCASE Challenge Task 2 has made unsupervised anomalous sound detection, training on normal operation only, its standard benchmark. The 2026 edition adds a second microphone placed farther from each machine, on the premise that the far channel supplies a standing noise reference that should improve robustness to environmental noise. This thesis tests that premise on the seven development machine types, using two reproduced detectors as fixed measuring instruments: the official autoencoder baseline (System A) and a training-free frozen-embedding nearest-neighbour detector in the GenRep line (System B), which reaches parity with the baseline on the development set.

The study addresses two questions. First, does the second channel improve detection? Not by any of the fusion mechanisms evaluated here: uniform late fusion of the near and far channels lowers the official harmonic-mean score on both detectors, with paired-bootstrap intervals lying entirely below zero, the loss is not specific to the equal-weight blend, and no score-level or embedding-level variant rises above the near channel alone. Second, does the outcome replicate across detector families? It does: two detectors that share no architecture, scoring rule, or training procedure converge on the same negative, which makes it a property of the second channel and of the encoders that consume it rather than of one detector, and identifies cross-detector agreement as the check that a single-detector stereo claim lacks. A two-level account is offered for the result: the near/far premise holds for only two of the seven machines, and, on a proposed architectural explanation consistent with the failure of embedding-level fusion, the frozen monaural encoders project both channels into much the same monaurally learned representation and cannot preserve the inter-channel structure, which would explain why fusion fails even where a real channel gap exists. The contribution is a per-machine characterization of stereo anomalous sound detection that locates the failure in the data and the encoders rather than in the detectors.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Industrial machinery fails, and the cost of a failure depends almost entirely on when it is detected. Corrective maintenance waits for a breakdown and pays for unplanned downtime; preventive maintenance replaces components on a fixed schedule and discards remaining service life. Predictive maintenance (PdM) sits between these extremes: it monitors the actual condition of a machine and intervenes only when degradation is detected, performing each repair at the latest safe moment [YPY25]. The enabling component of any PdM system is a condition-monitoring pipeline that can register the early signs of a fault before it progresses to failure.

Sound is an attractive monitoring channel for this purpose. Rotating and reciprocating machines emit characteristic acoustic signatures, and incipient faults such as bearing wear, gear damage, imbalance, or cavitation alter those signatures, often before any change is measurable elsewhere [AUN+19]. Microphones are non-invasive, inexpensive, and capture the superposition of many fault modes from a single sensor, which makes acoustic condition monitoring practical to deploy at scale [PTI+19]. Because anomalous machine sounds are rare, diverse, and infeasible to collect exhaustively in advance, the problem is posed as unsupervised anomalous sound detection (ASD): a model is trained only on normal operation and must flag departures from that learned normality at test time, without ever having seen a fault [Nun21, KKI+20].

What turns this from an accuracy problem into a deployment problem is the cost of a false alarm. A monitoring system that interrupts production for inspections that find nothing healthy quickly loses the trust of maintenance teams and is switched off [WFI+25]. A useful detector must therefore separate anomalies from normal operation at a conservative decision threshold, where the false-positive rate is kept low. This requirement

is encoded directly in the evaluation used throughout this thesis, that of the DCASE Challenge. The Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge is an annual set of public benchmark tasks in machine listening; its Task 2 has been the reference benchmark for unsupervised anomalous sound detection since 2020, releasing a common dataset, protocol, and scoring each year on which participating systems are compared [KKI+20, NHT+26]. Alongside the global area under the ROC curve (AUC), DCASE Task 2 scores the partial AUC (pAUC) restricted to the low false-positive-rate region $[0, 0.1]$, and the official ranking metric is the harmonic mean of AUC and pAUC over all machine types and sections [KKI+20, NHT+26]. The pAUC term is what carries the false-alarm cost into the score, and it is the operative quantity throughout this thesis: a system that ranks anomalies well on average but cannot hold up under a strict threshold is of little industrial value. For reference, the official autoencoder baseline reaches a mean pAUC of only 0.545 [NHT+26], which this thesis reproduces at 0.547 (Table 5.1); either figure signals how far the low-false-alarm regime remains from solved. Section 2.1 develops the maintenance and acoustic background in full; this chapter states only what is needed to frame the contribution.

1.2 The 2026 Noise-Aware Task

The DCASE 2026 Challenge Task 2, "Noise-aware unsupervised anomalous sound detection for machine condition monitoring," introduces one structural change from every prior edition of the task [NHT+26]. Where previous editions provided single-channel audio, the 2026 training and test clips are two-channel recordings captured simultaneously by two microphones placed at different distances from the machine: one near the target and one farther away. All recordings are 16 kHz and last 6 to 16 seconds, and participants may use either or both channels.

The design rests on an explicit premise. The organizers state that the distant microphone "is expected to contain relatively stronger environmental noise and weaker direct machine sounds," so it "may help distinguish environmental noise components from the target machine sounds" [NHT+26]. In other words, the second channel is offered as a noise reference: the near microphone carries the signal, the far microphone carries comparatively more noise, and the contrast between them should make the detector more resilient to the environmental noise that has limited ASD performance in noisy conditions. This is an appealing assumption, and it is the assumption this thesis sets out to test rather than to take for granted. As the results will show, the near/far contrast is real for only a minority of machine types: the signal-to-noise ratio gap between the channels exceeds 3 dB for the two ToyCar variants (3.05 dB and 3.10 dB) but stays below 1.6 dB for the remaining five machine types, and is even slightly negative for the bearing machine, where the far microphone is marginally the cleaner of the two. The premise that motivates the entire

task therefore holds only partially, and characterizing exactly where it holds and where it breaks is a central concern of what follows.

1.3 Research Questions

The two detectors compared in this thesis must first exist and be trusted. Chapter 4 establishes that foundation: it builds both System A (the official autoencoder baseline) and System B (a GenRep-style frozen-embedding nearest-neighbour detector that pairs pretrained audio encoders with kNN scoring [SS25]) on the 2026 development set, together with the dual-channel fusion operator and the evaluation protocol common to both, providing two independent and trustworthy detector families. The thesis is organized around three research questions, the first of which is settled as a foundation rather than left open.

- **RQ1 (Can both detectors be reproduced and trusted?).** Can the official autoencoder baseline and a frozen-embedding nearest-neighbour detector both be reproduced on the 2026 development set, with System B reaching parity with the baseline, so that the two serve as reliable reference detectors? Chapter 4 answers this affirmatively and treats it as the foundation for the two questions that follow rather than as an open problem.
- **RQ2 (Does the second channel help?).** Does adding the second microphone improve detection over the near channel alone, and is the underlying near/far SNR premise empirically supported across the seven machine types?
- **RQ3 (Does the outcome replicate across detectors?).** Does the effect of adding the second channel replicate across detector families, so that two methodologically unrelated detectors agree on whether it helps or hurts, or is the outcome specific to a single detector?

1.4 Contributions

This thesis builds two competitive detectors for the 2026 task and uses them to characterize, per machine, how the stereo near/far microphone pair affects unsupervised ASD. Centrally, it reproduces a frozen-embedding nearest-neighbour detector in the design family that leads recent first-shot ASD [SS25] and shows that this detector matches or surpasses the official autoencoder baseline on the development set. On that footing, the work establishes whether the second channel contributes, finds that uniform near/far fusion does not improve detection on either detector, and explains the mechanism that governs that behaviour.

- **Two reproduced detectors, one of state-of-the-art design.** We reproduce both the official autoencoder baseline (System A) and a frozen-embedding kNN detector (System B) of the kind that leads recent first-shot ASD [SS25]. On the development set, System B matches or surpasses the official baseline, giving two reliable detector families to build the stereo study on.
- **A per-machine study of stereo ASD on the 2026 development set.** We evaluate whether the near/far microphone pair improves unsupervised ASD across both detector families, using per-machine analysis rather than aggregate scores alone.
- **A measurement of the near/far premise.** We quantify the per-channel signal-to-noise ratio for every machine type and show that the noise-reference premise holds for only 2 of the 7 development machines (the two ToyCar variants), with the other five showing channel gaps under 1.6 dB.
- **A cross-detector test of channel fusion.** We show that uniform near/far fusion does not improve detection on either detector family: both the autoencoder and the frozen-embedding detector register a statistically significant aggregate loss under channel fusion, and their per-machine outcomes largely agree rather than diverge. Two methodologically unrelated detectors converging on the same negative makes the result a property of the channel and the encoders rather than of one detector, and identifies cross-detector agreement as a necessary check for any stereo ASD claim.
- **A mechanistic hypothesis.** We propose, as an explanation consistent with the failure of embedding-level fusion, that frozen monaural pre-trained encoders project both channels into much the same monaurally learned subspace and do not preserve the spatial structure the second microphone provides. This would account for why uniform fusion fails to help, and on some machines significantly hurts, even where a genuine near/far contrast exists, and it points to channel-aware encoders as the route to exploiting the second microphone.

1.5 Thesis Outline

The remainder of the thesis proceeds as follows. Chapter 2 establishes the technical and industrial background: acoustic condition monitoring within predictive maintenance, the signal-processing representations used to turn audio into time-frequency features, the machine-learning methods that distinguish normal from anomalous operation, and the DCASE Task 2 challenge lineage that situates this work. Chapter 3 describes the 2026 development set, characterizes its two-channel structure, and quantifies the per-channel signal-to-noise ratio for every machine type to test the near/far premise directly. Chapter 4 specifies the two detector families, the autoencoder baseline (System A) and the

frozen-embedding kNN detector (System B), the dual-channel late-fusion operator, and the evaluation protocol held identical across systems, establishing the two trustworthy detector families on which the research questions are tested. Chapter 5 reports the full per-machine ablations and the cross-detector comparison, addressing both whether the second microphone improves detection over the near channel and whether any benefit replicates across detector families (RQ2 and RQ3). Chapter 6 concludes with the limitations of the study and directions for future work.

Chapter 2

Background

This chapter establishes the technical and industrial foundation for the anomalous sound detection (ASD) pipeline investigated in this thesis. It is organised into four parts. The first situates acoustic monitoring within the broader predictive maintenance context, explaining why machine sounds carry diagnostic information and how ASD fits within the condition-monitoring pipeline. The second covers the signal-processing representations used to transform raw audio into time-frequency features. The third surveys the machine-learning architectures that operate on those features to distinguish normal from anomalous operation. The fourth traces the DCASE Task 2 challenge series from its single-channel origins to the 2026 noise-aware setting and locates the present work within that lineage. Together, these sections motivate the features and models evaluated in the experimental chapters.

2.1 Predictive Maintenance and Acoustic Monitoring

2.1.1 The Maintenance Spectrum: From Scheduled to Predictive

Industrial machinery requires maintenance to remain safe and productive, and the strategy chosen for that maintenance has direct consequences for both operating cost and system reliability. Three broad strategies exist along a spectrum of increasing sophistication. *Corrective maintenance*, commonly called "run to failure", defers all intervention until a breakdown has already occurred. While it minimises planned downtime, unplanned failures are expensive, difficult to schedule, and in safety-critical settings potentially dangerous. *Preventive maintenance* addresses this by scheduling interventions at fixed time intervals or usage thresholds, regardless of actual machine condition. This avoids catastrophic failures but routinely replaces components that still have significant remaining service life, wasting

both materials and labour.

Predictive maintenance (PdM) represents the current state of the art in industrial practice. Rather than following a fixed schedule, PdM continuously monitors the actual condition of a machine and triggers maintenance only when indicators suggest that a failure is approaching. The objective is to perform each intervention at the latest safe moment: late enough to avoid unnecessary replacement, early enough to prevent an unplanned breakdown. When implemented effectively, PdM reduces maintenance costs, increases asset availability, and generates operational data that can improve future maintenance decisions [YPY25].

The core requirement of any PdM system is therefore a reliable *condition monitoring* pipeline, a set of sensors and algorithms that can detect the early signs of degradation before they progress to failure. The choice of monitoring modality matters greatly, and it is this choice that motivates the acoustic approach taken in this thesis.

2.1.2 Why Sound? The Case for Acoustic Condition Monitoring

Rotating and reciprocating machinery produces sound as an unavoidable byproduct of its operation. Bearings, gears, fans, pumps, and valves all generate characteristic acoustic signatures determined by their geometry, speed, and load. When a component begins to degrade (through wear, imbalance, misalignment, contamination, or structural damage), the acoustic signature changes in ways that are often detectable well before the machine fails or performance degrades measurably [AUN+19]. Acoustic condition monitoring exploits this relationship to infer machine health from sound [YPY25].

Several properties make acoustic monitoring particularly attractive in industrial contexts [YPY25]. First, it is *non-invasive*: microphones can be placed near a machine without any physical contact or modification to the equipment, unlike accelerometers that must be mounted directly to the machine housing. Second, it is *low-cost*: consumer-grade microphones suitable for condition monitoring are inexpensive and widely available, making large-scale deployment feasible. Third, it is *information-rich*: a single microphone captures the superposition of all acoustic events produced by a machine, meaning that multiple fault modes can in principle be detected from one sensor [PTI+19]. Finally, acoustic signals are *familiar* to human operators, who often detect anomalies by ear long before formal instrumentation confirms them, confirming that the acoustic channel carries genuine diagnostic information.

These advantages are not without trade-offs. Unlike vibration sensors mounted at the bearing housing, microphones capture all sound in the environment, including background noise from adjacent machines, ventilation systems, and human activity. Separating the target machine's contribution from this acoustic clutter is a central challenge in industrial

ASD, and one that strongly influences the design of the machine learning systems described in later sections of this chapter.

An older response to this challenge, and the one the 2026 task revives, is to use more than one microphone. When a reference sensor captures predominantly noise, the noise component in a primary signal can be estimated and subtracted, an idea formalised as adaptive noise cancelling [WGM+75] and, in single-channel form, as spectral subtraction [Bol79] and Wiener filtering [LO79]. With several sensors, spatial filtering, or beamforming, steers sensitivity toward the source and away from interference [VVB88], and relative-transfer-function methods exploit the fixed acoustic path between microphones to separate a target from spatially distinct noise [GBW01]. The 2026 near/far microphone pair studied in this thesis is a minimal instance of this classical setup, a signal-bearing channel alongside a comparatively noisier reference, so the question of whether the second channel helps is one that this literature has long addressed in speech and array processing.

2.1.3 Fault Mechanisms and Their Acoustic Signatures

Different physical degradation mechanisms produce acoustically distinct signatures, and understanding this mapping is important for choosing appropriate signal representations and detection methods. The seven machine types in the DCASE 2026 development set span several of these regimes; four of them illustrate the range, namely the bearing, the fan, the valve, and the ToyCar.

Bearing faults are among the most acoustically distinctive failure modes. As a bearing race, rolling element, or cage degrades, each rotation produces a series of impulsive contacts at characteristic frequencies set by the bearing geometry and rotational speed. These appear as periodic, narrow-band energy bursts, often most pronounced in high-frequency bands (above 1 kHz) where background noise is lower. Early-stage faults may produce only subtle amplitude modulation of the noise floor, while advanced faults generate clearly audible impacts [AUN+19]. The same impulsive, high-frequency character governs the gear faults of the gearbox machine, where broken or chipped teeth excite the gear-mesh frequency and its harmonics and produce sidebands around them [AUN+19].

Rotating aerodynamic machinery, represented by the fan, emits a continuous and broadly stationary sound: tonal components at the rotational frequency and its harmonics superimposed on broadband aerodynamic noise [PTI+19]. Imbalance, misalignment, or contamination shift or strengthen the tonal components, but these changes are low in frequency and easily masked by the aerodynamic floor, which makes them harder to separate from benign speed variation than a sharp bearing impact.

Intermittent and transient sources, represented by the valve and the slide rail, produce sound mainly during actuation. A solenoid valve emits short impulsive clicks as it opens

and closes, so its normal signature is a sparse train of transients rather than a continuous tone, and faults alter the shape, timing, or energy of those transients [PTI+19]. The slide rail adds sliding and friction noise from linear motion. Because the informative events are brief and separated by near-silence, these machines stress detectors that assume a stationary signal.

Composite miniature machinery, represented by the ToyCar, combines several of the above in a controlled package: a small motor, a gear train, and running gear whose normal sound mixes motor whine, gear noise, and rolling contact [KSU+19a]. The ToyCar and its emulated counterpart act as repeatable stand-ins for full-scale rotating machinery, with faults introduced into the drive train.

This diversity implies that no single acoustic feature or detection algorithm is universally optimal. A representation tuned to high-frequency impulsive content suits the bearing and gearbox but is less discriminative for the low-frequency tonal shifts of an unbalanced fan, and a model that assumes a stationary input is poorly matched to the sparse transients of the valve. This observation motivates the multi-representation and learned-feature approaches reviewed in the remainder of this chapter, and it reappears in the per-machine analysis of later chapters, where machine type strongly conditions which detector and representation perform best.

2.1.4 Anomalous Sound Detection as a Predictive Maintenance Task

Within the PdM pipeline, Anomalous Sound Detection (ASD) occupies the role of the *fault detection* stage: its output is a binary or continuous-valued anomaly score indicating whether the machine is operating normally or shows signs of deviation. This is distinct from *fault diagnosis*, which would identify the specific component or failure mode responsible, and from *prognosis*, which would estimate how long the machine can continue to operate safely (the Remaining Useful Life, or RUL).

ASD is formulated as an anomaly detection problem rather than a supervised classification problem for a fundamental practical reason: *anomalous operating sounds are rare, diverse, and difficult to collect in advance*. A factory cannot be expected to deliberately damage its equipment to generate training data for every possible failure mode. Consequently, ASD systems are trained exclusively on recordings of healthy machine operation and must identify departures from that learned normal state at test time, without ever having observed an example of failure [Nun21, KKI+20]. This one-class or unsupervised formulation is the defining characteristic of the problem and the primary driver of the architectural choices reviewed in Section 2.3.

The challenge is compounded by the fact that "normal" is not a fixed concept. The

same machine operating at different speeds, loads, or in different acoustic environments will produce different sounds, all of which are equally normal. An effective ASD system must therefore learn a model of normality that is *sensitive* to fault-induced changes while remaining *insensitive* to benign operational variations. Achieving this balance under realistic industrial conditions, including domain shifts between training and deployment environments, is the central open problem that motivates the DCASE benchmark series and the research reviewed in this thesis [WFI+25, KIK+21].

2.2 Feature Extraction

Audio signals are captured as one-dimensional time-domain waveforms sampled at fixed rates. They are the most basic representation and carry complete signal information, but they are sensitive to noise, amplitude variations, and recording conditions. While some recent deep learning approaches process raw waveforms directly using 1D convolutional networks, most methods transform the signal into alternative representations that better expose relevant acoustic structure. This section covers the spectrogram-based time-frequency representations used in this thesis; the learnable representations produced by pre-trained neural encoders are treated separately in Section 2.3. Methods rooted in classical signal processing rely on well-established mathematical transformations with fixed parameters, making them interpretable, computationally efficient, and suitable for scenarios with limited training data.

2.2.1 Digital Audio and Sampling

A microphone converts sound pressure into a continuous voltage, which an analog-to-digital converter records as a sequence of numbers by measuring it at regular time intervals. The number of measurements per second is the sampling rate f_s, expressed in hertz. The choice of f_s fixes the highest frequency the recording can represent: the Nyquist-Shannon sampling theorem states that a signal sampled at f_s can faithfully capture frequency content only up to the Nyquist frequency $f_s/2$, and any energy above that limit is aliased, appearing as a spurious lower frequency that cannot afterwards be removed. Sampling rate therefore trades fidelity against data volume. Consumer music is typically sampled at 44.1 kHz, which covers the roughly 20 kHz upper limit of human hearing, whereas speech and many machine-sound applications use lower rates because the informative content lies in a narrower band. The DCASE Task 2 recordings, including the 2026 development set used in this thesis, are sampled at 16 kHz [NHT+26], so all frequency content up to 8 kHz is retained. At this rate a single 10-second clip is a vector of 160,000 samples.

That number is the reason raw waveforms are rarely fed to a detector directly. Neigh-

bouring samples of an audio signal are strongly correlated, and a periodic machine sound repeats near-identical structure many times within a clip, so the 160,000-dimensional vector carries far less independent information than its length suggests: it is highly redundant. The representation is also unstable in ways that do not matter perceptually or diagnostically. Two recordings of the same sound that differ only by a small time shift, or by the phase of a periodic component, produce very different sample sequences even though they are acoustically identical, which forces a model working on raw samples to learn invariances that a suitable transform supplies for free. For these reasons most ASD systems first map the waveform to a time-frequency representation that compresses the redundancy and exposes the spectral structure in which faults are visible, as described next.

2.2.2 Time-Frequency Analysis

Time-frequency analysis is a powerful signal processing approach used to characterize non-stationary signals, those whose frequency content changes over time, such as audio from mechanical systems. Unlike purely time-domain or frequency-domain methods, time-frequency analysis provides a joint representation that captures how the spectral components of a signal evolve, enabling a more detailed understanding of transient events and periodic structures [YPY25]. Techniques such as the Short-Time Fourier Transform (STFT) and spectrograms are commonly employed to generate these representations. In the context of machine sound monitoring and anomaly detection, time-frequency analysis is particularly valuable, as faults often manifest as subtle or abrupt changes in spectral patterns over time [AUN+19]. These representations not only serve as rich sources of information for manual inspection but also form the input for machine learning models, making them a foundational element of modern audio feature extraction pipelines.

2.2.3 Short-Time Fourier Transform (STFT)

The Short-Time Fourier Transform is a technique used to determine how the frequency content of a signal changes over time by dividing a longer signal into shorter segments and computing the Fourier transform separately on each. This approach addresses the limitation of the conventional Fourier Transform, which provides no temporal localisation and is therefore unsuitable for non-stationary signals.

The discrete STFT of a signal $x[n]$ is mathematically defined as:

$$X(m,k) = \sum_{n=0}^{N-1} x[n + mH]\, w[n]\, \exp\left(-\frac{2\pi i k n}{N}\right) \tag{2.1}$$

where n is a sample index within the analysis window, $w[n]$ is the window function, m is the frame index, k is the frequency bin index, N is the window length, and H is the hop size (the number of samples by which the window is shifted between successive frames).

The STFT involves a fundamental trade-off between time and frequency resolution: short windows provide higher time resolution for capturing rapid changes but result in lower frequency resolution, while longer windows offer better frequency resolution at the cost of reduced temporal precision. This trade-off, governed by the Heisenberg uncertainty principle, is an important consideration when designing sound analysis systems for different applications.

2.2.4 Window Functions and Overlap

The choice of window function strongly affects the quality of time-frequency analysis. Commonly used window functions include Hann, Hamming, and Gaussian windows, which taper at the edges to reduce spectral leakage. The window length typically ranges from 20 to 50 milliseconds for speech and audio applications, though this may be adjusted based on the specific characteristics of the signal being analysed. Overlapping windows reduces information loss at window boundaries and produces smoother, more continuous time-frequency representations; typical overlap ratios range from 50% to 75%. However, increased overlap raises computational cost because more frames must be processed.

An audio signal carries information in both the time and frequency domains, so a joint representation is necessary. Spectrogram-based representations are energy-based time-frequency mappings derived from the magnitude or power of the STFT, with optional perceptual or task-specific post-processing. Spectrograms are the dominant feature for audio analysis because they present sound as a two-dimensional image, enabling the use of computer vision techniques such as CNNs [TCP22]. The most widely used spectrogram variants are described below.

2.2.5 Mel-Frequency Scaling

The Mel scale is a perceptually motivated frequency scaling based on how humans perceive pitch differences. While it is foundational in speech processing, its application in machine sound analysis is more contested. The rationale is that frequency bands are warped to mimic human auditory resolution, denser at low frequencies and sparser at high frequencies. This scaling is mathematically defined as:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.2}$$

However, industrial machines are not designed for or interpreted by human hearing. Recent work on learnable audio front-ends has shown that data-driven representations can outperform fixed perceptual scales such as the Mel filterbank on audio classification tasks [ZTdCQT21]. These findings suggest that while Mel scaling is computationally convenient and well established, its effectiveness for detecting subtle machine faults may be limited when compared to data-driven alternatives.

2.2.6 Log-Mel Spectrograms

Log-Mel spectrograms are two-dimensional time-frequency representations derived by applying a Mel filter bank to the power spectrogram and then taking the logarithm. They are widely used in deep learning-based audio systems. Typical configurations involve 64, 128, or 256 Mel bands, with frame lengths between 20 and 50 ms and 50% overlap.

The popularity of log-Mel spectrograms stems from their compatibility with image-based convolutional neural networks. They provide a perceptually smooth yet information-rich input, and their two-dimensional structure aligns well with models pre-trained on natural image datasets such as ImageNet. They also offer a practical balance between resolution and computational load, making them a de facto standard in audio-based anomaly detection pipelines [KKI+20, KSU+19a].

2.3 Machine Learning

Building upon the feature extraction techniques discussed above, machine learning (ML) frameworks provide the computational logic to automatically model normal and abnormal acoustic behaviour in industrial systems. While signal processing transforms raw waveforms into structured time-frequency representations, machine learning acts as the interpretative layer that translates those features into predictive insights.

The methods in this chapter rest on a distinction worth stating plainly, because it separates ASD from the more familiar supervised setting. In *classification*, a model is trained on labelled examples drawn from every class it must later recognise, and it learns a decision boundary that assigns each new input to one of those known classes; predicting a fault this way requires having observed faults during training. *Anomaly detection*, and its close variant *novelty detection*, remove that requirement: the training set contains only the normal class, the model builds a description of normal behaviour, and any input that departs from it is flagged. The two settings differ in what they consume and what they emit. A classifier consumes labelled examples of every class and emits a class label, whereas an anomaly detector consumes only normal data and emits a continuous anomaly score,

which a threshold then converts into a normal-or-anomalous decision. The *novelty-detection* qualifier is the precise one for ASD as posed here: the training data are assumed to be entirely normal, uncontaminated by unlabelled faults, so the task is to recognise departures from a clean model of normality rather than to separate two populations observed side by side.

It follows from this distinction that, in the specific context of Anomalous Sound Detection (ASD) for predictive maintenance, the objective is not classification but the identification of deviations from learned patterns of healthy operation. This is important in industrial environments where fault data are scarce, highly imbalanced, or entirely unavailable during training. Consequently, contemporary approaches prioritise learning compact representations of normal machine sounds, identifying anomalies as outliers within the learned feature space. The evolution from handcrafted statistical models to data-driven learning frameworks has improved robustness against background noise, variable operating conditions, and cross-machine generalisation gaps [Nun21]. The rapid advancement of these methodologies has been heavily driven by the annual DCASE challenges [KKI+20, KIK+21, DIH+22], which have established the primary public datasets and benchmarks for industrial machine condition monitoring and are traced in full in Section 2.4.

2.3.1 Deep Reconstruction-Based Architectures

Reconstruction-based methods operate on the assumption that a neural network trained exclusively on normal operational sounds will learn to represent the range of sounds that normal operation produces (their data manifold, the region of the signal space that normal clips occupy). When presented with an anomalous input that falls outside this learned range, the model reconstructs it poorly, and this reconstruction error, the difference between the input and the network's attempted copy of it, serves as an anomaly score [KKI+20].

2.3.1.1 Autoencoders

An Autoencoder (AE) is a neural network trained to copy its input to its output through a narrow intermediate layer. That narrow layer, the *bottleneck*, forces the network to squeeze the input into a compact, low-dimensional code, its *latent space*, and then rebuild the input from that code, so it reconstructs well only the patterns it saw often during training. The reconstruction error then serves as the anomaly score. While effective for stationary signals, simple AEs struggle with non-stationary machine sounds, such as those produced by valves or slider rails. In these cases, the inherent variance of normal sound may exceed the variance introduced by an anomaly, leading to false negatives. The official DCASE Task 2 baseline, and System A in this thesis, is an autoencoder of exactly this

kind [NHT+26]. Reconstruction-based ASD in this vein was established early in the field, both as a formal detection objective derived from the Neyman–Pearson lemma [KSU+19b] and through architectural variants such as the interpolation network, which predicts a masked centre frame rather than reconstructing the whole input, sharpening sensitivity to non-stationary events [SNP+20].

2.3.2 Large-Scale Pre-trained Models and Transfer Learning

The "first-shot" paradigm, in which a system must generalise to an entirely new machine type with no machine-specific tuning, has made *transfer learning*, reusing a model already trained on a large general dataset instead of training one from scratch, a central strategy in state-of-the-art ASD [DIH+23, NHN+24]. When training data for a new machine type is extremely limited, the most effective strategy available is to reuse an *audio encoder*, a neural network that maps a sound clip to a fixed-length vector of numbers (its *embedding*) summarising the clip's content, pre-trained on a large general-purpose sound dataset such as AudioSet [HJZ+25].

2.3.2.1 Transformer-Based Backbones: AST and BEATs

The Audio Spectrogram Transformer (AST) applies the vision transformer (ViT) framework, a neural architecture built from *self-attention* layers that let every part of the input exchange information with every other part, to acoustic data by treating the spectrogram as an image: it is split into overlapping patches that are processed through self-attention, allowing the model to relate distant points in time and frequency [GCG21].

The BEATs (Bidirectional Encoder representation from Audio Transformers) framework uses an iterative pre-training approach that jointly learns an acoustic tokeniser and an audio self-supervised model. BEATs has performed well in noisy environments, especially when fine-tuned using Blind Source Separation (BSS) as an auxiliary task [XTL25].

2.3.2.2 Self-Supervised and Distilled Encoders

These encoders differ mainly in how they are pre-trained, and three families recur below. In *self-supervised* learning the model creates its own training targets from unlabelled audio, for instance by hiding part of a spectrogram and learning to predict it, a *masked* objective. In *contrastive* learning the model is trained to place related inputs close together and unrelated ones far apart in the embedding space. In *distillation* a smaller model is trained to imitate the outputs of a larger model or an ensemble. Beyond AST and BEATs, the detector developed in this thesis draws on four further pre-trained encoders, selected

so that the set spans distinct pre-training objectives rather than several variants of one. EAT (Efficient Audio Transformer) is a self-supervised model pre-trained on AudioSet with a masked-spectrogram objective in a bootstrapped self-distillation framework, combining an utterance-level and a frame-level loss under a high masking ratio so that pre-training is computationally inexpensive; the large variant is used here [CLM+24]. M2D-CLAP couples Masked Modeling Duo, a masked-prediction scheme in which an online encoder predicts the latent representations that a momentum target encoder assigns to masked patches, with CLAP-style audio-language contrastive learning, so its embedding space is organized both by reconstruction and by natural-language supervision [NTO+24]. SSLAM (Self-Supervised Learning from Audio Mixtures) extends masked self-supervision to overlapping mixtures of sounds rather than isolated clips, targeting the polyphonic soundscapes typical of real recordings, which suits industrial audio where the target machine is heard against background noise [AAM+25]. CED (Consistent Ensemble Distillation), by contrast, is a supervised audio-tagging model distilled from an ensemble of teachers under consistent augmentation; the tiny variant used here is the smallest and cheapest encoder in the set, included as a discriminatively trained counterpoint to the self-supervised models [DWY+24].

The set therefore mixes masked self-distillation (BEATs, EAT), audio-language contrastive learning (M2D-CLAP), mixture-based self-supervision (SSLAM), and supervised ensemble distillation (CED), with embedding widths ranging from 192 to 1024 dimensions. The intent is that the frozen embeddings carry complementary inductive biases, so that combining their anomaly scores draws on partly independent views of the same clip rather than correlated ones.

2.3.2.3 Patch-Based Spatial-Temporal Representations

For few-shot, training-free ASD, researchers have proposed methods that extract patch-based representations directly from frozen pre-trained AST models [WLK+25]. Rather than relying on a single global feature vector, the input spectrogram is divided into localised patches and processed through the transformer. The resulting latent representations are compared against a *memory bank*, a stored collection of embeddings taken from normal clips, using a similarity measure such as *cosine similarity* (the cosine of the angle between two vectors, which is 1 when they point the same way and near 0 when they are unrelated). Scoring a clip by its closeness to the most similar stored normal examples is a *nearest-neighbour* formulation whose roots lie in distance-based outlier detection [RRS00] and which directly parallels the memory-bank approach that leads industrial visual anomaly detection [RPZ+22]. The anomaly score is derived by aggregating these similarity values, effectively producing an anomaly map that highlights localised deviations in the soundscape.

2.3.2.4 Frozen-Embedding Nearest-Neighbour Detection

A complementary use of pre-trained encoders avoids fine-tuning altogether. A frozen encoder maps each recording to a fixed embedding, the embeddings of the normal training clips are stored in a memory bank, and a test clip is scored by its distance to the nearest stored embeddings: a clip that lies far from every normal example in the embedding space is flagged as anomalous [SS25]. Because no parameters are updated, the detector operates first-shot on a new machine type, and this design is the basis for System B in this thesis. System B builds on five pre-trained audio encoders used as frozen feature extractors: BEATs, EAT-large, M2D-CLAP, SSLAM, and CED-tiny, each pre-trained on large-scale audio with a self-supervised or, in the case of CED, a supervised-distillation objective.

The main difficulty is the domain imbalance built into the task. Each machine provides roughly 990 source-domain clips but only about 10 target-domain clips [NHT+26], so a raw nearest-neighbour score systematically reports the sparsely populated target region as anomalous. GenRep, the system reproduced as System B, addresses this with two devices: MemMixup, which augments the small target memory bank using its nearest source neighbours, and a domain normalization step that places the source and target distances on a comparable scale so the two regions are treated even-handedly [SS25]. The result is a training-free detector that, as reported in Section 2.4, ranked second in the 2025 challenge while using no labelled anomalies.

2.3.3 Domain Shift and Domain Generalization

A difficulty cuts across all of the architectures above: the conditions under which a machine is recorded change between training and deployment, and a detector tuned to one set of conditions can misjudge another. The DCASE task formalizes this as a split between a source domain, recorded under one set of conditions with abundant normal data, and a target domain, recorded under different conditions (a changed operating speed or load, or a different background) with only a handful of normal clips. Two formulations follow. In *domain adaptation* the target domain is known and the model may be specialized to it; in *domain generalization*, the setting the task has used since 2022, the domain label is withheld at test time, so a single detector and a single decision threshold must serve both domains at once [KIK+21, DIH+22].

Domain generalization is the harder and more realistic setting, and it is the central difficulty of the present work. With roughly 990 source-normal clips against about 10 target-normal clips per machine [NHT+26], a detector that learns normality from the abundant source data alone will systematically misjudge the sparsely sampled target domain, flagging normal target clips as anomalous merely because they are under-represented. Methods in this family attack the imbalance in different ways: by learning representations invariant to

the factors that distinguish the domains, by standardizing or normalizing anomaly scores per domain so that the two are placed on a comparable scale, or, as in the frozen-embedding detector reproduced as System B, by maintaining separate per-domain references and combining them. Because the cost of mishandling this is concentrated in the target domain, the experimental chapters report the two domains separately rather than pooling them, and treat target-domain metrics as the higher-variance of the two.

2.3.4 Evaluation Metrics and Industrial Performance Requirements

In the industrial context, an ASD system's utility is measured not only by overall accuracy but also by its reliability at low false-positive rates (FPR). Maintenance teams cannot afford frequent false alarms, as these lead to alarm fatigue and wasted operational costs [WFI+25].

2.3.4.1 AUC, pAUC, and the Official Score

A detector does not output a label but a real-valued anomaly score, and a decision requires a *threshold*: clips scoring above it are declared anomalous, those below it normal. Each threshold produces two rates on a labelled test set, the *true-positive rate* (TPR), the fraction of genuine anomalies correctly flagged, and the *false-positive rate* (FPR), the fraction of normal clips wrongly flagged. Sweeping the threshold from strict to permissive traces the *receiver operating characteristic* (ROC) curve, which plots TPR against FPR. The Area Under this Curve (AUC) condenses the curve into a single threshold-independent number: it equals the probability that a randomly chosen anomaly receives a higher score than a randomly chosen normal clip, so it is 0.5 for random guessing and 1.0 for a detector that ranks every anomaly above every normal clip [KKI+20]. AUC thus measures ranking quality across all operating points at once. However, the partial AUC (pAUC) [McC89], calculated over the FPR range $[0, 0.1]$, is the more industrially relevant metric. A high pAUC indicates that the model can identify anomalies even when the decision threshold is set extremely conservatively to minimise false alarms, which is the operative requirement in continuous machine monitoring [WFI+25].

These per-domain and per-section measurements are condensed into a single official score Ω, defined as their harmonic mean over all machine types m, sections n, and domains $d \in \{\text{source}, \text{target}\}$ [NHT+26]:

$$\Omega = h\big\{\mathrm{AUC}_{m,n,d},\ \mathrm{pAUC}_{m,n}\big\}\,,$$

where $h\{\cdot\}$ denotes the harmonic mean. The harmonic mean is used rather than the arithmetic mean because it is dominated by its smallest arguments: a system cannot offset

a machine type or domain it handles poorly by scoring highly on the rest. Every machine-domain cell must be reasonable for Ω to be high. This property is the reason per-machine analysis, not the aggregate score alone, is used throughout this thesis, since a single weak machine can hold Ω down while the average looks healthy.

2.4 DCASE Task 2 Lineage and Position of This Work

The methods reviewed above have been shaped, year by year, by the DCASE Challenge Task 2 on unsupervised ASD, which has set the field's public benchmarks on the MIMII [PTI+19] and ToyADMOS [KSU+19a] machine-sound datasets. Each edition changed one element of the setting, and the progression moves steadily toward the noisier, lower-supervision conditions of real deployment:

- **2020.** The first edition fixed the core protocol: train on normal sound only, and score detectors by the AUC together with the partial AUC over the low false-positive region $[0, 0.1]$ [KKI+20].
- **2021.** Domain shift was introduced, splitting each machine's data into a source recording condition and a sparsely sampled target condition, with the domain of each test clip known [KIK+21].
- **2022.** This was tightened into domain generalization, withholding the domain label at test time so that a single detector and a single threshold must serve both conditions [DIH+22].
- **2023.** The challenge adopted the first-shot setting, requiring systems to generalize to entirely new machine types with no machine-specific tuning [DIH+23, NHN+24].
- **2024.** Each machine type was consolidated into a single section [NHT+26].
- **2025.** Supplemental recordings, either clean target sound or noise-only clips, were supplied to support noise robustness in scenarios where the machine or the noise source can be stopped long enough to capture them [NHN+25].
- **2026.** That assumption is replaced with a two-channel recording from a near and a far microphone, on the premise that the far microphone supplies a standing noise reference even when no separate noise recording is available [NHT+26].

The most recent completed edition, DCASE 2025, shows where the state of the art now sits and which method family leads. Ranked by the official score (the harmonic mean of AUC and pAUC over all machines on the evaluation set), the top two systems

sit within a hair of each other: a fine-tuned audio-encoder system with a discriminative classification head won at 61.63%, and the training-free GenRep frozen-embedding detector ranked second at 61.57%, against an autoencoder baseline of 56.51% [DCA25, SS25]. That a training-free nearest-neighbour detector comes within 0.06 points of a fine-tuned winner, with none of the latter's training cost, is the empirical basis for this thesis taking the frozen-embedding line as System B rather than a fine-tuning pipeline.

This thesis works entirely within the 2026 noise-aware setting, on the seven development machine types. It does not pursue a leaderboard ranking, and the strongest recent systems depend on encoder fine-tuning pipelines. Instead, it takes the two detector families that bracket current practice, a reconstruction autoencoder (System A), which is the official baseline [NHT+26], and a training-free frozen-embedding nearest-neighbour detector (System B) in the GenRep line that has led recent first-shot editions [SS25], and uses them as fixed, reproduced reference points. On that footing, the work addresses the question the 2026 design leaves open: whether the far microphone actually delivers the noise robustness its premise promises, and whether either detector family can convert the second channel into improved detection. The remainder of the thesis answers that question per machine and per detector.

Chapter 3

Dataset and Exploratory Analysis

The 2026 task rests on a single empirical premise: that the far microphone supplies a usable noise reference because it captures comparatively more environmental noise and less direct machine sound than the near microphone [NHT+26]. Whether that premise holds is not a modelling choice but a property of the recordings, and it can be measured before any detector is built. This chapter establishes the data on which the rest of the thesis runs and then tests the premise directly. The first section describes the 2026 development set and its two-channel structure. The second quantifies the per-channel signal-to-noise ratio (SNR) for every machine type and reports how large the near/far contrast actually is. The third reads those measurements as a statement about channel heterogeneity and draws out what they imply for the two-channel strategies developed in Chapter 4.

3.1 The DCASE 2026 Task 2 Development Set

The development set contains the normal and anomalous operating sounds of seven machine types: `ToyCar`, `ToyCarEmu`, `fan`, `bearingEmu`, `gearboxEmu`, `sliderEmu`, and `valveEmu` [NHT+26]. Five of the seven carry the `Emu` suffix, which in this dataset marks an emulated machine whose recordings are simulated rather than captured from a physical unit operating in a real acoustic environment; the two without the suffix, `ToyCar` and `fan`, are real recordings [NHT+26]. The distinction matters only insofar as two real and five emulated machine types coexist in the same benchmark and are treated identically by the split below. These machines span the impulsive, tonal, and transient fault regimes described in Section 2.1.3, and each machine type is provided as a single section, in keeping with the first-shot protocol adopted from 2024 onward [NHT+26].

The structural change that defines the 2026 edition is the move from single-channel

Table 3.1: Composition of the DCASE 2026 Task 2 development set. Each machine type forms a single section with a source (Src) and a target (Tgt) domain. Durations are representative per machine; all clips are two-channel and sampled at 16 kHz.

			Train (normal)		Test (normal)		Test (anomalous)		
Machine	Type	Dur. (s)	Src	Tgt	Src	Tgt	Src	Tgt	Total
`ToyCar`	real	11	990	10	50	50	50	50	1200
`ToyCarEmu`	emu	12	990	10	50	50	50	50	1200
`bearingEmu`	emu	10	990	10	50	50	50	50	1200
`fan`	real	10	990	10	50	50	50	50	1200
`gearboxEmu`	emu	10	990	10	50	50	50	50	1200
`sliderEmu`	emu	10	990	10	50	50	50	50	1200
`valveEmu`	emu	10	990	10	50	50	50	50	1200
Total			6930	70	350	350	350	350	8400

to two-channel audio. Where every prior edition supplied one microphone, each 2026 clip contains two synchronised channels recorded at different distances from the machine, one microphone near the target and one farther away, and participants may use either or both [NHT+26]. All recordings are two-channel, last between 6 and 16 seconds, and are sampled at 16 kHz [NHT+26]. The microphone arrangement differs across machine types but is held fixed within each type, so the near/far geometry is a per-machine constant rather than a per-clip variable. For the five emulated machines the two channels are produced by convolving recorded impulse responses with sounds captured near the machine, while for the two real machines, `ToyCar` and `fan`, the two channels are captured directly with synchronised microphones [NHT+26]. Throughout this thesis, channel 0 (ch0) denotes the near microphone and channel 1 (ch1) the far microphone. As the measurements in Section 3.2 show, ch0 has the higher signal-to-noise ratio on every machine except `bearingEmu`, where the two channels are effectively tied and the far microphone is marginally the cleaner of the two.

Table 3.1 reports the per-machine composition. The structure is uniform across machines: 990 source-domain and 10 target-domain normal clips for training, and a test set of 50 normal and 50 anomalous clips in each of the two domains, for 1200 clips per machine and 8400 in total.

The task retains the domain-shift structure of the 2022 to 2025 editions: each section splits into a source domain with sufficient training data and a target domain with only a handful of clips, and the domain label is withheld at test time so that one detector and one threshold must serve both [NHT+26]. The most consequential fact about this composition is the scarcity of target-domain training data. Each machine provides only 10 target normal

clips against 990 source clips, yet the test set evaluates the two domains in equal measure. A detector must therefore characterise target-domain normality from roughly ten examples, which makes target-domain metrics in later chapters high-variance: a target AUC that moves by a few points between configurations is more likely to reflect sampling noise than a real difference in detector quality. The experimental chapters consequently treat small target-domain differences as noise and report per-machine results with confidence intervals rather than point estimates alone.

3.2 Per-Channel Signal-to-Noise Ratio Analysis

The premise of the two-channel design is quantitative, so it admits a quantitative test. If the far microphone is to act as a noise reference, the near channel must carry a measurably stronger machine signal relative to noise than the far channel, and the size of that difference determines how much spatial information the second channel can add. This section measures the per-channel SNR for each machine type and reports the near/far gap.

3.2.1 Estimation Procedure

The signal-to-noise ratio is estimated per channel directly from each clip, without a separate reference noise recording. Each channel is divided into frames of 1024 samples with a hop of 512 samples, which at the 16 kHz sampling rate corresponds to frames of 64 ms advanced by 32 ms, and the power of each frame is its mean squared amplitude. Writing P_{90} and P_{10} for the 90th and 10th percentiles of the per-frame power across a clip, the SNR is

$$\mathrm{SNR} = 10 \log_{10}\left(\frac{P_{90}}{P_{10}}\right) \quad [\mathrm{dB}].$$

This is a percentile estimate of the within-clip dynamic range: P_{90} stands in for the high-energy frames dominated by the machine and P_{10} for the quiet frames that approximate the noise floor. It is a proxy for the true signal-to-noise ratio, not a decomposition of target sound and background obtained from a reference noise track, and should be read as such. The per-machine figures are the mean of this per-clip estimate over the source-normal clips, with the standard deviation taken across those clips.

3.2.2 Per-Channel SNR Across Machine Types

Table 3.2 reports the mean SNR of each channel, the near-minus-far gap (ch0 minus ch1), and the standard deviation of that gap across clips. The near/far contrast the task assumes

Table 3.2: Per-channel SNR (percentile estimate, mean over source-normal clips), the near/far gap (ch0 − ch1), and the across-clip standard deviation of the gap, ordered by gap. Channel 0 is the near microphone, channel 1 the far one.

Machine	SNR ch0 (dB)	SNR ch1 (dB)	Gap (dB)	Gap SD (dB)
`ToyCarEmu`	7.61	4.51	3.10	1.50
`ToyCar`	7.64	4.59	3.05	1.42
`valveEmu`	5.95	4.38	1.57	0.41
`gearboxEmu`	6.25	4.90	1.35	0.69
`fan`	4.08	3.78	0.30	0.19
`sliderEmu`	5.79	5.70	0.08	0.47
`bearingEmu`	4.54	4.65	−0.11	0.62

is present for only a minority of the development machines. The ToyCar family shows a large and consistent gap of about 3 dB (`ToyCarEmu` 3.10 dB, `ToyCar` 3.05 dB), the regime in which the far channel behaves as the intended noise reference. `valveEmu` (1.57 ± 0.41 dB) and `gearboxEmu` (1.35 ± 0.69 dB) show smaller but clearly positive gaps. At the other end, `fan` (0.30 ± 0.19 dB), `sliderEmu` (0.08 ± 0.47 dB), and `bearingEmu` (-0.11 ± 0.62 dB) have gaps that straddle zero, so their two channels carry essentially the same signal-to-noise ratio, and for `bearingEmu` the far microphone is marginally the cleaner of the two. The premise that motivates the entire 2026 task therefore holds clearly for two of seven machines and weakly or not at all for the rest.

3.3 Channel Heterogeneity and Its Implications

The measurements in Section 3.2 describe a heterogeneous set of machines rather than a uniform two-channel benefit. The near/far SNR gap ranges from about 3 dB on the ToyCar family down to a value near zero, and slightly negative, on `bearingEmu`. The same split appears in the inter-channel level and correlation: the `ToyCar` near channel is roughly 14 dB louder than its far channel with near-zero inter-channel correlation, whereas the `fan` channels are close to balanced and strongly correlated (about 0.87). This spread is a direct consequence of the recording design noted in Section 3.1: the microphone geometry is fixed per machine but differs across machines [NHT+26], so the degree to which the far microphone is genuinely the noisier of the two is itself a per-machine property. There is no single near/far relationship that holds across the development set.

The level difference tells only part of the story; the correlation between the two channels tells the rest. Table 3.3 reports, for every machine, the mean level of each channel, their difference, and the inter-channel correlation, and four regimes emerge. On the two ToyCar

Table 3.3: Inter-channel level and correlation per machine, averaged over forty source-normal clips. Levels are mean RMS in dBFS; the level difference is ch0 − ch1, so a positive value means the near microphone is louder. The correlation is the Pearson correlation between the two channel waveforms. The machines group into four regimes.

Machine	ch0 (dBFS)	ch1 (dBFS)	Level diff. (dB)	Inter-chan. corr.	Regime
`ToyCar`	−24.9	−39.6	+14.8	0.03	weak, decorrelated far
`ToyCarEmu`	−25.2	−27.7	+2.7	0.04	weak, decorrelated far
`bearingEmu`	−38.3	−38.6	+0.4	0.30	balanced, partly indep.
`sliderEmu`	−24.6	−23.9	−0.7	0.47	balanced, partly indep.
`fan`	−24.3	−21.4	−2.8	0.87	near-mono (far ≈ near)
`gearboxEmu`	−26.3	−26.5	+0.3	−0.11	balanced, anti-correlated
`valveEmu`	−42.7	−42.9	+0.3	−0.50	balanced, anti-correlated

variants the far channel is both quieter and essentially uncorrelated with the near channel (correlation ≈ 0.03), so it carries a different, weak noise field rather than a copy of the signal, a contrast visible in the near and far spectrograms of Figure 3.1. On `bearingEmu` and `sliderEmu` the levels are within a decibel and the channels are moderately correlated (0.30 to 0.47), partly independent but similar. On `fan` the two channels are near-duplicates, correlated at 0.87 with the far channel marginally the louder, effectively one microphone recorded twice. On `gearboxEmu` and `valveEmu` the levels match but the correlation is negative (−0.11 and −0.50), a spatially opposed capture. The scalar SNR gap of Section 3.2 does not capture this structure: a small gap can accompany either a near-duplicate channel (`fan`) or an anti-correlated one (`valveEmu`), which are very different situations for a fusion scheme. What independent information the second channel can add is therefore a question of correlation as much as of level, and the consequence for detection is taken up in Chapter 5.

This heterogeneity has a direct consequence for how the second channel can be used and evaluated. A two-channel strategy premised on a clean noise reference can only help where such a reference exists, which the SNR analysis localises to the two ToyCar variants and, more weakly, to `valveEmu` and `gearboxEmu`. On the machines whose channels differ by only a few tenths of a decibel, the two microphones carry nearly the same signal-to-noise content, so the second channel offers little independent information for a noise-referencing scheme

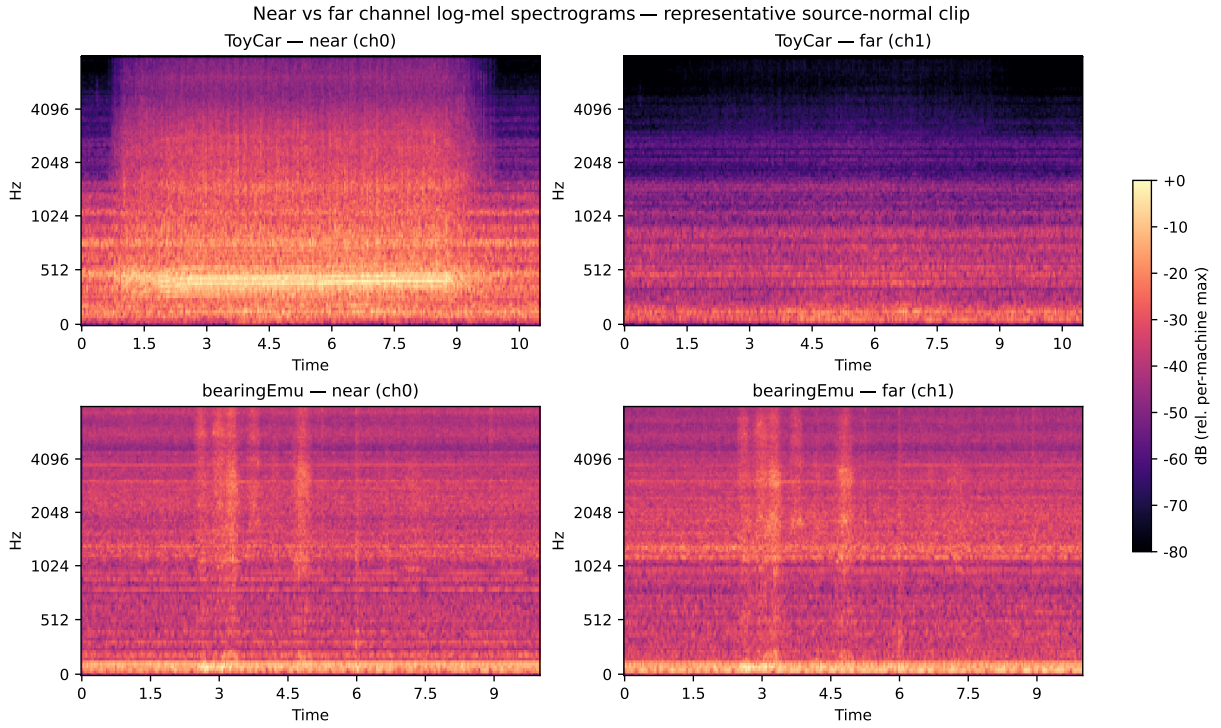

Figure 3.1: Near (ch0) versus far (ch1) log-mel spectrograms of a representative source-normal clip, for `ToyCar` (top) and `bearingEmu` (bottom). Colour is decibels relative to the per-machine maximum across both channels, so the level gap is preserved rather than normalized away per panel. The `ToyCar` far channel is visibly washed out, consistent with its large level gap and near-zero correlation, whereas the `bearingEmu` channels are similar in both level and structure.

to exploit, and on `bearingEmu` the nominal roles of the two microphones are effectively reversed. Any benefit that channel fusion produces must therefore be read per machine: an aggregate improvement averaged over all seven types can be driven entirely by the machines where the premise holds while concealing that it does nothing, or harms, on the others. This is the reason the experimental chapters report per-machine results rather than a single averaged score, and it frames the central question that the systems of Chapter 4 are built to answer, whether the second channel delivers a detection gain even on the machines where a real near/far contrast is present. Whether it does is a modelling result, not a property of the data, and is deferred to the experimental chapters that follow.

Chapter 4

Systems and Methods

This chapter specifies the two detectors and the measurement apparatus used to answer the research questions. It builds two systems that bracket current practice in unsupervised ASD, a reconstruction autoencoder and a frozen-embedding nearest-neighbour detector, and fixes a single evaluation protocol that both share, so that any difference observed later is a property of the detector or the channel and not of the way it was measured. The chapter is organised as follows. Section 4.1 states the experimental design and the comparison matrix. Section 4.2 describes System A, the official autoencoder baseline used unmodified. Section 4.3 describes System B, the frozen-embedding kNN detector, including the contamination-sensitive question of how its encoders and layers were selected. Section 4.4 defines the dual-channel late-fusion operator. Section 4.5 sets out the metrics, the per-machine reporting lens, and the bootstrap procedure used for the fusion verdict.

4.1 Experimental Design

The study compares two detector families rather than tuning one system to its limit. System A is a reconstruction autoencoder: it learns to reproduce normal log-mel frames and scores a clip by how badly it reconstructs them. System B is a frozen-embedding nearest-neighbour detector: it maps each clip to a fixed embedding from a pre-trained audio encoder and scores it by distance to the stored normal examples. The two represent the dominant lines in current unsupervised ASD, a trained reconstruction model and a training-free embedding model, and choosing one of each is deliberate: a property of the second microphone that holds for only one family is a property of that family, not of the channel, and only a system that appears in both families can be attributed to the data. Neither detector is tuned in this thesis. Both are reproduced and then frozen, and they are treated as fixed measuring instruments whose role is to register a delta between channels,

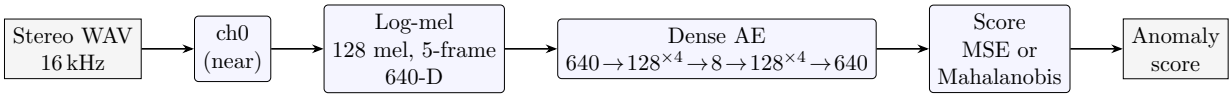

Figure 4.1: System A pipeline. The near channel is converted to a log-mel representation with five-frame context, reconstructed by a dense autoencoder, and scored by reconstruction error (MSE) or selective Mahalanobis distance. The far channel enters only through the late-fusion configuration of Section 4.4.

not to chase an absolute score.

What is held identical across the two systems is the part that makes the comparison valid. Both run on the same DCASE 2026 development set described in Chapter 3, with the same single section per machine, the same source and target domains, and the same test clips. Both are scored with the same metrics, AUC, partial AUC at a false-positive rate of 0.1, and the official harmonic-mean score Ω defined in Section 2.3. Both use the same decision protocol and the same per-machine reporting. Only the detector internals differ. The evaluation protocol that fixes these choices is given in Section 4.5.

On this common footing the experiment is a three-by-two comparison, evaluated per machine. Each detector is run in three input configurations, the near channel alone (ch0), the far channel alone (ch1), and the dual-channel late fusion of the two (Section 4.4), and the same three configurations are run on both System A and System B. The three configurations and two detectors map onto the questions directly. The ch0 configuration on each detector establishes the single-channel parity foundation, the trustworthy reference against which everything else is measured. The contrast between fusion and ch0 within a single detector answers RQ2, whether the second channel improves detection over the near channel alone. The contrast of that fusion-versus-ch0 outcome across the two detectors answers RQ3, whether any improvement is a reproducible property of the channel or an artefact of one detector family. The results of this matrix are reported per machine in Chapter 5; the present chapter defines the systems and the protocol that produce it.

4.2 System A: Reconstruction Autoencoder Baseline

System A is the official DCASE 2026 Task 2 autoencoder baseline [NHT+26], used without modification. It is included as a fixed, widely understood reference point rather than as a contribution of this thesis: keeping it byte-for-byte identical to the released baseline is what lets the second channel and the second detector be measured against a known quantity. Its configuration is reproduced here only so that the comparison is fully specified, and is summarized in Figure 4.1.

Front end. Each clip is converted to a log-mel spectrogram with 128 mel bins, computed from a short-time Fourier transform with a window of 1024 samples and a hop of 512 samples (Equation 2.1). Each input vector to the network is formed by concatenating a context of five consecutive frames, giving a $128 \times 5 = 640$-dimensional input. At the 16 kHz sampling rate of the dataset, the analysis window spans 64 ms and advances by 32 ms.

Architecture. The network is a dense (fully connected) autoencoder. The encoder maps the 640-dimensional input through four hidden layers of width 128 to a bottleneck of dimension 8, and the decoder mirrors this back to the 640-dimensional output, so the full stack is $640 \to 128^{\times 4} \to 8 \to 128^{\times 4} \to 640$. Every hidden layer applies batch normalization followed by a ReLU activation. The narrow bottleneck is the working part of the model: it forces the network to keep only the regularities shared across many normal frames and to discard idiosyncratic detail, so that frames unlike the training distribution are reconstructed poorly.

Training. A separate model is trained for each machine type on that machine's source-normal and target-normal clips. Training runs for 100 epochs with the Adam optimizer at a learning rate of 10^{-3}, a batch size of 256, and a mean-squared-error reconstruction loss. The random seed is fixed at 13711.

Scoring. Two scoring modes share the same trained weights. In the MSE mode the anomaly score of a clip is its mean reconstruction error. In the Selective Mahalanobis mode, after training, one further pass over the training data collects per-domain covariance statistics, and a clip is scored by the smaller of its Mahalanobis distances [Mah36] to the source-domain and target-domain normal models, $\min(\mathrm{Maha}_{\mathrm{src}}, \mathrm{Maha}_{\mathrm{tgt}})$. The decision threshold for both modes is set as the 0.9 quantile of a gamma distribution fitted to the training-set scores, and the partial AUC is computed with a maximum false-positive rate of 0.1. Both modes are reported, because they expose different failure modes of the same model.

Input. For the single-channel reference, System A consumes the near channel (ch0) only, matching the released baseline. The dual-channel configuration of Section 4.4 runs the identical model independently on ch0 and ch1 and fuses the two resulting score streams.

4.3 System B: Frozen-Embedding Nearest-Neighbour Detector

System B detects anomalies without training a model on the task. A pre-trained audio encoder, frozen, maps each clip to a fixed embedding; the embeddings of the normal training

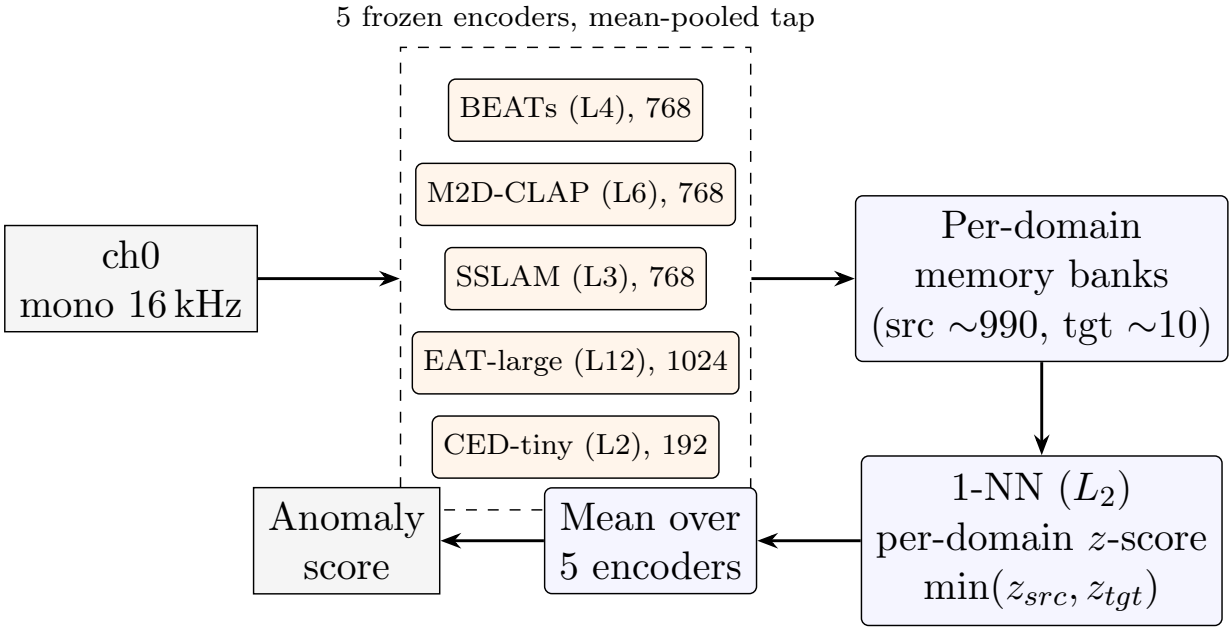

Figure 4.2: System B pipeline. The near channel is embedded by five frozen encoders, each tapped at a fixed transformer block and mean-pooled; a test clip is scored against per-domain memory banks by its nearest-neighbour Euclidean distance, standardized per domain and reduced by $\min(z_{src}, z_{tgt})$, and the five per-encoder scores are averaged. Layer taps and the selection procedure are given in Section 4.3.3.

clips are stored in a memory bank; and a test clip is scored by its distance to the nearest stored normal embedding. Because no parameters are updated, the detector operates first-shot on a new machine type, with none of the cost of a training run and none of the reproducibility risk that attaches to a fine-tuning schedule, where learning rate, warm-up, and stopping point each move the result. This is the design's claim to a place in the study on its own terms: a training-free nearest-neighbour detector in this family finished a close second to the fine-tuned winner of the most recent completed challenge, effectively matching the winning official score at none of the training cost (Section 2.4). System B reproduces the GenRep approach [SS25] in its single-channel, layer-tuned configuration, summarized in Figure 4.2.

4.3.1 Encoders and Embedding Extraction

System B uses five pre-trained audio encoders as frozen feature extractors: BEATs [CWW+23], EAT-large [CLM+24], M2D-CLAP [NTO+24], SSLAM [AAM+25], and CED-tiny [DWY+24]. Each was pre-trained on large-scale general audio with a self-supervised or distillation ob-

Table 4.1: System B encoder roster. Each encoder contributes one utterance embedding, obtained by time-mean pooling the tapped transformer block (one-indexed). Embedding widths differ across encoders; because fusion is at the score level (Section 4.4), the widths never have to be reconciled.

Encoder	Tapped block	Total blocks	Embed. dim.
BEATs	4	12	768
M2D-CLAP	6	12	768
SSLAM	3	12	768
EAT-large	12	24	1024
CED-tiny	2	12	192

jective, and each is used here with its weights held fixed. An encoder maps a 16 kHz mono waveform to a single utterance-level vector by mean-pooling the token (time) axis of one tapped transformer block; the pooled vector is the embedding. The mean is taken over the full token sequence of the tapped block, with no class token and no exceptions across the five encoders. For M2D-CLAP, the contrastive audio projection head is deliberately bypassed, so that scoring uses the layer-tap representation rather than the projection.

The tapped block differs by encoder, as does the embedding width. Table 4.1 lists, for each encoder, the block used (one-indexed, so block 1 is the first transformer block), the encoder's total block count, and the resulting embedding dimension. The layers were not chosen uniformly; the selection procedure, and the part of it that depends on development-set labels, is set out in Section 4.3.3.

4.3.2 Memory Bank and Scoring

For each machine type, the normal training clips are split by their explicit source or target domain label into two memory banks, roughly 990 source embeddings and 10 target embeddings, rather than by a positional cut of the data. A test clip is scored against each bank by the distance to its single nearest neighbour, that is, $k = 1$, so the usual mean-of-k-nearest-distances reduces to the nearest-neighbour distance. The distance is Euclidean (L_2); the embeddings are not normalized to unit length before scoring, so the metric is genuine Euclidean distance and not cosine similarity.

The domain imbalance built into the task, 990 source clips against 10 target clips, would otherwise make the sparsely populated target region read as anomalous for every clip. The original GenRep system addresses this with two devices: MemMixup, which augments the small target bank with interpolations toward its nearest source neighbours, and a domain normalization step that places the source and target distances on a comparable

scale [SS25]. System B reproduces the normalization but omits MemMixup; the target bank is used as recorded. Omitting it also leaves the scoring path deterministic, with no stochastic component for the fixed seed to act on.

The normalization is applied as follows. For a given encoder and machine, the nearest-neighbour distances of the test clips to the source bank and to the target bank are each standardized to zero mean and unit variance, and the per-clip score is the smaller of the two standardized values, $\min(z_{\mathrm{src}}, z_{\mathrm{tgt}})$. One point of terminology matters here. Although this step is named for local density in the original implementation, it is a per-domain z-score computed across the test set, not an estimate of the local density of the memory bank; it is described as a z-normalized nearest-neighbour score throughout this thesis to avoid implying a density estimator. The standardization uses only the distribution of the (unlabelled) test distances and never the anomaly labels, so it introduces no label leakage; it is, however, transductive, in that the whole machine's test set is needed to compute the per-domain mean and variance. This follows the convention of the reproduced system [SS25]. The statistical source of these standardization constants is treated in full in Section 4.4, because the same mechanism carries over to channel fusion.

The five encoders are combined into a single System B score by an unweighted arithmetic mean of their per-clip scores. There are no per-encoder weights, so the combination introduces no quantity tuned on the data. Because the combination is at the score level, the mixed embedding widths of Table 4.1 (192, 768, and 1024) never need to be reconciled: each encoder emits its own scalar score, and only the scores are averaged.

4.3.3 Encoder and Layer Selection

How the encoders and their layers were chosen determines whether System B's development-set numbers are an honest estimate or an optimistic one, so the procedure is stated plainly, including the part that depends on development labels.

The set of five encoders is a fixed design choice and not the outcome of a search over encoder subsets. BEATs is the encoder shipped by the upstream GenRep system; the other four were integrated for this work, and all five are always used together, applied identically to every machine type. There is no leave-one-encoder-out selection feeding the configuration, and per-machine encoder routing, while a natural extension, is not part of System B and is left to future work.

The per-encoder layer taps fall into two cases. BEATs uses its paper-pinned layer 4, taken from the established convention for that encoder and never swept, so its choice does not depend on this dataset. The remaining four taps, for M2D-CLAP, SSLAM, EAT-large, and CED-tiny, were selected by scoring each candidate layer on the development set and taking, per encoder, the layer with the highest single-encoder harmonic-mean score.

Table 4.2: Per-encoder layer selection for the four swept encoders. Each candidate layer was scored by its single-encoder harmonic mean on the development set, and the argmax (in bold) was selected. The grids are narrow and mostly flat, which bounds the selection effect. BEATs is excluded because its layer 4 is paper-pinned, not swept.

Encoder	Candidate layers: dev hmean			Selected
M2D-CLAP	3: 0.5563	4: 0.5626	6: **0.5710**	6
SSLAM	3: **0.5748**	4: 0.5718	6: 0.5739	3
EAT-large	6: 0.5617	8: 0.5669	12: **0.5746**	12
CED-tiny	2: **0.5607**	3: 0.5607	4: 0.5601	2

Table 4.2 lists the candidate grid and the development scores that drove each choice. This selection used the development-set anomaly labels, and is therefore the same class of in-sample, label-dependent selection as a development-tuned fusion weight; it is not label-free, and is not described as such.

One honest statement bounds the consequence of this. The development figure is selection-biased and should be read as an in-sample, optimistic estimate rather than an unbiased measure of generalization. The bias is small and bounded: the candidate grids span only three points and are mostly flat, and at the ensemble level the layer optimization moves the uniform five-encoder development harmonic mean from approximately 0.5742, with last-block taps, to approximately 0.5780, with the selected taps, a shift of about +0.004. That delta is the magnitude of the selection effect, and it is small enough that the selected configuration cannot be more than a few thousandths of a point better than an unselected one.

4.4 Dual-Channel Late Fusion

The thesis evaluates a single mechanism for using the second microphone: score-level, or late, fusion. Each detector is run independently on the near and the far channel, producing two anomaly-score streams per clip, and the two streams are combined into one score. Fusion is kept at the score level deliberately. Earlier explorations of waveform-level dual-channel preprocessing and of embedding-level fusion did not yield a usable detection gain and are not adopted as the method here; the embedding-level results are reported alongside the score-level ablation in Chapter 5, while the waveform-level preprocessing is set aside and not pursued further. Late fusion is the clean, minimal way to ask whether the second channel adds anything, and it is the only fusion method this thesis reports as its result.

The operator. The fusion is an equal-weight combination of the two channels. For each detector, the per-clip anomaly scores from ch0 and from ch1 are first put on a common scale by the same per-domain z-score used inside System B (Section 4.3.2), and the standardized channel scores are then averaged with equal weight. Writing s_{ch0} and s_{ch1} for the standardized per-clip scores of the two channels, the fused score is their unweighted mean,

$$s_{\text{fused}} = \tfrac{1}{2}\left(s_{\text{ch0}} + s_{\text{ch1}}\right).$$

The standardization constants are estimated transductively on the test set, from the distribution of the (unlabelled) test scores per domain, exactly as in single-channel scoring; they never use the anomaly labels.

Why uniform, and why not tuned. The equal weight is a methodological choice, not a fitted parameter. A per-machine channel weight, selected by searching the development set for the mixture that maximizes the score, would have no held-out validation behind it, and would be contaminated in the same way as a development-tuned threshold: the reported gain would partly measure the search rather than the channel. An earlier autoencoder study that tuned a per-machine channel weight on the development set is treated as exactly this kind of optimistic baseline and is not used as a thesis result. Uniform fusion needs no tuning, so it has no development-set dependence to disclose, which is precisely why it is the configuration reported here. It also sets a deliberately conservative bar: if equal-weight fusion does not help, a weighting that would have to be learned per machine cannot be claimed to help without paying for that learning honestly.

Applied to both detectors. The same operator is applied to System A and to System B without change, so that the dual-channel comparison is made on equal terms across the two detector families. This is what allows the cross-detector test of RQ3: a fusion outcome can be compared between detectors only because the fusion rule itself is identical for both.

4.5 Evaluation Protocol

Metrics. Every configuration is scored with the challenge metrics. For each machine type, three quantities are computed: the source-domain AUC (source-normal clips against all anomalies), the target-domain AUC (target-normal clips against all anomalies), and the partial AUC at a maximum false-positive rate of 0.1. These per-machine, per-domain quantities are condensed into the single official score Ω, the harmonic mean over all machines and domains, defined in Section 2.3 and not repeated here. The harmonic mean is reported for completeness but is not the primary lens, for the reason given when it was introduced: it is dominated by its weakest cells, so a single broken machine-domain combination can hold it down while the arithmetic average looks healthy, and conversely an aggregate can look healthy while concealing a machine on which a method does nothing.

Reporting lens. The primary unit of analysis is therefore the individual machine and domain, not the aggregate. Results in Chapter 5 are read per machine, with the source and target domains kept separate, and the aggregate Ω is reported alongside as a secondary summary. The target-domain figures are treated with particular caution: each machine provides only about ten target-normal training clips (Chapter 3), so target metrics are high-variance, and a target AUC that moves by a few points between configurations is read as sampling noise unless a confidence interval separates it from zero.

Statistical testing. The verdict on whether fusion beats the near channel is made with a paired bootstrap rather than from point estimates. For each machine, the test clips are resampled with replacement 1000 times, with a fixed seed, and stratified within each of the four (domain $\times$ label) cells, source-normal, source-anomaly, target-normal, and target-anomaly, so that every replicate contains the clips needed for each AUC and partial AUC to be defined. The resampling unit is the test clip, stratified by domain and label, and not the machine. The bootstrap indices are drawn independently per machine but reused across the configurations being compared, so that the difference between two configurations, for example the fusion-minus-ch0 change in Ω, is a paired difference in which the shared test-set noise cancels. Within each replicate the per-machine metrics are recomputed and aggregated exactly as in scoring, including the transductive per-domain standardization on the resampled test set, and the reported confidence interval is the 2.5 to 97.5 percentile range over the 1000 replicates [Efr79]. The per-machine intervals are reported without correction for multiple comparisons; across seven machines and several fusion variants, roughly one interval in twenty is expected to exclude zero by chance alone, so no individual per-machine interval is read as decisive on its own, and the verdict on fusion rests on the aggregate paired deltas.

Fusion baseline. The reference for every fusion comparison is the ch0 configuration of the same detector. Measuring a System A fusion against the System A near channel, and a System B fusion against the System B near channel, keeps the fusion gain from being credited against the wrong baseline, and is what makes the cross-detector comparison of RQ3 a comparison of like with like.

Development-set scope. All AUC, partial AUC, and bootstrap quantities reported in this thesis are development-set measurements, computed on the seven development machine types with the ground-truth labels released for that set.

Chapter 5

Results and Discussion

This chapter reports the development-set results of the comparison defined in Chapter 4 and uses them to answer RQ2 and RQ3. Every number below is a development-set measurement. The chapter first fixes the single-channel foundation, opens the aggregate score to show which cells set it, then turns to the dual-channel question and the cross-detector comparison that is the core of the thesis, the mechanism that accounts for the outcome, and finally the one result that is deliberately excluded.

5.1 Reading Guide

The comparison is the same throughout: the near channel (ch0), the far channel (ch1), and their uniform late fusion, each run on System A and System B and read per machine. The official harmonic-mean score Ω (Section 4.5) is reported for comparability between systems, but the primary lens is the per-machine, per-domain result, not the aggregate. The reason is built into the metric: because Ω is a harmonic mean over every machine-domain cell, it reports the weakest cell rather than average competence, so a single configuration's Ω can move for opposite reasons on different machines, and two systems with the same Ω can disagree completely about which cells they handle well. That one property licenses the entire emphasis of this chapter on the per-machine, per-domain view. Aggregates are therefore used to state the headline direction, and the per-machine tables are used to establish what is actually happening underneath it. The metric definitions are not repeated here; they are given in Section 4.5.

Table 5.1: Single-channel (ch0) aggregate scores on the development set. AUC values are arithmetic means over the seven machine types; Ω is the official harmonic mean over all machine-domain cells. System A is shown in both scoring modes; System B is the layer-optimized frozen-embedding detector.

System	Scoring	AUC_{src}	AUC_{tgt}	pAUC	Ω
System A	MSE	0.688	0.536	0.547	0.5716
System A	Selective Mahalanobis	0.674	0.565	0.538	0.5798
System B	frozen-embedding kNN	0.625	0.617	0.523	0.5780

5.2 Single-Channel Reproduction

Both detectors are reproduced on the near channel before any stereo question is asked, so that every later fusion result is measured against a trustworthy single-channel reference. Table 5.1 reports the aggregate scores. System A, the official baseline, reaches $\Omega = 0.5716$ in its MSE scoring mode and 0.5798 in Selective Mahalanobis mode. System B, the frozen-embedding detector in its layer-optimized configuration, reaches $\Omega = 0.5780$.

These numbers earn one claim, and it is a foundation rather than a finding. System B exceeds the official autoencoder under its MSE scoring (0.5780 against 0.5716) and sits within a fraction of a point of its stronger Mahalanobis scoring (0.5780 against 0.5798), so the training-free detector is reproduced at a strength comparable to the official baseline. The two detector families are therefore both trustworthy and close enough in aggregate to make a like-for-like stereo comparison meaningful. Their internal profiles differ, which is the point of using both: System A leans on the source domain (AUC_{src} 0.674 against AUC_{tgt} 0.565), whereas System B is more domain-balanced (AUC_{src} 0.625, AUC_{tgt} 0.617), so an effect that appears in both is unlikely to be an artefact of one detector's domain bias.

5.3 Single-Channel Per-Machine Results

Table 5.2 gives the per-machine ch0 results for both detectors, with each detector occupying its own block of columns so that the two can be read separately as well as side by side. This is the reference against which every fusion delta in the following sections is measured; no fusion is applied yet, and nothing here speaks to whether the second channel helps. What it shows is heterogeneity. Per-machine Ω spans 0.517 to 0.625 on System A and 0.499 to 0.656 on System B, a range of more than ten points within a single detector, so the aggregate near 0.58 is an average over machines that the detector handles very differently rather than a uniform level of competence.

Table 5.2: Per-machine single-channel (ch0) results. System A in Selective Mahalanobis mode; System B in the layer-optimized configuration. Ω is the per-machine harmonic mean of the three quantities to its left. In the bottom row, the AUC and pAUC entries are arithmetic means over the seven machines, while Ω is the official harmonic mean over all machine-domain cells, so the aggregate Ω is not the mean of the per-machine Ω column.

	System A (Mahalanobis)				System B (frozen-embedding)			
Machine	AUC_{src}	AUC_{tgt}	pAUC	Ω	AUC_{src}	AUC_{tgt}	pAUC	Ω
`ToyCar`	0.781	0.550	0.579	0.622	0.739	0.735	0.537	0.656
`ToyCarEmu`	0.676	0.733	0.512	0.625	0.638	0.687	0.537	0.614
`bearingEmu`	0.647	0.621	0.601	0.622	0.584	0.591	0.578	0.585
`fan`	0.592	0.452	0.527	0.517	0.510	0.480	0.508	0.499
`gearboxEmu`	0.765	0.554	0.544	0.606	0.659	0.594	0.501	0.577
`sliderEmu`	0.681	0.495	0.502	0.547	0.590	0.536	0.488	0.535
`valveEmu`	0.577	0.551	0.501	0.541	0.658	0.691	0.511	0.609
Aggregate	0.674	0.565	0.538	0.5798	0.625	0.617	0.523	0.5780

The two detectors do not rank the machines the same way. System A is strongest on the cleaner, source-dominated machines (`ToyCar`, `ToyCarEmu`, `bearingEmu` all above $\Omega = 0.62$) and weakest on `fan` (0.517), while System B peaks on `ToyCar` (0.656) and `valveEmu` (0.609) and also bottoms out on `fan` (0.499). The contrast on `valveEmu` is the sharpest: it is one of System A's weakest machines (0.541) and one of System B's strongest (0.609), which is the kind of detector-specific profile that makes a two-detector study worthwhile. The one point of agreement is the shared weakness on `fan`, the most stationary and tonal machine, consistent with the difficulty noted for it in Chapter 3. Section 5.4 takes this heterogeneity apart to show that it is concentrated in particular cells rather than spread evenly, which is what sets the aggregate.

5.4 What Governs the Aggregate Score

Both detectors aggregate to $\Omega \approx 0.58$ on the near channel, but that number is not an average of competence. It is a harmonic mean over the twenty-one per-machine, per-domain quantities, the source AUC, the target AUC, and the partial AUC of each of the seven machines, and a harmonic mean is set by its smallest arguments. This section opens the aggregate to show which cells set it and what they have in common, because the answer is what the mechanism in Section 5.7 has to explain.

The first measurement is how much the weak cells cost. Table 5.3 reports, for each

Table 5.3: Composition of the single-channel aggregate. Ω is the official harmonic mean over the twenty-one per-machine, per-domain cells; the arithmetic mean is over the same cells; the gap is what the harmonic mean charges for unevenness. The right block gives the mean of each axis over the seven machines, showing the partial AUC as the global floor on both detectors.

	Aggregate			Mean per axis		
Detector	Ω	arith. mean	gap	$\mathrm{AUC}_{\mathrm{src}}$	$\mathrm{AUC}_{\mathrm{tgt}}$	pAUC
System A (Mahalanobis)	0.5798	0.5924	0.0126	0.674	0.565	0.538
System B (frozen-embed.)	0.5780	0.5883	0.0103	0.625	0.617	0.523

detector, the official Ω and the plain arithmetic mean of the same twenty-one cells. The arithmetic mean is higher than Ω on both detectors, by 0.0126 on System A and 0.0103 on System B, and that gap is exactly the penalty the harmonic mean charges for unevenness: it is the amount by which the weak cells pull the official score below the level the detector reaches on average. The gap is moderate here only because no cell is catastrophic. Were any single cell to fall toward 0.40, Ω would track it down far faster than the arithmetic mean would, which is the property that makes the worst cell, not the average, the thing worth examining.

The second measurement is what the weak cells have in common, and the answer is the same on both detectors: they are not whole machines, they are particular axes. The three quantities fall in a fixed order when averaged over machines (right block of Table 5.3): the source AUC is highest, the target AUC next, and the partial AUC lowest, on both detectors. Ranking all twenty-one cells from worst to best makes this concrete. On System A the seven lowest cells are five partial-AUC cells and two target-AUC cells, with no source cell among them; the three worst are `fan` target (0.452), `sliderEmu` target (0.495), and `valveEmu` partial AUC (0.501). On System B the seven lowest are four partial-AUC, two target, and one source cell, the three worst being `fan` target (0.480), `sliderEmu` partial AUC (0.488), and `gearboxEmu` partial AUC (0.501). The floor is therefore the low-false-alarm region and the target domain, spread across the harder machines, rather than a few machines collapsing across every axis.

This reframes a loose intuition into a measured property. The aggregate is held near 0.58 by the partial-AUC and target-domain cells of the hard machines, and it would not rise even if the strong machines scored higher, because a harmonic mean cannot be bought up by its best arguments. The one machine that is weak on two axes at once, `fan` (target and partial AUC on both detectors), is the genuine worst machine rather than a single worst cell, which is why it surfaced already as the shared weakness in Section 5.3. Why these particular cells are weak, the sparse transients of the valve, the tonal masking of the fan, and the ten-clip target domain, is an acoustic question taken up in Section 5.7.

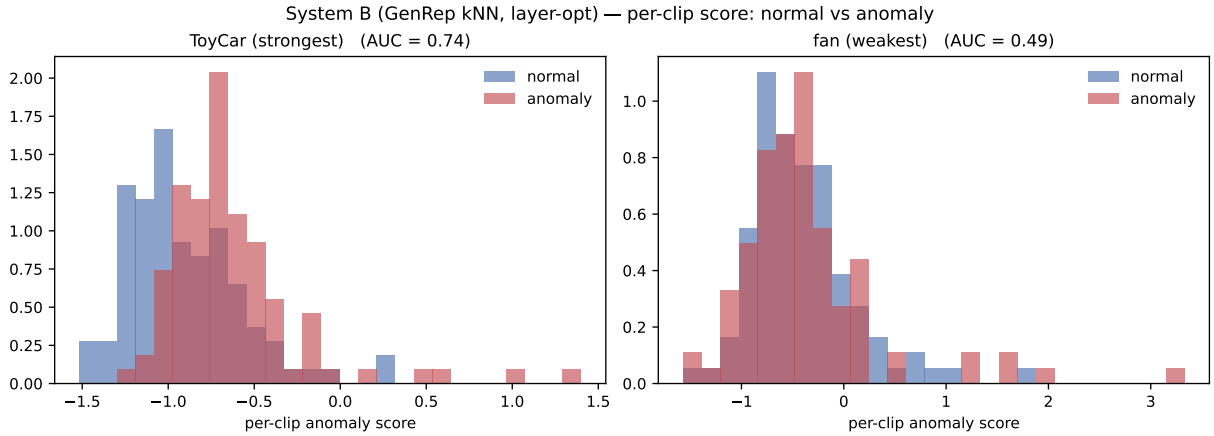

Figure 5.1: Per-clip anomaly-score distributions, normal versus anomaly, under System B for the strongest machine (`ToyCar`, left) and the weakest (`fan`, right). The `ToyCar` distributions are offset; the `fan` distributions coincide, so the detector's two classes are indistinguishable in score. The printed AUC is an illustrative mixed-domain separation, not the per-domain Ω.

The point here is structural: the aggregate is a statement about the worst cells, so the per-machine, per-domain reading used throughout this chapter is not a stylistic choice but the only way to see what the single number is actually reporting.

The score distributions behind these cells make the point concrete. Figure 5.1 contrasts the per-clip anomaly scores of the strongest and weakest machines under System B: for `ToyCar` the normal and anomaly distributions are visibly offset, whereas for `fan` they sit almost exactly on top of one another. The floor is therefore not a thresholding artefact that a better operating point could rescue; on the weak machine the two classes are not separated in score at all, which is the strongest form a weak cell can take. The separation values printed on the figure are illustrative mixed-domain AUCs and are distinct from the per-domain Ω of Table 5.2.

5.5 Dual-Channel Fusion (RQ2)

The sections above measured how well each detector performs per machine on the near channel. This section asks a different question on a different axis, and the two must not be run together: whether the second microphone adds anything is independent of how strong the near channel is. The strong per-machine results of Section 5.3 are near-channel results, and none of what follows turns them into fusion gains. The dual-channel question is whether adding the far microphone improves on the near channel alone. It does not. On

Table 5.4: Aggregate Ω across the three input configurations on both detectors (DCASE convention). The fusion row is uniform equal-weight late fusion. Deltas are the fusion-minus-ch0 change with the 95% paired-bootstrap interval. System B is shown here at its last-layer ch0 (0.5742), the anchor of the paired fusion comparison; its layer-optimized single-channel score (0.5780) is the one reported in Section 5.3, and the relationship between the two is set out in the note below.

Configuration	System A (Mahalanobis)	System B (frozen-embedding)
ch0 (near)	0.5798	0.5742
ch1 (far)	0.5076	0.5278
uniform fusion	0.5524	0.5595
fusion − ch0	-0.0274 $[-0.0490, -0.0081]$	-0.0147 $[-0.0260, -0.0035]$

both detectors, uniform late fusion lowers the aggregate score relative to ch0, and on both the loss is statistically significant.

Table 5.4 places the two detectors side by side across the three input configurations. The far channel alone is markedly weaker than the near channel on both detectors (System A $\Omega = 0.508$, System B 0.528, against ch0 near 0.58), and uniform fusion lands between the two channels rather than above the better one: System A falls from 0.5798 to 0.5524, a change of -0.0274 with a 95% paired-bootstrap interval of $[-0.0490, -0.0081]$, and System B falls from 0.5742 to 0.5595, a change of -0.0147 with an interval of $[-0.0260, -0.0035]$. Both intervals lie entirely below zero, so neither loss is attributable to sampling noise. The far channel costs System A more than System B, which follows from System A's far channel being the weaker of the two ($\Omega = 0.508$ against System B's 0.528).

Table 5.5 breaks the far channel down per machine, and three regularities matter for what follows. First, the far channel is the weaker detector input on thirteen of the fourteen machine-detector pairs; the only exception is `valveEmu` under System B, where ch1 is marginally the stronger channel (0.618 against 0.609), and `fan` under System B is effectively tied (0.493 against 0.499). Second, the per-machine fusion outcomes of Section 5.6 broadly track this strength gap: the largest significant fusion losses fall on the machines with the largest ch1 deficits (`ToyCar` under System B, deficit -0.141; `fan` under System A, deficit -0.104), and the positive point estimates cluster where the deficit is small. This is the behaviour classifier-combination theory predicts when an equal-weight rule fuses scorers of unequal strength [KHDM98, Kun04], and Section 5.7 treats it as the proximate layer of the explanation. The tracking is broad rather than exact, and its clearest exception is instructive: on `valveEmu` under System B the two channels are of essentially equal strength, yet fusion still loses significantly (-0.0265, Table 5.7), so channel strength alone does not decide the fusion outcome, and this is the case the deeper levels of Section 5.7 must carry. Third, the far channel's weakness is not fully predicted by the channel analysis

Table 5.5: Per-machine far-channel (ch1) score Ω and the ch1 − ch0 strength gap for both detectors. System A is in Selective Mahalanobis mode. System B ch1 values are computed on the last-layer caches, the anchor of the fusion study (Table 5.4); its ch0 references are the layer-optimized per-machine values of Table 5.2, so the System B gaps carry a small configuration offset, about 0.004 at the aggregate (Section 4.3.3). A negative gap means the far channel is the weaker detector input on that machine.

	System A (Mahalanobis)		System B (frozen-embedding)	
Machine	Ω (ch1)	ch1 − ch0	Ω (ch1)	ch1 − ch0
`ToyCar`	0.536	−0.086	0.515	−0.141
`ToyCarEmu`	0.610	−0.015	0.568	−0.046
`bearingEmu`	0.531	−0.091	0.529	−0.056
`fan`	0.413	−0.104	0.493	−0.006
`gearboxEmu`	0.501	−0.105	0.504	−0.073
`sliderEmu`	0.497	−0.050	0.490	−0.045
`valveEmu`	0.507	−0.034	0.618	+0.009
Aggregate	0.508	–	0.528	–

of Chapter 3: on `bearingEmu`, where the SNR proxy reports the two channels as tied, ch1 is still clearly the weaker input on both detectors (0.531 against 0.622 on System A, 0.529 against 0.585 on System B), so per-channel detection quality is governed by more than the scalar SNR contrast.

The pattern is not specific to the equal-weight blend. Table 5.6 reports, against the near-channel baseline, the far channel alone and five fusion variants of System B: three score-level blends at different weights and two embedding-level concatenations. The far channel alone and four of the five fusion variants have a 95% paired-bootstrap interval lying entirely below zero, and no variant rises above ch0. The single fusion variant whose interval includes zero is `score_fusion_w0.7`, the most near-channel-dominant blend, whose interval $[-0.0132, +0.0005]$ barely crosses it; it does not improve on ch0, it merely fails to significantly damage it by leaning almost entirely on the near channel. Embedding-level fusion fares no better than score-level fusion (-0.0167 and -0.0176), which matters for the mechanism discussed in Section 5.7: combining the channels earlier, inside the encoder's representation, does not recover anything.

A note on the System B anchor. The single-channel result of Section 5.3 uses the layer-optimized configuration ($\Omega = 0.5780$), whereas the dual-channel variants in this section are computed on the earlier last-layer caches, where ch0 scores 0.5742. The two channels in the fusion study therefore use the same last-layer taps, which keeps each fusion delta a clean paired comparison against its own ch0; the small 0.5742-versus-0.5780 gap is the

Table 5.6: The far channel alone (`ch1_only`) and five fusion variants of System B against the near-channel baseline ($\Omega = 0.5742$, pre-layer-optimization caches). $\Delta\Omega$ is the change from ch0 with its 95% paired-bootstrap interval (1000 resamples, stratified by domain and label, paired per machine). The uniform equal-weight blend is `score_fusion_w0.5`.

Variant	Ω	$\Delta\Omega$ vs ch0	95% CI of $\Delta\Omega$
`ch0_only` (reference)	0.5742	–	–
`ch1_only`	0.5278	-0.0464	$[-0.0697, -0.0255]$
`score_fusion_w0.3`	0.5490	-0.0252	$[-0.0409, -0.0095]$
`score_fusion_w0.5` (uniform)	0.5595	-0.0147	$[-0.0260, -0.0035]$
`score_fusion_w0.7`	0.5676	-0.0066	$[-0.0132, +0.0005]$
`embedding_fusion`	0.5575	-0.0167	$[-0.0281, -0.0047]$
`embedding_fusion_l2`	0.5566	-0.0176	$[-0.0288, -0.0057]$

Table 5.7: Per-machine uniform-fusion change $\Delta\Omega$ (fusion − ch0) for both detectors. A star marks a 95% paired-bootstrap interval that excludes zero. The final column records whether the two detectors agree on the sign of the effect.

Machine	System A $\Delta\Omega$	System B $\Delta\Omega$	Sign agreement
`ToyCar`	-0.0319	$-0.0454^\star$	yes
`ToyCarEmu`	$+0.0138$	$+0.0036$	yes
`bearingEmu`	-0.0073	-0.0129	yes
`fan`	$-0.0912^\star$	$+0.0073$	no
`gearboxEmu`	-0.0207	-0.0278	yes
`sliderEmu`	-0.0217	-0.0187	yes
`valveEmu`	$+0.0061$	$-0.0265^\star$	no

effect of layer optimization on the single channel, not a fusion effect. The fusion deltas, being paired within a single configuration, are unaffected by this choice.

5.6 Cross-Detector Comparison (RQ3)

RQ3 asks whether the dual-channel outcome is a property of the second channel or an accident of one detector. The test is to place the per-machine fusion effect of the two detectors side by side and ask whether they agree. Table 5.7 and Figure 5.2 report the per-machine uniform-fusion change $\Delta\Omega$ for both detectors.

Three facts stand out, and together they settle RQ3. First, the two detectors agree on

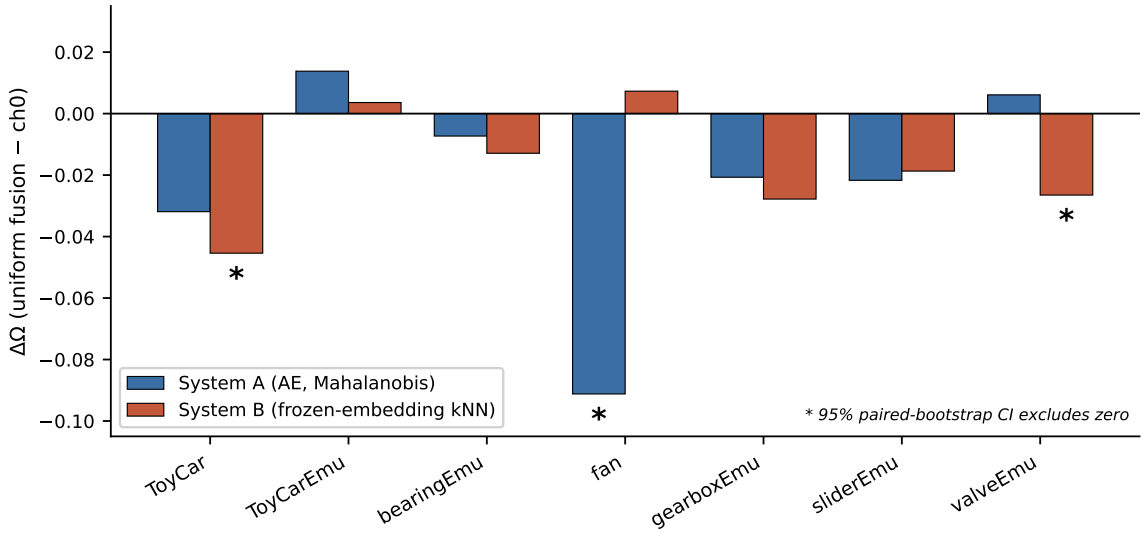

Figure 5.2: Per-machine uniform-fusion change $\Delta\Omega$ (fusion − ch0) for System A (autoencoder, Mahalanobis) and System B (frozen-embedding kNN). Bars below the zero line indicate that fusion hurts. A star marks a 95% paired-bootstrap interval that excludes zero. The two detectors agree in sign on five of seven machines, and every significant effect is a loss.

the sign of the fusion effect on five of the seven machines (`ToyCar`, `ToyCarEmu`, `bearingEmu`, `gearboxEmu`, `sliderEmu`), and the two disagreements, `fan` and `valveEmu`, are each non-significant on at least one detector. Second, no machine improves significantly under either detector: every positive $\Delta\Omega$ has an interval that crosses zero, and the only intervals that exclude zero are losses (`fan` under System A; `ToyCar` and `valveEmu` under System B). Third, the few machines with a positive point estimate are not detector-specific in a way that would suggest the channel helps somewhere: the positive-estimate sets are {`ToyCarEmu`, `valveEmu`} for System A and {`ToyCarEmu`, `fan`} for System B, which share `ToyCarEmu` and in any case lie within noise.

The contribution follows in one line. Two methodologically unrelated detectors, a reconstruction autoencoder and a frozen-embedding nearest-neighbour model, converge on the same negative: uniform near/far fusion does not help, and where it has a significant effect it hurts. Because the result replicates across two detectors that share no architecture, scoring rule, or training procedure, it is a property of the second channel and the encoders that consume it, not an artefact of one detector family. The corollary for the field is that a fusion result reported on a single detector cannot be trusted as a statement about the channel; cross-detector agreement is the necessary check, and it is exactly what is missing from a single-detector stereo claim.

One shared component qualifies this attribution and should be stated rather than left for the reader to find. The fusion operator and the per-domain standardization are identical across the two detectors by design (Section 4.4), so a negative caused by the combination rule itself, rather than by the channel, would also replicate across them. The variant study of Section 5.5 is what closes this gap: embedding-level fusion bypasses the score-averaging rule entirely and fails in the same way, and no reweighting of the blend rises above the near channel, so the convergent negative is not specific to the shared operator.

5.7 Why Fusion Fails

At the surface the failure is arithmetic. Uniform fusion averages the near channel with a far channel that Table 5.5 shows to be the weaker detector input on essentially every machine, and an equal-weight rule that fuses scorers of unequal strength is expected to degrade the stronger one [KHDM98, Kun04]; the per-machine fusion losses accordingly track the ch1 deficit. What this arithmetic does not explain is why the far channel is the weaker input in the first place, and why fusion fails even where the deficit is small or absent, as on `valveEmu` under System B. That explanation has two levels, and the two levels reinforce rather than compete with each other.

The first level is the data. As established in Chapter 3, the near/far SNR premise holds for only the two ToyCar variants, where the gap is about 3 dB, and it is essentially absent for the other five machines, falling below 1.6 dB and inverting slightly on `bearingEmu`, where the far microphone is marginally the cleaner of the two. On five of seven machines there is therefore almost no inter-channel contrast for any fusion scheme to exploit, so the second channel can only dilute the first. This alone would explain the losses on those five machines.

It does not, however, explain `ToyCar`. There a real 3 dB gap exists, yet fusion still fails, and significantly so under System B. For this we offer a second, architectural explanation, as a hypothesis rather than a demonstrated result. The encoders in both detectors were pre-trained on monaural audio and are applied to each channel independently, so they plausibly project the near and the far channel into much the same monaurally learned representation. On this account, whatever spatial relationship distinguishes the two microphones is not preserved by an encoder that never learned to represent two channels jointly, so the far channel contributes a noisier copy of the same information rather than a complementary view, and averaging the two can only move the score toward the weaker channel, the behaviour that classifier-combination theory predicts when an unweighted rule fuses scorers of unequal strength [KHDM98, Kun04]. The hypothesis is consistent with, though not proven by, the embedding-level variants of Section 5.5: fusing the channels earlier, inside the representation, fails just as score-level fusion does ($\Delta\Omega = -0.0167$ and -0.0176), which

is what one would expect if the representation, rather than the score-combination step, is where the inter-channel information is lost. A direct test would require an encoder that ingests both channels jointly, which this study does not build, so the account remains a hypothesis the evidence is consistent with, not an established mechanism.

Taken together, the two levels cover every fusion mode in both detector families. On the five low-contrast machines the channel carries little to add; on the two high-contrast machines the second channel still fails to help, which the architectural hypothesis attributes to the frozen monaural encoder's inability to represent what it adds. There is no configuration in the study, score-level or embedding-level, weighted or uniform, on either detector, in which the second channel improves detection, and the two-level account is consistent with why none does.

A separate thread connects the absolute per-cell difficulty of Section 5.4 to the acoustics of the machines, and it explains why the floor sits where it does. The weakest cells are not arbitrary. The valve produces sound only in brief actuation transients separated by near-silence, so its anomalies are sparse events that a clip-level score averages away, which depresses its partial AUC, the metric most sensitive to the high-confidence tail. The fan is the opposite case: a continuous tonal source whose fault signatures are low-frequency shifts buried under a strong aerodynamic floor, which masks anomalies and leaves it the weakest machine on both detectors. The target domain is weak across machines for a reason that is not acoustic but statistical: with only ten target-normal training clips against roughly nine hundred and ninety source clips (Chapter 3), the target distribution is undersampled, so its cells are both lower and noisier than the source cells. These three causes, sparse transients, tonal masking, and target undersampling, account for the composition of the floor identified in Section 5.4: partial-AUC and target cells on the transient and tonal machines.

Two further measurements make the channel-level half of the account concrete. First, the inter-channel analysis of Section 3.3 (Table 3.3) shows that the far channel rarely carries complementary information to begin with: on the ToyCar variants it is a weak, decorrelated copy of the near channel, and on `fan` it is a near-duplicate, correlated at 0.87 with the near channel and marginally louder, as the near and far `fan` spectrograms in Figure 5.3 make visible. Fusing a clean near channel with a channel that is either weak-and-decorrelated or a near-duplicate can only inject noise or repeat what is already present, which is what the uniform-fusion losses register, and it is a structure the scalar SNR gap cannot see. Second, the difficulty of the weakest machine is a property of the data rather than of one detector. Figure 5.4 shows that under System A, as under System B (Figure 5.1), the per-clip normal and anomaly score distributions for `fan` coincide: neither a reconstruction model nor a frozen-encoder nearest-neighbour detector finds class-separating structure in fan's sound. A difficulty shared by two unrelated detectors sits in the signal, not the method, which is the same logic that gives the convergent-negative fusion result its force.

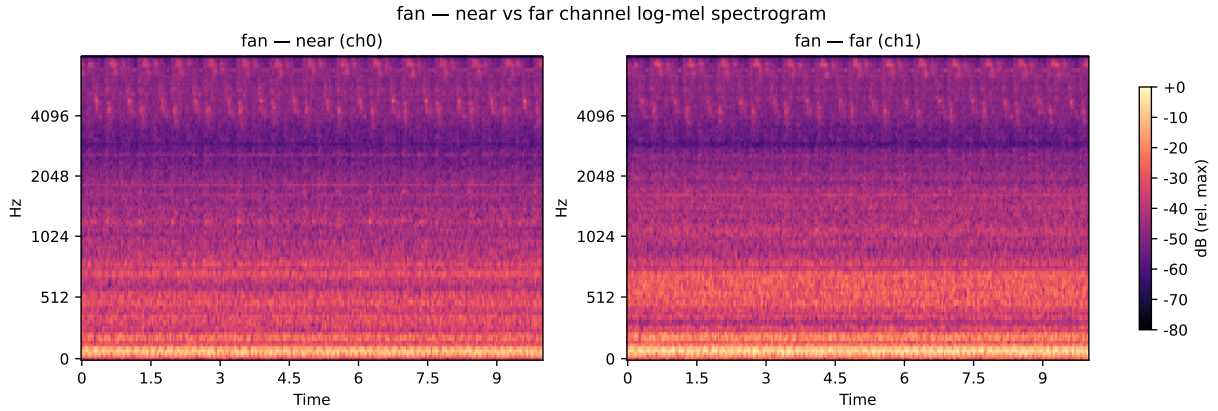

Figure 5.3: Near (ch0) versus far (ch1) log-mel spectrograms of a representative `fan` clip. The two channels are visually almost interchangeable, consistent with their inter-channel correlation of 0.87 (Table 3.3): the far microphone supplies a near-duplicate of the near channel rather than a complementary view.

5.8 Validity and the Excluded Result

One number is deliberately absent from the results above, and the reason is methodological. An earlier dual-channel analysis on System A selected a separate fusion weight for each machine by searching the development set for the value that maximized the score. That procedure reached an apparent $\Omega = 0.5881$, a nominal improvement over the near channel. It is not reported as a result of this thesis, because the per-machine weights were fitted on the same development set used to evaluate them, with no held-out split: the apparent gain measures the search as much as the channel, and a weight chosen this way carries no guarantee of generalizing. It is the same contamination class that motivated the decision, in Chapter 4, to report only the uniform equal-weight fusion, which has no tuned quantity to overfit. Stating the excluded number plainly is the more rigorous choice: a reader can see that the only configuration that appears to help is the one whose apparent help cannot be separated from the tuning, which is precisely why uniform fusion is the honest comparison and why the conclusion of this chapter rests on it.

A second number invites the same scrutiny and is treated more leniently for a specific, bounded reason. System B's headline score of 0.5780 comes from its layer-optimized configuration, in which four of the five encoder taps (all but the paper-pinned BEATs layer) were chosen by maximizing the single-encoder development harmonic mean with no held-out split (Chapter 4, Table 4.2). That selection is, strictly, the same contamination class as the per-machine fusion weight: it consults the development labels to choose a configuration and then reports that configuration on the same data. It is reported here, where the fusion weight is not, because its effect is small and quantified. The last-layer ensemble, which

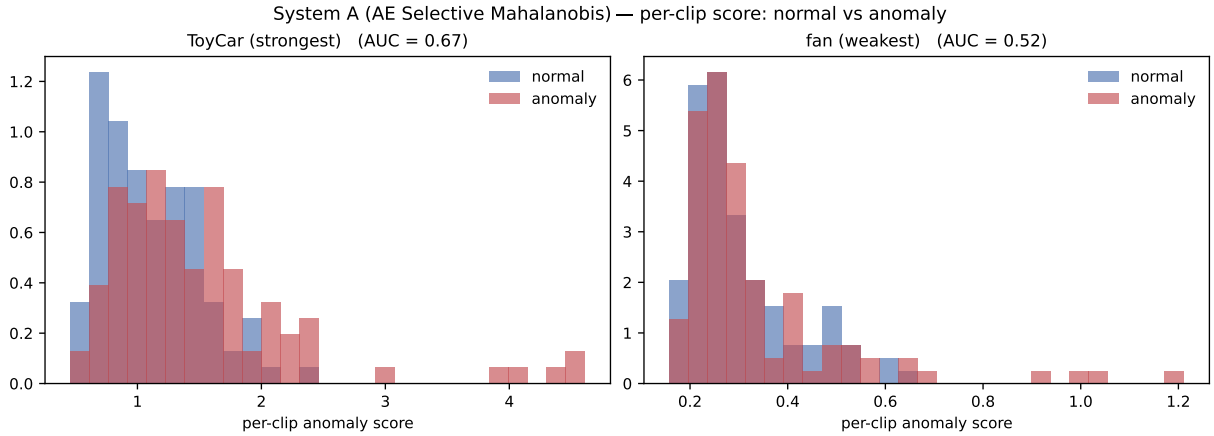

Figure 5.4: Per-clip anomaly-score distributions, normal versus anomaly, under System A for `ToyCar` (left) and `fan` (right). As with System B (Figure 5.1), `fan`'s two classes coincide in score, so its difficulty replicates across the two unrelated detectors and is a property of the data. The printed AUC is an illustrative mixed-domain separation, not the per-domain Ω.

involves no layer selection, scores $\Omega = 0.5742$; layer optimization raises this to 0.5780, a difference of about 0.004, so the selection bias carried by the System B headline is at most a few thousandths of a point. The difference from the fusion weight is not the magnitude of the apparent gain, which is comparable, but its nature: the layer choice is a selection among three discrete, nearly tied options per encoder, with little room to overfit, the candidate grids being narrow and flat (spanning at most 0.015 in single-encoder Ω), whereas the per-machine fusion weight is a continuous parameter fitted independently for each of the seven machines, with no such bound on how far the search can chase development-set noise. The honest summary is that 0.5780 is a mildly optimistic development estimate with a bias bounded below half a hundredth of a point, reported with that caveat, whereas 0.5881 is an unbounded artefact of per-machine search and is reported as excluded.

Chapter 6

Conclusion

This thesis built two detectors for the 2026 noise-aware task and used them as fixed instruments to test whether the second microphone improves unsupervised anomalous sound detection. This chapter summarizes what that test established, the limits within which it holds, and the directions those limits point to.

6.1 Summary of Findings

The foundation, which settles no research question on its own, is that both detectors were reproduced and are trustworthy: the frozen-embedding nearest-neighbour detector (System B) reaches parity with the official autoencoder baseline (System A) on the development set, giving two independent detector families on which the stereo question could be tested.

On that foundation, the answer to RQ2 is negative: the second channel does not help under any fusion mechanism evaluated. Uniform late fusion of the near and far channels lowers the official score Ω on both detectors, by 0.0274 on System A and by 0.0147 on System B, and in both cases the 95% paired-bootstrap interval lies entirely below zero. The loss is not an artefact of the equal-weight blend, since four of the five fusion variants of System B, at score level and embedding level alike, fall significantly below the near-channel baseline and none rises above it.

The answer to RQ3 is that this negative replicates. Two detectors that share no architecture, scoring rule, or training procedure agree on the sign of the fusion effect on five of the seven machines and show no significant gain on any. Because the outcome is the same across two unrelated detectors, it is a property of the second channel and of the encoders that consume it, not an accident of one detector family. The corollary is the methodological point this thesis contributes: cross-detector agreement is the check that a single-detector

stereo claim lacks, and without it a fusion gain measured on one detector cannot be read as a statement about the channel.

A two-level account is offered for why fusion fails. At the data level, the near/far premise that motivates the task holds for only the two ToyCar variants, where the inter-channel gap is about 3 dB, and is absent or inverted on the other five machines. At the architectural level, a proposed explanation, consistent with the failure of embedding-level fusion but not directly tested here, is that the frozen monaural encoders project both channels into much the same monaurally learned representation and cannot preserve the inter-channel structure, which would account for fusion adding nothing even on the ToyCar machines where a real gap exists (Section 5.7).

A secondary result of the single-channel analysis supports the per-machine lens used throughout. The aggregate score of each detector is held near its level not by uniform competence but by the partial-AUC and target-domain cells of the hardest machines (Section 5.4); the harmonic-mean aggregate therefore reports the weakest cells rather than the average, and only a per-machine, per-domain reading reveals what the single number conceals.

6.2 Limitations

Several limitations bound these conclusions, each noted where it arises in the body. The results are development-set measurements only (Section 4.5). Only late, uniform fusion was evaluated, so the study shows that this fusion does not help, not that no learned fusion could. The signal-to-noise figure is a within-clip percentile proxy rather than a calibrated denoising gate, and is read only as such (Section 3.2.1). The per-domain standardization is transductive, so a clip's score depends on the test batch it is normalized within rather than on training statistics alone (Section 4.3.2). System B omits the MemMixup augmentation of the original GenRep design, so its handling of the sparse target domain rests on the bank standardization alone. Each detector family is represented by a single instance, one autoencoder and one frozen-encoder detector, so the cross-detector replication rests on two systems rather than many. And the target domain provides only ten normal clips per machine, which makes every target-domain quantity high-variance.

6.3 Future Work

Each direction below addresses one of those limitations or the mechanism behind the central result. The binding constraint, on the account proposed in Section 5.7, is architectural: if frozen monaural encoders cannot preserve the inter-channel structure, the primary direc-

tion is to replace them with channel-aware or stereo-pretrained encoders that represent the two microphones jointly, which would be the route by which a genuine near/far gap could be turned into a detection gain. Any learned fusion weight, the possibility the uniform-only design deliberately leaves open, should be fitted on a held-out validation split rather than on the evaluation data, which is the fix for the contamination that forced the uniform comparison. Finally, per-machine encoder or layer routing, already set aside in the methods (Section 4.3.3), is a natural extension given the per-machine heterogeneity documented here, provided any such selection is made without consulting test labels.

Bibliography

[AAM+25] Tony Alex, Sara Atito, Armin Mustafa, Muhammad Awais, and Philip J. B. Jackson. SSLAM: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL: https://openreview.net/forum?id=odU59TxdiB, arXiv:2506.12222.

[AUN+19] Muhammad Altaf, Muhammad Uzair, Muhammad Naeem, Adeel Ahmad, Syed Badshah, Jawad A. Shah, and Almas Anjum. Automatic and efficient fault detection in rotating machinery using sound signals. *Acoustics Australia*, 47(2):125–139, 2019. doi:10.1007/s40857-019-00153-6.

[Bol79] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979. doi:10.1109/TASSP.1979.1163209.

[CLM+24] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. EAT: Self-supervised pre-training with efficient audio transformer. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3807–3815, 2024. arXiv:2401.03497, doi:10.24963/ijcai.2024/421.

[CWW+23] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 4672–4712, 2023. arXiv:2212.09058.

[DCA25] DCASE Community. DCASE 2025 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring – Results. https://dcase.community/challenge2025/task-first-shot-unsupervised-anomalous-sound-detection-for-machine-condition-monitoring-results, 2025. Accessed 2026-06-22.

[DIH+22] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Takashi Endo, Masaaki Yamamoto, and Yohei Kawaguchi. Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2022. arXiv:2206.05876. doi:10.48550/arXiv.2206.05876.

[DIH+23] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei Kawaguchi. Description and Discussion on DCASE 2023 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. arXiv preprint arXiv:2305.07828, 2023. URL: https://arxiv.org/abs/2305.07828.

[DWY+24] Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. CED: Consistent ensemble distillation for audio tagging. In *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 291–295, 2024. arXiv:2308.11957.

[Efr79] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi:10.1214/aos/1176344552.

[GBW01] Sharon Gannot, David Burshtein, and Ehud Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626, 2001. doi:10.1109/78.934132.

[GCG21] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. doi:10.21437/Interspeech.2021-698.

[HJZ+25] Bing Han, Anbai Jiang, Xinhu Zheng, Wei-Qiang Zhang, Jia Liu, Pingyi Fan, and Yanmin Qian. Exploring Self-Supervised Audio Models for Generalized Anomalous Sound Detection. arXiv preprint arXiv:2508.12230, 2025. URL: https://arxiv.org/abs/2508.12230.

[KHDM98] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. doi:10.1109/34.667881.

[KIK+21] Yohei Kawaguchi, Keisuke Imoto, Yuma Koizumi, Noboru Harada, Daisuke Niizumi, Kota Dohi, Ryo Tanabe, Harsh Purohit, and Takashi Endo.

Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain-shifted conditions. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021. arXiv:2106.04492. doi:10.48550/arXiv.2106.04492.

[KKI+20] Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiki Nakamura, Yuki Nikaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, and Noboru Harada. Description and discussion on DCASE 2020 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020. arXiv:2006.05822. doi:10.48550/arXiv.2006.05822.

[KSU+19a] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 313–317, 2019. doi:10.1109/WASPAA.2019.8937164.

[KSU+19b] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. Unsupervised detection of anomalous sound based on deep learning and the Neyman–Pearson lemma. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):212–224, 2019. doi:10.1109/TASLP.2018.2877258.

[Kun04] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[LO79] Jae S. Lim and Alan V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979. doi:10.1109/PROC.1979.11540.

[Mah36] Prasanta Chandra Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

[McC89] Donna Katzman McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989. doi:10.1177/0272989X8900900307.

[NHN+24] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi. Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring. In *Proc. Detection*

and Classification of Acoustic Scenes and Events Workshop (DCASE), 2024. arXiv:2406.07250. doi:10.48550/arXiv.2406.07250.

[NHN+25] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi. Description and Discussion on DCASE 2025 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. arXiv preprint arXiv:2506.10097, 2025. URL: https://arxiv.org/abs/2506.10097.

[NHT+26] Tomoya Nishida, Noboru Harada, Daiki Takeuchi, Daisuke Niizumi, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi. Description and Discussion on DCASE 2026 Challenge Task 2: Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring, 2026. URL: https://arxiv.org/abs/2606.01578, arXiv:2606.01578.

[NTO+24] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, Masahiro Yasuda, Shunsuke Tsubaki, and Keisuke Imoto. M2D-CLAP: Masked modeling duo meets CLAP for learning general-purpose audio-language representation. In *Proc. Interspeech 2024*, pages 57–61, 2024. arXiv:2406.02032, doi:10.21437/Interspeech.2024-29.

[Nun21] Eduardo C. Nunes. Anomalous Sound Detection with Machine Learning: A Systematic Review. arXiv preprint arXiv:2102.07820, 2021. URL: https://arxiv.org/abs/2102.07820.

[PTI+19] Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 209–213, 2019. doi:10.33682/m76f-d618.

[RPZ+22] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, 2022.

[RRS00] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 427–438, 2000. doi:10.1145/342009.335437.

[SNP+20] Kaori Suefusa, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei Kawaguchi. Anomalous sound detection based on interpolation deep neural network. In *ICASSP 2020 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 271–275, 2020. doi:10.1109/ICASSP40776.2020.9054344.

[SS25] Phurich Saengthong and Takahiro Shinozaki. Deep Generic Representations for Domain-Generalized Anomalous Sound Detection. arXiv preprint arXiv:2409.05035; DCASE 2025 Challenge Task 2 Technical Report, 2025. URL: https://arxiv.org/abs/2409.05035.

[TCP22] Syed M. Tayyab, Simon Chatterton, and Paolo Pennacchi. Image-processing-based intelligent defect diagnosis of rolling element bearings using spectrogram images. *Machines*, 10(10):908, 2022. doi:10.3390/machines10100908.

[VVB88] Barry D. Van Veen and Kevin M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988. doi:10.1109/53.665.

[WFI+25] Kevin Wilkinghoff, Takuya Fujimura, Keisuke Imoto, Jonathan Le Roux, Zheng-Hua Tan, and Tomoki Toda. Handling Domain Shifts for Anomalous Sound Detection: A Review of DCASE-Related Work. Technical Report TR2025-157, Mitsubishi Electric Research Laboratories, 2025. URL: https://www.merl.com/publications/docs/TR2025-157.pdf.

[WGM+75] Bernard Widrow, John R. Glover, John M. McCool, John Kaunitz, Charles S. Williams, Robert H. Hearn, James R. Zeidler, Eugene Dong, and Robert C. Goodlin. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12):1692–1716, 1975. doi:10.1109/PROC.1975.10036.

[WLK+25] Ho-Hsiang Wu, Wei-Cheng Lin, Abinaya Kumar, Luca Bondi, Shabnam Ghaffarzadegan, and Juan Pablo Bello. Towards Few-Shot Training-Free Anomaly Sound Detection. In *Proc. Interspeech 2025*, pages 3384–3388, 2025. doi:10.21437/Interspeech.2025-467.

[XTL25] Haifeng Xu, Yizhou Tan, and Shengchen Li. Robust Anomaly Sound Detection via BSS-Augmented Pre-trained Model Fine-tuning. Tech. rep., DCASE2025 Challenge, 2025. URL: https://dcase.community/documents/challenge2025/technical_reports/DCASE2025_Li_47_t2.pdf.

[YPY25] Tian Ye, Tong Peng, and Lihua Yang. Review on sound-based industrial predictive maintenance: From feature engineering to deep learning. *Mathematics*, 13(11):1724, 2025. doi:10.3390/math13111724.

[ZTdCQT21] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. LEAF: A learnable frontend for audio classification. In *International Conference on Learning Representations (ICLR)*, 2021.