



Politecnico di Torino

Department of Mechanical and Aerospace Engineering

Master of Science in Mechanical Engineering

Machine Learning and Deep Learning Approaches for Anomaly Detection and Fault Diagnostics in Rotating Machinery

Supervisors:

Dr. Fahimeh Mashayekhi
Politecnico di Torino

Prof. Maryam Ghandchi Tehrani
University of Groningen

Candidate:

Mohammadkarim Najah

July 2026

Abstract

Rolling element bearings are among the most likely sources of unplanned failures, and the cost of vibration instrumentation constrains failures in electromechanical drive systems and their health monitoring systems. This thesis works on investigating whether the motor phase currents that already exist inside the PLC logs or every frequency inverter can be replaced, or they are just complementary for external accelerometers, and whether reliable fault detection is achievable without any labeled fault data. The Paderborn University Dataset has been used for carrying out all experiments and collected from the benchmark, which provides synchronously recorded vibration and two-phase motor current signals for 32 states: six healthy, twelve artificially damaged, and fourteen real-damaged bearing states under four operating conditions at different speeds, loads, and torques.

We have organized this study into two phases. First, 178-dimensional feature sets spanning six signal analysis domains (time domain, frequency domain, wavelet-packet energies, envelope spectra, motor-current signature sidebands, and cepstral rahmonics) are extracted from half second windows of both conditions, and four classical classifiers (SVM, k -NN, Random Forest, XGBoost) are trained under the different conditions: recording grouped cross-validation, artificial-only and real-only sub-domains, artificial to real transfer, cross operating condition testing, and leave one bearing out (LOBO) validation, respectively. For the second phase, a hybrid architecture is developed in which, at the first step, anomaly detectors trained solely on healthy data (Mahalanobis and k -NN distance, deterministic and variational autoencoders, and a temporal LSTM-VAE) are able to flag anomalous patterns. Then, assigning the fault type to detected windows by a compact classifier trained on only twelve labeled recordings per fault class.

We have obtained three important findings. First, the choice of evaluation condition has priority over the choice of model: pooled cross-validation leads to macro F1 scores up to 0.998, while LOBO evaluation of the same classifier drops to 0.57, which demonstrates that conventional protocols strongly overstate industrial performance. As the second finding, we can see that only motor current signals support anomaly detection: we achieved an AUROC of 0.98 using only current features with an autoencoder trained only on healthy data that separates faulty from healthy recordings, and AUROC of 0.998 when the same model is trained with both vibration and current features. Third, the anomaly score increases monotonically with the documented severity of the damage, showing that the detector output can act as

the preliminary condition indicator. We have also documented two negative aspects; raw 64 kHz waveforms overcome both reconstruction-based detection and few-shot CNN classification. Moreover, it has been indicated that the diversity of damage states plays a crucial role rather than damage severity, in terms of the performance for generalization.

Acknowledgements

I am profoundly grateful to my supervisors at the University of Groningen, Prof. Maryam Ghandchi Tehrani and Dr. Mahmoud Askari of the Dynamics and Vibration Research Unit. They welcomed me into their group and guided this work from beginning to end with their expertise, thoughtful discussions, and generous advice. Their supervision shaped both the direction of this research and my growth as an engineer, and this thesis would not have been possible without their continuous support.

I would also like to express my sincere gratitude to my supervisor at Politecnico di Torino, Dr. Fahimeh Mashayekhi, for her constant guidance, insightful feedback, and encouragement throughout every stage of this thesis.

My sincere thanks go as well to Dr. Bahman Askari of Vanderlande Group in the Netherlands, who acted as one of my supervisors. His excellent advice and guidance throughout this project were invaluable, and I am deeply grateful for the time and knowledge he so generously shared with me.

Finally, my heartfelt gratitude goes to my wife, Bahar, for her unconditional love, patience, and encouragement throughout my studies.

Declaration on the Use of Artificial Intelligence Tools

In accordance with the guidelines of Politecnico di Torino on the transparent use of generative artificial intelligence in theses and student works, the author declares that AI-based tools were used during the preparation of this thesis, as specified below.

- **Language and writing support.** Generative AI assistants (Anthropic Claude and OpenAI ChatGPT) were used to improve the grammar, clarity, and style of the English text and to assist with the $\text{\LaTeX}$ formatting of the document.
- **Programming support.** AI coding assistants were used as a support tool during the implementation of the analysis code, for example for debugging, code review, and the generation of plotting routines.

No AI tool was used to generate the research idea, the methodology, the experimental design, or the scientific conclusions of this work. All experiments, data analyses, and interpretations of the results were conceived, executed, and verified by the author, who critically reviewed every AI-assisted output and assumes full responsibility for the entire content of this thesis.

Contents

Acknowledgements	i
Declaration on the Use of Artificial Intelligence Tools	ii
1 Introduction	1
1.1 Industrial and Economic Motivation	1
1.2 Condition Monitoring: From Vibration to Motor Currents	2
1.3 Data-Driven Methods and the Role of Benchmark Datasets	3
1.4 Identified Research Gaps	3
1.5 Research Objectives	4
1.6 Thesis Scope and Limitations	5
1.7 Thesis Organisation	5
2 Literature Review	6
2.1 Overview and Scope	6
2.2 Bearing Condition Monitoring: Foundations	6
2.3 The Paderborn University Bearing Dataset	7
2.4 Signal Processing for Bearing Diagnostics	8
2.4.1 Vibration Analysis and Envelope Demodulation	8
2.4.2 Motor Current Signature Analysis	8
2.5 Feature Engineering and Interpretability	9
2.6 Classical Machine Learning Classifiers	9
2.7 Deep Learning for Bearing Fault Detection	10
2.8 Unsupervised Anomaly Detection	11
2.8.1 Problem Framing	11
2.8.2 One-Class Methods	11
2.9 Evaluation Protocols and Domain Adaptation	11
2.10 Research Gaps and Positioning of This Thesis	12

3	Methodology	13
3.1	Dataset Description: The Paderborn KAt Bearing Dataset	13
3.1.1	Selection Rationale	13
3.1.2	Bearing Type and Geometry	13
3.1.3	Bearing Damage States	14
3.2	Experimental Setup and Operating Conditions	17
3.2.1	Test Rig and Instrumentation	17
3.2.2	Operating Conditions	18
3.3	Exploratory Data Analysis	18
3.3.1	Raw Signal Behaviour	19
3.3.2	Spectral Analysis	19
3.3.3	Motor Current Spectral Signatures	22
3.3.4	Time–Frequency and Cepstral Views	25
3.3.5	Statistical Summary and Dataset-Wide Views	25
3.3.6	The Operating-Condition Effect	31
3.3.7	Implications for the Methodology	31
3.4	Preprocessing and Segmentation	34
3.4.1	Windowing	34
3.4.2	Leakage Control	35
3.4.3	Per-Recording Physical Parameters	35
3.5	Feature Engineering	35
3.5.1	Time-Domain Features	35
3.5.2	Frequency-Domain Features	37
3.5.3	Time–Frequency Features: Wavelet Packet Energies	37
3.5.4	Envelope Features	38
3.5.5	MCSA Features	38
3.5.6	Cepstral Features	39
3.5.7	Feature Selection	39
3.6	Phase 1: Supervised Baselines and Evaluation Protocols	40
3.6.1	Model Set	40
3.6.2	Evaluation Metrics	45
3.6.3	Evaluation Protocol Suite	45
3.7	Phase 2: Hybrid Anomaly Detection and Few-Shot Diagnosis	46
3.7.1	Rationale and Architecture	46
3.7.2	Input Representation: A Negative Result First	46
3.7.3	Stage-1 Detector Set	47
3.7.4	Thresholding and Calibration	48

3.7.5	Stage 2: Few-Shot Fault Classification	48
3.7.6	Extended Analyses	49
3.7.7	Implementation	49
4	Results and Discussion	50
4.1	Supervised Baselines under the Standard Protocol	50
4.2	What the Models Learn: Feature Importance and SHAP	52
4.3	Damage-Domain Scenarios and Transfer	55
4.4	Cross-Condition Generalisation	57
4.5	Leave-One-Bearing-Out: The Evaluation-Protocol Effect	60
4.6	Stage 1: Healthy-Only Anomaly Detection	62
4.7	The Two-Stage System and Threshold Calibration	67
4.8	Severity-Aware Evaluation	69
4.9	Label-Budget Analysis	71
4.10	Representation Study: Two Negative Results	72
4.11	Robustness to Random Initialisation	73
4.12	Comparison with Published Results	73
4.13	Synthesis	74
5	Conclusions and Future Work	75
5.1	Summary of the Work	75
5.2	Principal Findings	75
5.3	Contributions	76
5.4	Limitations	77
5.5	Future Work	77

List of Figures

3.1	Schematic of the Paderborn KAt modular test rig. Modules: (1) drive motor (PMSM, 425 W); (2) torque-measurement shaft; (3) rolling-bearing test module with accelerometer adapter; (4) flywheel; (5) load motor (PMSM, 1.7 kW). Adapted from [1].	17
3.2	Raw vibration overlay (100 ms excerpt) for the six health states at the reference operating condition. Artificial defects show periodic impulses; real damage is visually indistinguishable from healthy. . . .	19
3.3	Per-class vibration signals with RMS and kurtosis annotations. The kurtosis of the artificial defects (11.58 and 11.76 for outer and inner race) collapses to 2.61 and 2.20 for their naturally damaged counterparts.	20
3.4	Outer-race comparison: artificial defect ($\text{RMS} = 0.539 \text{ g}$, $K = 11.58$) versus naturally developed damage ($\text{RMS} = 0.477 \text{ g}$, $K = 2.61$). . . .	21
3.5	Inner-race comparison: artificial defect ($\text{RMS} = 0.687 \text{ g}$, $K = 11.76$) versus naturally developed damage ($\text{RMS} = 0.352 \text{ g}$, $K = 2.20$). The real fault sits below the healthy baseline in both statistics.	21
3.6	Raw phase current I_a overlay (100 ms). All health states produce visually identical waveforms.	21
3.7	Raw phase current I_b overlay (100 ms). As for I_a , no class differences are visible in the time domain.	21
3.8	Welch PSD (0–10 kHz, log scale) per health state. Red dashed lines: BPFO harmonics; blue dashed lines: BPFI harmonics. All faulty states elevate the 2–5 kHz resonance band.	22
3.9	PSD overlays against the healthy reference. Artificial defects raise the spectrum by more than an order of magnitude at defect-frequency harmonics; real damage produces only subtle broadband changes. . .	23
3.10	Full-record FFT spectra ($\Delta f = 0.25 \text{ Hz}$). Red dashed: BPFO harmonics; blue dashed: BPFI harmonics.	23

3.11	Envelope spectra (band-pass 2–10 kHz, Hilbert transform). BPFO peak amplitudes: OR_Artificial ≈ 0.30 , OR_Real ≈ 0.11 ; the real inner-race fault yields only ≈ 0.065 at BPFI.	24
3.12	Phase-current I_a spectra per health state (log scale) with the data-estimated electrical fundamental f_e and the theoretical sideband positions $f_e \pm k \cdot \text{BPFO}/\text{BPFI}$. Fault components are orders of magnitude below the fundamental.	26
3.13	Envelope spectra of the demodulated phase current I_a per health state. Bearing-induced modulation appears as lines directly at the defect frequencies (dashed markers).	27
3.14	Short-time Fourier spectrograms of the vibration signal (first second) per health state. Faulty states show periodic striations across the excited resonance bands.	28
3.15	Continuous wavelet scalograms of the vibration signal (200 ms) per health state. Artificial defects produce periodic impulse ridges; real damage produces diffuse texture.	29
3.16	Real cepstra of the vibration signal with the expected rahmonic positions k/BPFO and k/BPFI marked. Harmonic periodicity is clearly detectable only for the artificial defects.	30
3.17	Dataset composition: recordings per class (4- and 6-class views) and the balanced class $\times$ operating-condition design.	31
3.18	Distributions of RMS and kurtosis per health state over all recordings and operating conditions. Vibration separates the artificial subclasses only; the current statistics are indistinguishable across all states.	32
3.19	All 2,559 recordings in the vibration RMS-kurtosis plane. Artificial damage (triangles) separates cleanly; naturally developed damage (squares) overlaps the healthy cluster (circles).	32
3.20	One naturally damaged bearing (KA04) measured under the four operating conditions. The condition changes the signal almost as much as the damage does.	33
3.21	Mean vibration RMS per health state (rows) and operating condition (columns) over the whole dataset. Condition-induced variation rivals fault-induced variation.	33
3.22	Top-30 features by mutual information with the 4-class label. Envelope and defect-frequency amplitudes dominate the vibration side; demodulated-current features carry the current side.	41

3.23	Pearson correlation matrix of the top-25 features by mutual information. Harmonic families of the same defect frequency form strongly correlated blocks, motivating the redundancy filter.	42
3.24	Class-conditional distributions of the six features with the highest mutual information. Each feature separates some class pairs but none separates all, which is why the classifiers operate on the combined selection.	43
3.25	PCA (top) and t-SNE (bottom) embeddings of the feature space for vibration-only, current-only, and fused subsets, coloured by class. Class structure is visible before any supervised training.	43
4.1	Accuracy and macro F1 for all model and modality combinations (mean $\pm$ std over five grouped folds). The modality ranking fused, then vibration, then current holds for every model.	51
4.2	Row-normalised confusion matrices on the held-out fold for all model and modality combinations. The Combined class is the dominant error source, most visibly for the current modality.	53
4.3	Per-class F1 of the best model per modality on the held-out fold. Fusion mainly repairs the Combined class of the current channel. . . .	54
4.4	Top-15 XGBoost feature importances per modality (blue: vibration features; red: current features). The models select the physically expected envelope and sideband features.	54
4.5	Mean absolute SHAP values of the fused XGBoost model. The attribution confirms the gain-based ranking: envelope and defect-frequency features carry the decisions.	55
4.6	Best-model macro F1 across the five damage-domain scenarios. Within-domain diagnosis is nearly saturated; cross-domain transfer collapses for every modality.	56
4.7	Macro F1 of all four models and three modalities in the artificial-only, real-only, and pooled scenarios (mean $\pm$ std over grouped folds). The within-domain conclusions do not depend on the choice of model. . .	57
4.8	Row-normalised confusion matrices of the cross-domain transfer experiments (best model per direction). The diagnosis of naturally developed inner-race and combined damage breaks down after artificial-only training.	58
4.9	Cross-condition accuracy heatmaps per modality (XGBoost). Speed changes (first row and column) are the hardest shift for both modalities; torque changes additionally cripple the current channel.	58

4.10	LOBO recall for each held-out bearing (colour = true class). Failures concentrate on specific individuals: KA30, KB27, and KA15, the plastic-deformation indentation bearings, and the healthy bearing K002 are diagnosed at or near zero recall.	61
4.11	Pooled confusion matrices over all 32 LOBO folds per modality (XGBoost, row-normalised). The Combined class scatters into the single-fault classes once its bearings are unseen.	62
4.12	Receiver operating characteristics of all Stage-1 detectors and modalities at window level. Line style encodes the detector, colour the modality.	63
4.13	Anomaly-score distributions of the healthy test set (green) and faulty test set (red) per modality, with the 99th-percentile threshold. Fusion produces the cleanest separation.	64
4.14	PCA projection of the fused VAE latent means, coloured by the (never seen) 6-class label. The healthy cluster separates, and fault sub-types organise without supervision.	65
4.15	Window-level versus recording-level AUROC for all detectors and modalities. The 4s aggregated decision is uniformly stronger.	65
4.16	Detection rate of the fixed-threshold detector per fault sub-class and per operating condition. Naturally damaged inner-race bearings and the reduced-speed condition are the weakest spots.	66
4.17	Stage-1 AUROC under leave-one-healthy-bearing-out (mean $\pm$ std over six folds). Healthy-bearing individuality widens the uncertainty substantially, especially for the current modality.	66
4.18	End-to-end macro F1 versus the Stage-1 threshold percentile, at window level (left) and recording level with majority voting (right). Stars mark the calibrated operating points; the dashed line marks the conventional p99 rule.	67
4.19	Detection rate versus false-alarm rate along the threshold sweep, per modality. Stars mark the calibrated operating points.	68
4.20	Row-normalised confusion matrices of the end-to-end two-stage system per modality (MLP Stage 2). Residual errors concentrate in the Combined class and in faulty windows suppressed by Stage 1.	69
4.21	Anomaly-score distributions per extent-of-damage level (level 0 = healthy) and modality. The vibration and fused scores shift upward monotonically with the level.	70

4.22	Supervised LOBO recall versus extent-of-damage level (fused XG-Boost, one marker per bearing). Failures do not follow severity; they cluster on the plastic-deformation bearings KA30, KB27, and KA15. .	70
4.23	Diagnosis quality versus label budget: macro F1 of the Stage-2 classifier alone (left) and of the end-to-end system (right) for $K \in \{1, 2, 4, 8, 12\}$ labelled recordings per class (mean $\pm$ std over three draws).	71
4.24	Seed-to-seed spread of the Stage-1 detectors (left, AUROC) and of the Stage-2 and end-to-end systems (right, macro F1) over five random seeds.	73

List of Tables

3.1	Geometric and rating parameters of the type 6203 ball bearing [1]. . .	14
3.2	Operating history of the healthy reference bearings [1].	15
3.3	Artificially damaged bearings: location, extent level, and preparation method [1]. OR: outer ring; IR: inner ring.	15
3.4	Damage extent levels defined by the damaged-surface length as a fraction of the pitch circumference [1].	16
3.5	Bearings with naturally developed damage [1]. S: single; R: repetitive; M: multiple damage.	16
3.6	Composition of the dataset used in this work: bearing states and 4 s recordings per class (4 operating conditions $\times$ 20 repetitions each). .	17
3.7	Instrumentation of the Paderborn test rig [1].	18
3.8	Operating conditions of the Paderborn experiments [1].	18
3.9	Time-domain statistics of the six representative recordings (reference condition; 256,000 samples). CF = crest factor.	28
3.10	Multi-domain feature set extracted per 0.5 s window (68 vibration + 110 motor-current features).	36
4.1	Supervised 4-class diagnosis, all operating conditions pooled: macro F1 (mean $\pm$ std over 5 recording-grouped CV folds).	51
4.2	Damage-domain scenario study: best-model macro F1 per scenario and modality (recording-grouped splits).	56
4.3	Cross-condition generalisation (XGBoost): accuracy when training on one operating condition (rows) and testing on another (columns). .	59
4.4	Effect of the evaluation protocol on supervised diagnosis (XGBoost, macro F1): recording-grouped cross-validation versus leave-one-bearing-out (LOBO). The gap quantifies bearing-identity inflation.	60
4.5	Stage-1 anomaly detection, window level (0.5 s): AUROC per detector and modality with healthy-only training. Deep models are reported as mean $\pm$ std over 5 seeds.	63

4.6	Stage-1 anomaly detection, recording level (4s decision): AUROC per detector and modality.	63
4.7	Stage-1 detectors under leave-one-healthy-bearing-out: AUROC (mean $\pm$ std over 6 folds).	65
4.8	Few-shot Stage-2 diagnosis and end-to-end two-stage system: macro F1 (mean $\pm$ std over 5 seeds; 12 labelled recordings per fault class; default p99 threshold).	67
4.9	Stage-1 threshold calibration: operating points selected on the calibration set (healthy-validation plus few-shot windows; no test labels used) and the resulting end-to-end macro F1.	67
4.10	Severity-aware evaluation of the Stage-1 detector (Dense-AE, calibrated threshold): detection rate per extent-of-damage level and rank correlation between anomaly score and severity.	69
4.11	Label budget: Stage-2 MLP macro F1 versus the number of labelled recordings per fault class (mean over 3 random draws; identical test set for every K).	71
4.12	Representation study (fused modality, best configuration of each approach): raw-waveform deep models versus their feature-space equivalents.	72
4.13	Published results on the Paderborn benchmark and the corresponding figures of this thesis. Settings differ in class definitions and split rules; comparisons are indicative.	74

Chapter 1

Introduction

1.1 Industrial and Economic Motivation

Electromechanical drive systems form the foundation of modern manufacturing, process, and energy infrastructure. From compact servo drives in automotive assembly lines to multi-megawatt motors in cement mills and wind turbines, the uninterrupted operation of rotating machinery directly governs production throughput, energy consumption, and workplace safety. Within these systems, rolling-element bearings occupy a uniquely critical role. They are simultaneously among the most highly loaded and most wear-prone components, and their degradation is the single most frequent root cause of motor and gearbox failures. Industry surveys consistently attribute between 40% and 70% of unplanned drive-system breakdowns to bearing-related faults [2], a range that spans fractional-horsepower appliance motors through to the large-frame induction machines of petrochemical plants.

Maintenance strategy determines how much of this failure burden actually materialises as cost. Purely reactive maintenance waits for the failure and accepts the full secondary damage and downtime. Preventive maintenance replaces components on a fixed schedule, which avoids some failures but discards useful component life and does not protect against early failures within the interval. Predictive maintenance, the strategy this thesis serves, monitors the actual condition of the component and triggers intervention only when objectively warranted. The financial consequences of a bearing failure extend well beyond the price of the component. Unplanned downtime causes direct losses through halted production, idle labour, and expedited logistics, and indirect losses through quality deviations when a process is interrupted mid-cycle. A widely cited McKinsey Global Institute analysis estimated the annual economic toll of unplanned equipment failures in European manufacturing at more than \$240 billion, with drive systems representing a disproportionate share [3]. The

same analysis projects downtime reductions of 30 to 50 per cent and maintenance-cost reductions of 10 to 40 per cent for operators who adopt predictive maintenance at scale [3].

The practical bottleneck of predictive maintenance is not the decision logic but the sensing layer. Any condition-based strategy needs a signal that reflects component health, an algorithm that converts that signal into a health assessment, and enough confidence in the assessment to act on it. The present work addresses the first two elements for the most failure-prone component of the drive train, with an explicit focus on sensing that costs nothing to add.

1.2 Condition Monitoring: From Vibration to Motor Currents

For several decades the dominant sensing technology for bearing condition monitoring has been the piezoelectric accelerometer. Vibration measured at the bearing housing contains the mechanical impulse responses generated when rolling elements traverse surface defects, and mature signal processing methods such as envelope analysis, spectral kurtosis, and cepstral analysis exist to convert these impulses into actionable indicators [4, 5]. For large, high-value assets the cost of permanent accelerometer installations is easily justified. The economics change fundamentally, however, for the far larger population of small industrial motors that power pumps, fans, conveyors, and compressors. For machines rated below a few kilowatts, the cost of an accelerometer, its cabling, and the signal-conditioning chain can approach the replacement cost of the motor itself, which is why the overwhelming majority of these machines run without any condition monitoring at all.

Motor Current Signature Analysis (MCSA) removes this economic barrier by exploiting the stator phase currents that every frequency inverter already measures for control and protection purposes. Mechanical disturbances in the drive train modulate the electromagnetic air-gap torque and imprint characteristic sidebands onto the current spectrum [6, 7]. Because no additional hardware is required, MCSA is inherently deployable at fleet scale, and a monitoring function based on it amounts to a software upgrade of the existing drive. Its technical difficulty is equally well documented. The fault-induced sidebands are typically two to three orders of magnitude weaker than the supply-frequency component, and inverter switching noise further degrades the signal-to-noise ratio [1]. Whether current-based monitoring can approach the reliability of vibration-based monitoring is therefore not a question that can be settled by physical argument alone. It requires a controlled, like-for-like

empirical comparison, and providing that comparison is the central experimental theme of this thesis.

1.3 Data-Driven Methods and the Role of Benchmark Datasets

Affordable sensing, high-capacity data acquisition, and modern machine learning have shifted condition monitoring from physics-based models towards data-driven approaches that learn the statistical regularities of healthy and faulty operation directly from measurements [8, 9]. Data-driven pipelines are flexible, since they adapt to the signal characteristics of a specific installation, and scalable, since one algorithmic pipeline can serve a heterogeneous fleet [10]. Their weakness, repeatedly documented in the review literature, is that the reported performance depends strongly on how the evaluation is designed, and that models can silently exploit information that will not be available in deployment.

Fair scientific progress in this area depends on public benchmark datasets. The Case Western Reserve University (CWRU) dataset served the vibration-diagnostics community for two decades, but it contains only artificially seeded single-point defects, offers no current signals, and its common evaluation practice has been shown to overstate model performance [11]. The Paderborn University KAt-DataCenter dataset [1] was designed to overcome exactly these limitations. It provides synchronously recorded vibration and motor current signals, includes both artificially prepared and naturally developed damage, and documents each damaged bearing with a systematic morphological classification. It is the primary experimental platform of this thesis.

1.4 Identified Research Gaps

The literature review in Chapter 2 identifies four persistent gaps that this thesis addresses.

Like-for-like modality comparison. The diagnostic value of vibration versus motor current features has rarely been assessed with an identical feature-extraction pipeline, identical classifiers, and identical evaluation protocols on the same data. Most studies commit to one modality from the outset; the few comparative studies use separate pipelines per modality [1, 12].

Protocol-dependent performance inflation. Recent work shows that random or recording-level train-test splits allow classifiers to exploit bearing-specific signatures rather than fault physics [10,11]. The magnitude of this inflation on the Paderborn benchmark, as a function of signal modality and of the detection paradigm (supervised versus one-class), has not been quantified systematically.

Label-free detection from internal signals. One-class (healthy-only) anomaly detection on Paderborn data has been explored almost exclusively for the vibration channel. Whether bearing damage is detectable from motor current features alone, using detectors that never see a fault example, remains essentially unreported.

Severity- and label-budget-aware evaluation. The Paderborn documentation assigns every damaged bearing an extent-of-damage level and a damage-morphology class, yet published studies rarely use these labels. Likewise, hybrid detector-plus-classifier architectures are usually evaluated at a single, fixed label budget, leaving open how diagnostic quality scales with the number of labelled fault examples.

1.5 Research Objectives

The thesis pursues five operational objectives:

1. To characterise the vibration and current signals of the Paderborn dataset through an exploratory data analysis that documents the statistical and spectral differences between health states, damage origins, and operating conditions, and motivates every subsequent processing decision.
2. To construct a multi-domain feature set from both modalities, consisting of time statistics, spectral descriptors, wavelet-packet energies, envelope spectrum amplitudes at the bearing characteristic frequencies, MCSA sideband amplitudes, and cepstral harmonics, extracted with identical windowing for both channels.
3. To benchmark four classical classifiers (SVM, k -NN, Random Forest, XGBoost) for vibration-only, current-only, and fused feature sets under a graded suite of evaluation protocols: pooled recording-grouped cross-validation, artificial-only and real-only sub-domains, bidirectional artificial-to-real transfer, cross-condition testing, and leave-one-bearing-out validation.

4. To develop and evaluate a two-stage hybrid architecture in which detectors trained exclusively on healthy data (distance-based baselines, autoencoders, a variational autoencoder, and a temporal LSTM-VAE) flag anomalous windows, and a compact classifier trained on a small pool of labelled recordings assigns the fault type to the flagged windows.
5. To extend the evaluation along two axes rarely reported for this benchmark: the relation between the anomaly score and the documented damage-extent level, and the scaling of diagnostic quality with the number of labelled fault recordings.

1.6 Thesis Scope and Limitations

All analyses concern the bearing-fault detection and diagnosis task as defined by the Paderborn dataset, using the type 6203 deep-groove ball bearing under four prescribed operating conditions. Remaining-useful-life estimation and degradation-trajectory modelling are outside the scope, although the severity analysis in Chapter 4 provides a first step in that direction. No real-time hardware implementation is presented; all results derive from offline analysis of the recorded data. Deep models were trained on CPU hardware, which constrained architecture sizes to configurations trainable within hours; this constraint is made explicit where relevant.

1.7 Thesis Organisation

Chapter 2 reviews the relevant literature, from classical signal processing through classical and deep machine learning to one-class anomaly detection and domain adaptation. Chapter 3 describes the dataset, the exploratory data analysis, and the complete two-phase methodology, including the evaluation protocols. Chapter 4 presents and discusses the experimental results. Chapter 5 summarises the findings, states the contributions and limitations, and outlines future work.

Chapter 2

Literature Review

2.1 Overview and Scope

Research on data-driven bearing condition monitoring has evolved over two decades from hand-tuned spectral thresholds to end-to-end deep networks. This chapter surveys the strands of that literature on which the present work builds: the failure statistics and review literature that motivate the problem (Section 2.2); the Paderborn benchmark itself (Section 2.3); the signal-processing foundations for vibration and current analysis (Section 2.4); feature engineering and interpretability (Section 2.5); classical classifiers (Section 2.6); deep architectures (Section 2.7); one-class anomaly detection (Section 2.8); and evaluation protocols and domain adaptation (Section 2.9). Each section closes with the connection to the specific design choices made in this thesis.

2.2 Bearing Condition Monitoring: Foundations

Bonnett and Yung analysed a large population of motor failures reported to the IEEE Pulp and Paper Industry Committee and attributed 40 to 70 per cent of electromechanical drive breakdowns to bearing faults, identifying fatigue pitting, false brinelling, and lubrication degradation as the dominant modes [2]. Their conclusion, that detection must occur before surface damage propagates to spalling, remains the primary quantitative justification for bearing condition monitoring. The same study emphasised a point that is easy to overlook in algorithm-centred research: the value of a monitoring system is set by its behaviour on the failure modes that actually occur in service, which are dominated by fatigue and contamination rather than by the clean geometric defects that are convenient to manufacture in the laboratory.

Lei et al. mapped the evolution of machine-learning-based fault diagnosis across

more than two hundred studies and identified three persistent obstacles to deployment: scarce labelled fault data, sensitivity of trained models to operating-condition changes, and the absence of standardised evaluation protocols [8]. Singh et al. reached a sharper conclusion in their systematic review of bearing prognostics: the dominant practice of training and testing on data from the same experimental run substantially overstates generalisation, and leave-bearing-out or cross-condition protocols should be the minimum standard for honest reporting [10]. Both recommendations are implemented literally in this thesis, whose central methodological theme is the comparison of one model family across protocols of increasing strictness. For the CWRU benchmark, Smith and Randall demonstrated the same point from the signal-processing side, showing that several published claims dissolve once the evaluation is performed diagnostically rather than by pattern memorisation [11]. Their benchmark study is the closest methodological precedent for the protocol comparison carried out here, although it concerns a different dataset and does not cover the current modality or the one-class setting.

2.3 The Paderborn University Bearing Dataset

Lessmeier et al. introduced the KAt-DataCenter dataset to remedy two specific weaknesses of earlier public data: the absence of motor current signals and the exclusive reliance on artificial single-point defects [1]. Their modular test rig records housing vibration and two phase currents synchronously at 64 kHz for 32 bearing states, of which six are healthy, twelve artificially damaged, and fourteen naturally damaged through accelerated lifetime testing, under four operating conditions. Beyond the data themselves, the paper contributes a morphological damage taxonomy covering extent levels, single-point versus distributed character, and damage combination, and defines evaluation settings in which classifiers trained on artificial damage are tested on real damage.

Using seven classifiers, the authors reported that current-based classification reached 60.9% accuracy under the artificial-to-real setting against 75.0% for vibration-based ensembles, establishing both the feasibility and the difficulty of current-based diagnosis [1]. These figures are the historical reference points for the transfer experiments in Chapter 4. More recently, Tahan et al. pushed the artificial-to-real accuracy to 72.7% with a multi-branch network combining supervised, semi-supervised, and unsupervised objectives [13], which constitutes the strongest published baseline for that setting known to the author. At the other end of the difficulty spectrum, Firdausi and Ahmad reported detection rates above 99% for a small CNN evaluated

within a single damage domain [14]. The spread between these published numbers, from roughly 61% to above 99% on the same physical dataset, is itself evidence for the central claim of this thesis, namely that the evaluation protocol is a first-order determinant of the reported result.

2.4 Signal Processing for Bearing Diagnostics

2.4.1 Vibration Analysis and Envelope Demodulation

Randall’s monograph provides the standard treatment of vibration-based condition monitoring [4]: the kinematic derivation of the bearing characteristic frequencies, the mechanism by which defect impacts excite structural resonances, and envelope analysis (the high-frequency resonance technique) as the canonical demodulation method. The physical picture is worth stating because it explains several empirical results of this thesis. A localised defect generates a short impact each time a rolling element passes over it. The impact itself is far too short to be observed directly, but it excites the structural resonances of the bearing and housing, typically in the kilohertz range, and these resonances ring down between impacts. The measured vibration is therefore an amplitude-modulated carrier: the carrier is the resonance, and the modulation rate is the defect frequency. Demodulating the resonance band recovers the modulation and makes the defect frequency directly visible.

Randall, Antoni, and Chobsaard connected envelope analysis to the broader cyclostationary framework, showing that envelope-spectrum peaks at defect frequencies are manifestations of second-order cyclostationarity and that demodulation-band selection governs detection sensitivity [15]. Antoni’s spectral kurtosis [5] formalised the choice of that band as the region of maximal impulsive activity. In this thesis the demodulation band of 2 to 10 kHz is selected from the power-spectral-density analysis of the exploratory study, which identifies this region as the resonance band excited by defects on the Paderborn rig. Wavelet-based time-frequency analysis, reviewed by Peng and Chu [16], complements envelope analysis for non-stationary content and motivates the wavelet-packet energy features adopted here.

2.4.2 Motor Current Signature Analysis

Schoen et al. first demonstrated that bearing defects imprint measurable components onto the stator current at predictable combinations of the electrical supply frequency and the mechanical characteristic frequencies [7]. The physical mechanism parallels the vibration case: the defect produces a periodic torque disturbance,

the motor responds with a small speed and air-gap variation, and the stator current is amplitude- and phase-modulated at the disturbance rate. Benbouzid’s review consolidated the theory for the full range of induction-machine faults and stated the practical obstacle plainly: bearing-related sidebands lie two to three orders of magnitude below the fundamental, demanding high spectral resolution and careful signal conditioning [6].

Two consequences for the present work follow directly. First, window lengths must be long enough to resolve sidebands spaced tens of hertz from the fundamental, a requirement that fixes the 0.5 s analysis window used throughout. Second, on an inverter-fed permanent-magnet machine the electrical fundamental follows the commanded speed rather than the 50 Hz grid, so the fundamental is estimated from the spectrum of each recording rather than assumed.

2.5 Feature Engineering and Interpretability

Goyal et al. catalogued the time-domain statistical descriptors used in machinery diagnostics, including RMS, kurtosis, crest, shape and impulse factors, and analysed their sensitivity to different fault morphologies [17]. A key insight of that survey, confirmed empirically in Chapter 3, is that kurtosis responds strongly to isolated impulsive defects but only weakly to distributed roughness, which raises the broadband floor without producing heavy-tailed amplitude distributions. Multi-domain feature systems that combine time, frequency, and wavelet descriptors have repeatedly been shown to outperform any single domain. Wang and Wu paired such a system with XGBoost and SHAP-based attribution and reported that spectral entropy, RMS, and impulse factor dominated the model’s decisions [18]. The SHAP framework of Lundberg and Lee [19], which assigns each feature a game-theoretic contribution to each prediction, is used in this thesis to verify that the trained models rely on physically meaningful features rather than artefacts of the experimental setup.

2.6 Classical Machine Learning Classifiers

Jack and Nandi established support vector machines as the reference discriminative classifier for bearing diagnostics, showing that an RBF-kernel SVM outperformed comparably sized neural networks across defect severities and that genetic feature selection improved it further [20]. Breiman’s Random Forest [21] introduced variance reduction through decorrelated tree ensembles, a property that makes it robust for the moderate-sample, moderate-dimension regime typical of engineered fault

features. Chen and Guestrin’s XGBoost [22] added an explicitly regularised boosting objective and has become the strongest general-purpose learner for tabular data; in the diagnostics context it consistently ranks at or near the top of comparative studies [18]. These three models, together with the k -nearest-neighbour classifier as a non-parametric baseline, constitute the Phase-1 model set of this thesis. The selection deliberately spans margin-based, instance-based, bagging, and boosting paradigms, so that conclusions about modalities and protocols do not hinge on the idiosyncrasies of a single learning algorithm.

2.7 Deep Learning for Bearing Fault Detection

Convolutional networks applied to raw waveforms have achieved striking results when training and test data share the damage origin. Firdausi and Ahmad reported detection rates above 99% on Paderborn recordings with a two-layer 1D-CNN and no feature engineering at all [14]. Comprehensive reviews temper this picture: Zhang et al. surveyed deep architectures for bearing diagnostics and noted that reported accuracies degrade sharply under cross-domain evaluation and that comparisons against strong classical baselines are frequently absent [9]. This thesis contributes direct evidence on this point in the form of two documented negative results for raw-waveform deep models under realistic constraints.

The LSTM architecture of Hochreiter and Schmidhuber [23], designed to carry information across long sequences through gated cell states, underlies the sequence models used in this thesis. Sulaimon et al. applied an LSTM autoencoder to bearing monitoring in the one-class setting, training on healthy sequences only and thresholding the reconstruction error at a percentile of a healthy validation set [24], a threshold rule adopted directly in Phase 2 of this work. On the comparative question of modalities, Ayankoso et al. trained identical deep models on vibration and current data from a laboratory rig and found vibration individually stronger but fusion most robust [12]. Their study, conducted on a different test rig, is the closest published analogue to the controlled comparison performed here on the Paderborn benchmark.

2.8 Unsupervised Anomaly Detection

2.8.1 Problem Framing

Chandola, Banerjee, and Kumar’s survey distinguishes point, contextual, and collective anomalies [25]. Distributed bearing damage is a collective anomaly: no individual sample is extreme, but the joint statistical structure of the signal deviates from the healthy baseline. This framing argues for detectors that operate on rich statistical summaries of signal segments rather than on individual samples, and it anticipates the empirical finding of this thesis that detectors operating on engineered feature vectors succeed where raw-waveform reconstruction fails.

2.8.2 One-Class Methods

Schölkopf et al. introduced the one-class SVM, which estimates the support of the healthy distribution in a kernel feature space [26]. Liu, Ting, and Zhou proposed the Isolation Forest, which scores anomalies by the ease of isolating them under random axis-aligned partitioning [27]. Both established the one-class paradigm on which this thesis builds, although the detector set evaluated here favours reconstruction- and distance-based methods whose scores admit a direct probabilistic interpretation and whose failure modes are easier to diagnose. On the reconstruction side, the variational autoencoder of Kingma and Welling [28] learns a probabilistic latent representation whose evidence lower bound provides a principled anomaly score, and the β -weighted variant of Higgins et al. [29] controls the regularisation strength of the latent space. Nicchiotti et al. demonstrated the architectural template adopted in Phase 2: a one-class detector gates a downstream fault classifier, achieving 99.9% detection and 94.5% classification accuracy on turbomachinery data with healthy-only first-stage training [30].

2.9 Evaluation Protocols and Domain Adaptation

The generalisation gap between artificial and real damage is a special case of the domain-shift problem addressed by adversarial domain adaptation. Ganin et al.’s DANN learns representations that a domain discriminator cannot separate [31]. Long et al.’s CDAN conditions the discriminator on classifier predictions for a sharper alignment signal [32], and Zhu et al.’s CASSA framework, which adds structure-preserving projection and pseudo-labelling, reached 91.1% cross-domain accuracy on Paderborn data under class imbalance [33]. This thesis does not pursue adver-

sarial adaptation. It contributes the measurement side of the problem instead: a systematic quantification of how large the domain gaps actually are, per modality and per paradigm, under leakage-controlled protocols. Such measurements define the baseline that any adaptation method must beat, and Chapter 4 returns to the published adaptation results when the transfer experiments are discussed.

2.10 Research Gaps and Positioning of This Thesis

Four gaps emerge from the reviewed literature. First, no study reports a fully controlled vibration-versus-current comparison on the Paderborn benchmark with identical features, classifiers, and protocols for both modalities. Second, the inflation of reported performance caused by recording-level rather than bearing-level splitting has not been quantified on this dataset. Third, healthy-only anomaly detection from motor current features alone is essentially unreported. Fourth, the documented damage-extent levels and the label-budget dimension of hybrid architectures remain unused in published evaluations. The experimental programme of Chapters 3 and 4 addresses all four.

Chapter 3

Methodology

3.1 Dataset Description: The Paderborn KAt Bearing Dataset

3.1.1 Selection Rationale

All experiments use the Paderborn University KAt-DataCenter dataset [1]. Three properties make it uniquely suited to the research questions of this thesis. First, vibration and motor phase currents are recorded synchronously on the same rig under identical, documented operating conditions, enabling a like-for-like modality comparison. Second, the dataset contains both artificially prepared and naturally developed damage, so the artificial-to-real generalisation problem can be studied under controlled conditions. Third, every damaged bearing is documented morphologically, with damage mode, extent level, and single-point versus distributed character, which enables the severity-aware analysis of Chapter 4. Each measurement is stored as a MATLAB file whose name encodes the operating condition, bearing code, and repetition number (for example `N15_M07_F10_KA01_1.mat`).

3.1.2 Bearing Type and Geometry

All test specimens are single-row deep-groove ball bearings of type 6203 with the parameters of Table 3.1. Bearings from three manufacturers were used; their geometric differences are within 1 to 2 per cent, so the characteristic frequencies are practically identical across units [1].

The kinematic defect frequencies follow from this geometry. With shaft frequency f_r , ball diameter d , pitch diameter D , and contact angle ϕ , the ball-pass frequencies

Table 3.1: Geometric and rating parameters of the type 6203 ball bearing [1].

Parameter	Symbol	Value
Bearing type		Single-row deep-groove ball
Bearing designation		6203
Number of rolling elements	N_b	8
Inner raceway diameter	d_i	24.0 mm
Outer raceway diameter	d_o	33.1 mm
Pitch circle diameter	D	28.55 mm
Rolling element diameter	d	6.75 mm
Nominal contact angle	ϕ	0°
Static load rating	C_0	4,750 N
Dynamic load rating	C	9,500 N
Maximum speed		12,000 rpm

of the outer and inner race are

$$\text{BPFO} = \frac{N_b}{2} f_r \left(1 - \frac{d}{D} \cos \phi \right), \quad \text{BPFI} = \frac{N_b}{2} f_r \left(1 + \frac{d}{D} \cos \phi \right), \quad (3.1)$$

and the ball-spin and cage frequencies are

$$\text{BSF} = \frac{D}{2d} f_r \left(1 - \left(\frac{d}{D} \cos \phi \right)^2 \right), \quad \text{FTF} = \frac{f_r}{2} \left(1 - \frac{d}{D} \cos \phi \right). \quad (3.2)$$

At 1,500 rpm these evaluate to $\text{BPFO} = 76.4$ Hz, $\text{BPFI} = 123.6$ Hz, $\text{BSF} = 49.9$ Hz, and $\text{FTF} = 9.5$ Hz. Because all four scale linearly with speed, they are recomputed for every recording from the speed encoded in its filename, a detail that matters for the reduced-speed operating condition.

3.1.3 Bearing Damage States

The dataset comprises 32 bearing states: six healthy references, twelve artificially damaged, and fourteen naturally damaged bearings.

Healthy Reference Bearings

Six healthy bearings (K001 to K006) were run in for different durations and loads before measurement (Table 3.2). The differing run-in histories introduce a controlled degree of variability into the healthy class, a property that turns out to be consequential for the one-class experiments of Phase 2, where individual healthy bearings are shown to carry distinguishable signatures.

Table 3.2: Operating history of the healthy reference bearings [1].

Bearing Code	Run-in [h]	Radial Load [N]	Speed [rpm]
K001	>50	1,000–3,000	1,500–2,000
K002	19	3,000	2,900
K003	1	3,000	3,000
K004	5	3,000	3,000
K005	10	3,000	3,000
K006	16	3,000	2,900

Artificially Damaged Bearings

Twelve bearings carry controlled defects produced by three machining methods that differ systematically in morphology. Electrical discharge machining produces a sharply bounded trench of 0.25 mm length in the rolling direction. Drilling produces clean holes of 0.9 to 3 mm diameter. Manual electric engraving produces an irregular, shallower surface disturbance intended to resemble incipient pitting. All artificial damage is of the single-point type without repetition. Table 3.3 lists the full set, and Table 3.4 defines the extent levels used throughout this thesis.

Table 3.3: Artificially damaged bearings: location, extent level, and preparation method [1]. OR: outer ring; IR: inner ring.

Code	Component	Extent Level	Method
KA01	OR	1	EDM
KA03	OR	2	Electric engraver
KA05	OR	1	Electric engraver
KA06	OR	2	Electric engraver
KA07	OR	1	Drilling
KA08	OR	2	Drilling
KA09	OR	2	Drilling
KI01	IR	1	EDM
KI03	IR	1	Electric engraver
KI05	IR	1	Electric engraver
KI07	IR	2	Electric engraver
KI08	IR	2	Electric engraver

Naturally Damaged Bearings

Fourteen bearings were damaged in accelerated lifetime tests under elevated radial load and low-viscosity lubrication. Roughly 70% of the resulting damage is fatigue pitting; the remainder is plastic deformation in the form of indentations from de-

Table 3.4: Damage extent levels defined by the damaged-surface length as a fraction of the pitch circumference [1].

Damage Level	Percentage	Limit for 6203
1	0–2%	≤ 2 mm
2	2–10%	2–9 mm
3	10–50%	9–45 mm
4	50–100%	45–90 mm
5	>100%	>90 mm

bris particles. Table 3.5 summarises the set. Two properties deserve emphasis. First, several bearings exhibit multiple damage (M), in which symptoms combine across components; these produce superimposed signatures that are markedly harder to classify than clean single-point defects. Second, three bearings (KA30, KB24, KB27) carry distributed damage, the morphology most representative of industrial wear and, as Chapter 4 shows, the single most difficult category for every method evaluated in this thesis.

Table 3.5: Bearings with naturally developed damage [1]. S: single; R: repetitive; M: multiple damage.

Code	Damage Mode	Comp.	Comb.	Level	Characteristic
KA04	Fatigue: pitting	OR	S	1	Single point
KA15	Plastic def.: indent.	OR	S	1	Single point
KA16	Fatigue: pitting	OR	R	2	Single point
KA22	Fatigue: pitting	OR	S	1	Single point
KA30	Plastic def.: indent.	OR	R	1	Distributed
KB23	Fatigue: pitting	IR+OR	M	2	Single point
KB24	Fatigue: pitting	IR+OR	M	3	Distributed
KB27	Plastic def.: indent.	OR+IR	M	1	Distributed
KI04	Fatigue: pitting	IR	M	1	Single point
KI14	Fatigue: pitting	IR	M	1	Single point
KI16	Fatigue: pitting	IR	S	3	Single point
KI17	Fatigue: pitting	IR	R	1	Single point
KI18	Fatigue: pitting	IR	S	2	Single point
KI21	Fatigue: pitting	IR	S	1	Single point

For classification purposes the 32 states are grouped into four classes: *Healthy*, *Outer Ring*, *Inner Ring*, and *Combined* (the KB bearings, whose damage spans both raceways). Table 3.6 summarises the resulting corpus. One recording (N15_M01_F10_KA08_2) is corrupt in the public archive and is excluded, leaving 2,559 usable recordings.

Table 3.6: Composition of the dataset used in this work: bearing states and 4 s recordings per class (4 operating conditions $\times$ 20 repetitions each).

Class	Bearings	Recordings
Healthy	6	480
Outer Ring	12	960
Inner Ring	11	880
Combined	3	240
Total	32	2,560

3.2 Experimental Setup and Operating Conditions

3.2.1 Test Rig and Instrumentation

The data were acquired on the modular drive-train rig of the Chair of Design and Drive Technology at Paderborn University (Figure 3.1). The rig consists of a 425 W permanent-magnet synchronous drive motor fed by a standard industrial frequency inverter with a 16 kHz switching frequency, a torque-measurement shaft, the bearing test module with a spring-screw mechanism applying up to 10 kN of radial force, a flywheel, and a 1.7 kW load motor [1]. The use of an off-the-shelf inverter is a deliberate design choice: its switching noise makes the current signals representative of industrial installations rather than laboratory-clean measurements.

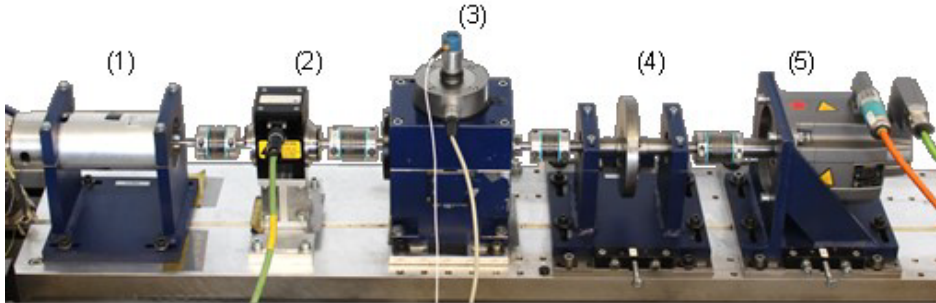


Figure 3.1: Schematic of the Paderborn KAt modular test rig. Modules: (1) drive motor (PMSM, 425 W); (2) torque-measurement shaft; (3) rolling-bearing test module with accelerometer adapter; (4) flywheel; (5) load motor (PMSM, 1.7 kW). Adapted from [1].

Two phase currents (I_a , I_b) are measured by LEM CKSR 15-NP transducers, low-pass filtered at 25 kHz, and digitised at 64 kHz. Housing vibration is measured by a piezoelectric accelerometer through a charge amplifier with a 30 kHz filter, digitised synchronously at 64 kHz. Shaft speed, torque, radial force, and oil temperature are recorded as supporting signals (Table 3.7).

Table 3.7: Instrumentation of the Paderborn test rig [1].

Quantity	Sensor / Device	Specification
Phase currents I_a, I_b	LEM CKSR 15-NP	$I_{PN} = 15$ A, accuracy 0.8%
Current sampling	DAQ board	64 kHz, 25 kHz LP filter
Housing vibration	Piezo accelerometer + charge amp.	64 kHz, 30 kHz LP filter
Shaft torque & speed	Magtrol 305	Rated 2 Nm
Radial force	Lorenz K11	Rated 10 kN
Oil temperature	Thermocouple	

3.2.2 Operating Conditions

Each bearing state was measured under the four fixed operating conditions of Table 3.8, with 20 repetitions of 4s each, giving 256,000 samples per channel per recording. Setting N15_M07_F10 is the nominal reference; the other three settings each reduce one parameter, so the individual influence of speed, torque, and radial force on diagnostic performance can be isolated. An important electrical detail follows from the drive architecture: the PMSM has four pole pairs, so the electrical fundamental of the current is 100 Hz at 1,500 rpm and 60 Hz at 900 rpm. All current-based features therefore estimate the fundamental from the measured spectrum instead of assuming a fixed supply frequency.

Table 3.8: Operating conditions of the Paderborn experiments [1].

Setting	No.	Speed [rpm]	Torque [Nm]	Radial Force [N]
N15_M07_F10	0	1,500	0.7	1,000
N09_M07_F10	1	900	0.7	1,000
N15_M01_F10	2	1,500	0.1	1,000
N15_M07_F04	3	1,500	0.7	400

3.3 Exploratory Data Analysis

Before committing to any learning pipeline, an exploratory data analysis (EDA) was performed to characterise the signals, verify data integrity, and ground every subsequent design decision in observed signal behaviour. The analyses below use the reference condition N15_M07_F10 and six representative health states: *Healthy*, *OR_Artificial*, *OR_Real*, *IR_Artificial*, *IR_Real*, and *Combined*.

3.3.1 Raw Signal Behaviour

Figure 3.2 overlays 100 ms excerpts of the raw vibration signal for all six states. Artificial defects produce clearly visible periodic impulse trains whose repetition rates match the kinematic frequencies of Equation (3.1). Naturally developed damage, by contrast, produces a diffuse, noise-like waveform that the eye cannot separate from the healthy baseline. The per-class view of Figure 3.3 quantifies this contrast: kurtosis reaches 11.6 to 11.8 for the artificial defects but only 2.2 to 2.7 for the real ones, at or below the Gaussian reference value of 3.

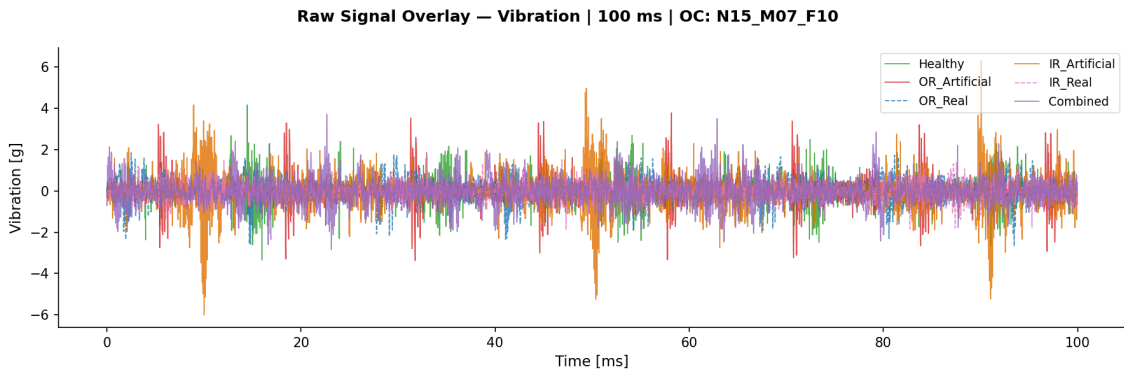


Figure 3.2: Raw vibration overlay (100 ms excerpt) for the six health states at the reference operating condition. Artificial defects show periodic impulses; real damage is visually indistinguishable from healthy.

Figures 3.4 and 3.5 compare the two damage origins directly for each fault location. The inner-race case is particularly instructive. The RMS of the naturally damaged bearing (0.352 g) is lower than the healthy reference (0.407 g), and its kurtosis (2.20) is the lowest of all six classes. Distributed micro-pitting evidently smooths the rolling contact rather than exciting sharp impulses, which means that no simple amplitude- or impulsiveness-based indicator can detect this damage class.

The corresponding current waveforms (Figures 3.6 and 3.7) are visually identical sinusoids for every health state. Whatever diagnostic information the current carries is confined to spectral fine structure far below the waveform amplitude scale.

3.3.2 Spectral Analysis

Power spectral densities were estimated by Welch’s method [34] with a 4,096-sample Hann window and 50% overlap. Figure 3.8 shows the per-class PSDs and Figure 3.9 the overlays against the healthy reference. Three observations shape the later methodology. Every faulty class elevates the energy in the 2 to 5 kHz structural-resonance band, which fixes the demodulation band for envelope analysis. The

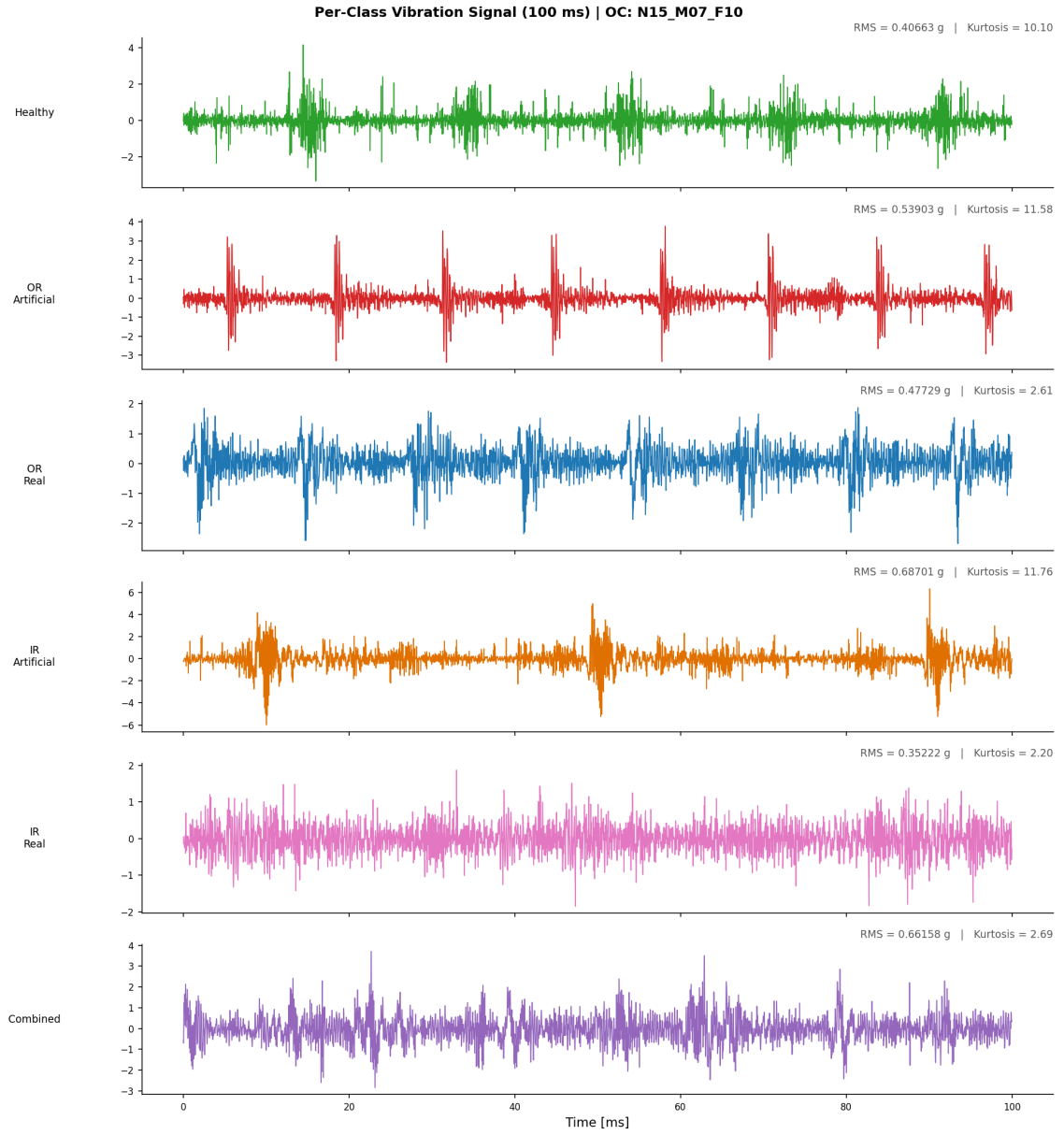


Figure 3.3: Per-class vibration signals with RMS and kurtosis annotations. The kurtosis of the artificial defects (11.58 and 11.76 for outer and inner race) collapses to 2.61 and 2.20 for their naturally damaged counterparts.

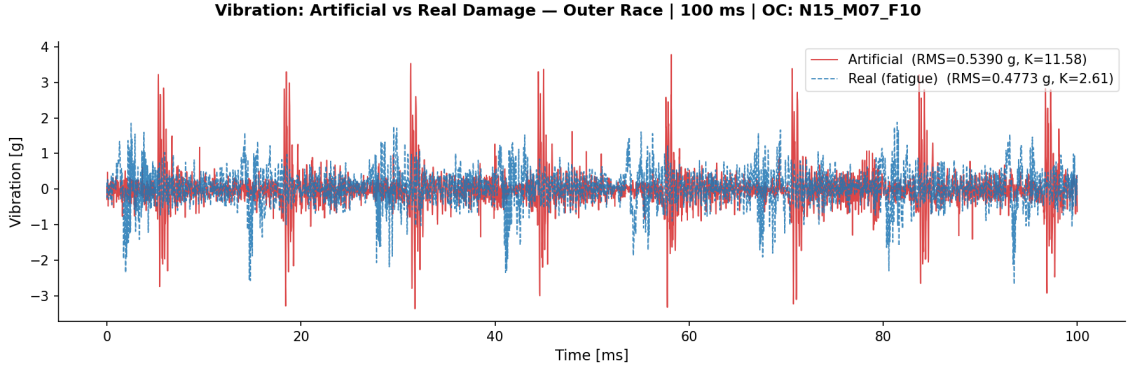


Figure 3.4: Outer-race comparison: artificial defect (RMS = 0.539 g, $K = 11.58$) versus naturally developed damage (RMS = 0.477 g, $K = 2.61$).

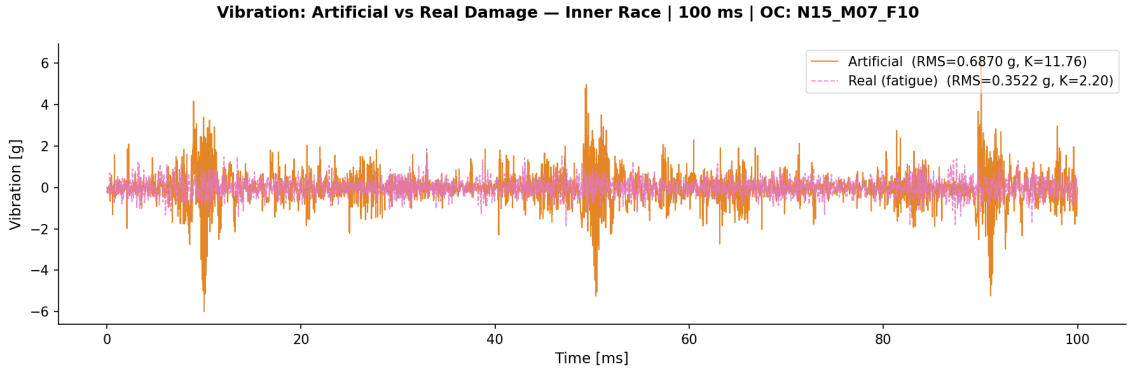


Figure 3.5: Inner-race comparison: artificial defect (RMS = 0.687 g, $K = 11.76$) versus naturally developed damage (RMS = 0.352 g, $K = 2.20$). The real fault sits below the healthy baseline in both statistics.

Figure 3.6: Raw phase current I_a overlay (100 ms). All health states produce visually identical waveforms.

Figure 3.7: Raw phase current I_b overlay (100 ms). As for I_a , no class differences are visible in the time domain.

artificial outer-race defect produces a pronounced harmonic comb at BPFO multiples. The artificial inner-race defect spreads energy beyond 9 kHz, consistent with load-zone amplitude modulation at the shaft rate.

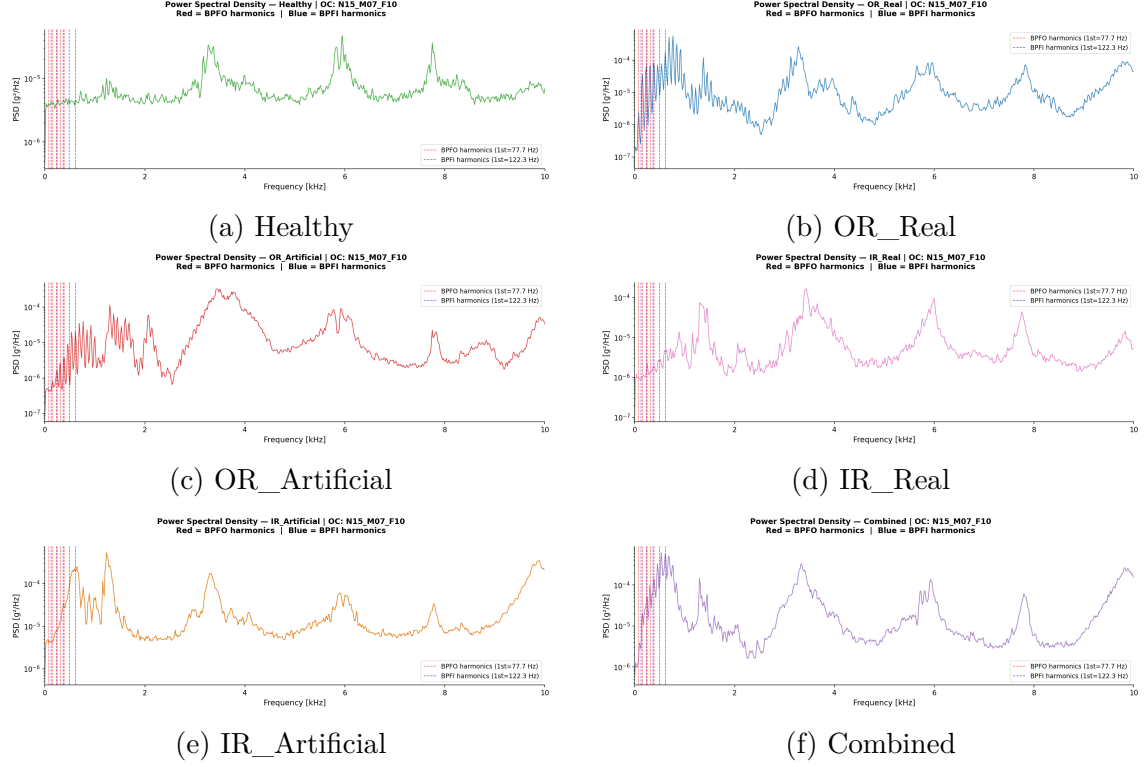


Figure 3.8: Welch PSD (0–10 kHz, log scale) per health state. Red dashed lines: BPFO harmonics; blue dashed lines: BPFI harmonics. All faulty states elevate the 2–5 kHz resonance band.

Full-resolution FFT spectra with $\Delta f = 0.25$ Hz over the complete 4 s records (Figure 3.10) resolve the individual defect lines, and the envelope spectra of Figure 3.11, computed by band-pass filtering to 2 to 10 kHz, Hilbert demodulation, and Fourier transformation of the envelope, provide the classical physics-based diagnosis picture. The envelope peaks quantify the difficulty gradient of the dataset. The artificial outer-race defect produces a BPFO peak near 0.30, its naturally damaged counterpart only 0.11, and the real inner-race fault a barely detectable 0.065.

3.3.3 Motor Current Spectral Signatures

Because the raw current waveforms are featureless, the current channel is examined in the frequency domain. Figure 3.12 shows the spectrum of phase current I_a for each health state on a logarithmic amplitude scale, with the electrical fundamental estimated from the data. The estimate lands at 100 Hz at the 1,500 rpm condition,

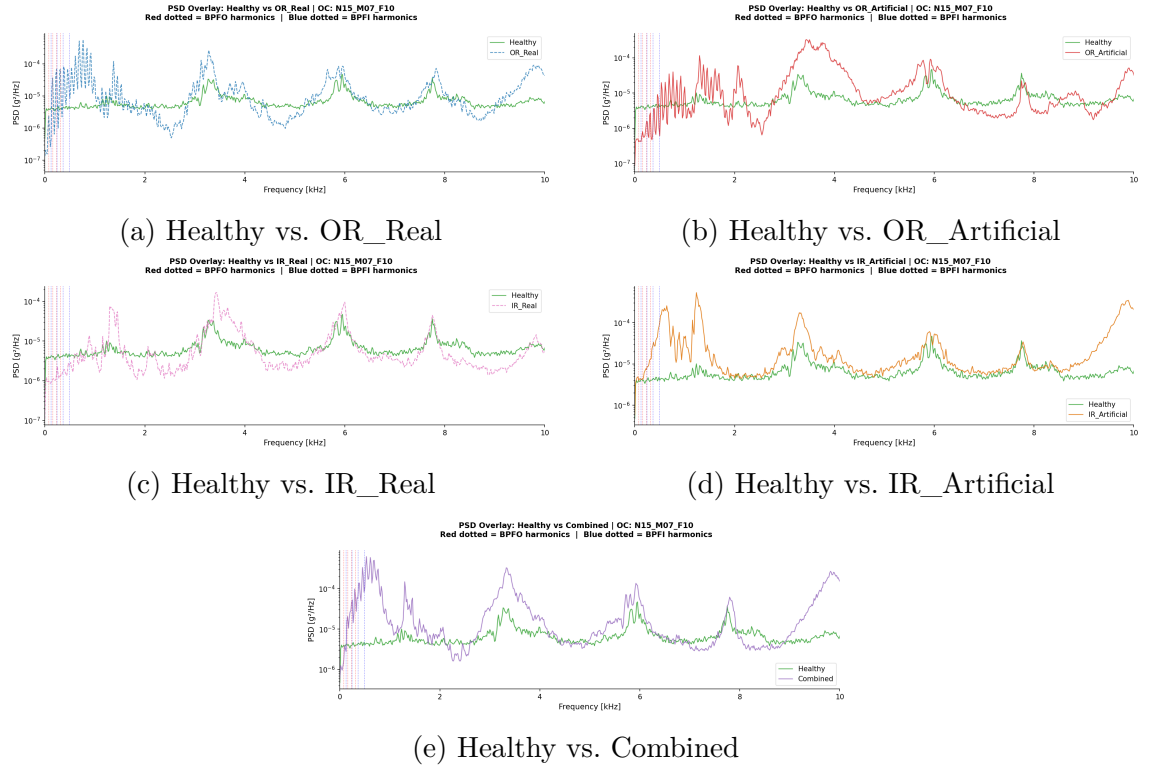


Figure 3.9: PSD overlays against the healthy reference. Artificial defects raise the spectrum by more than an order of magnitude at defect-frequency harmonics; real damage produces only subtle broadband changes.

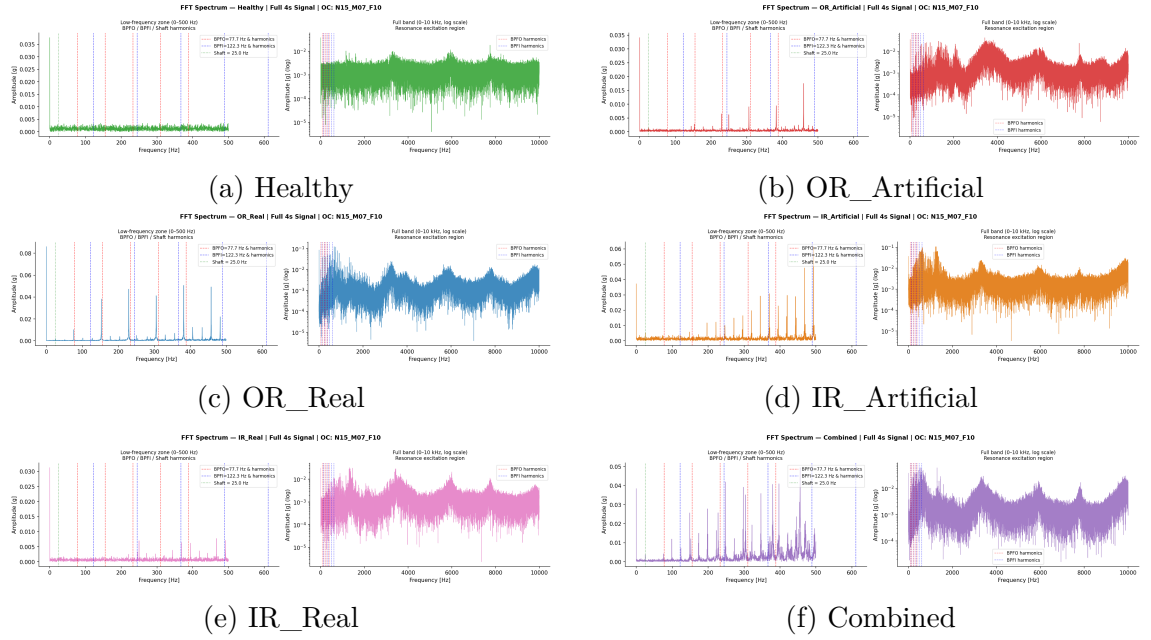
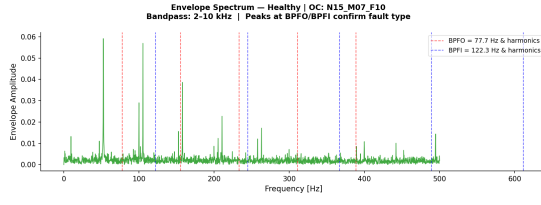
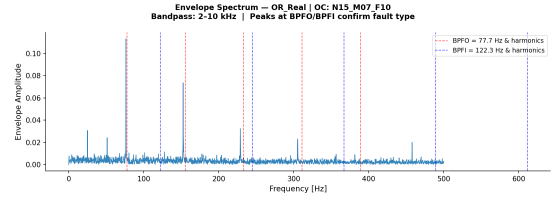


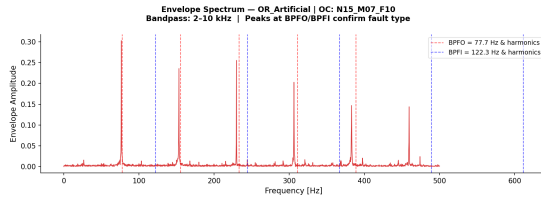
Figure 3.10: Full-record FFT spectra ($\Delta f = 0.25$ Hz). Red dashed: BPFO harmonics; blue dashed: BPFI harmonics.



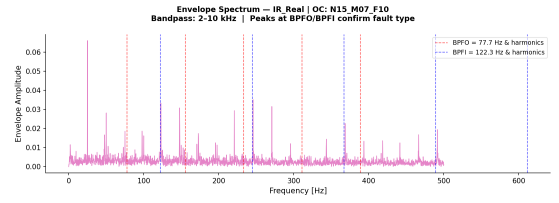
(a) Healthy



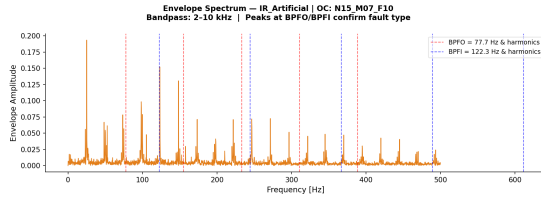
(b) OR_Real



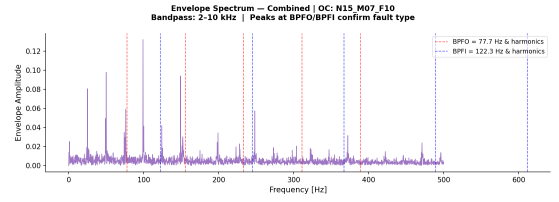
(c) OR_Artificial



(d) IR_Real



(e) IR_Artificial



(f) Combined

Figure 3.11: Envelope spectra (band-pass 2–10 kHz, Hilbert transform). BPFO peak amplitudes: OR_Artificial ≈ 0.30 , OR_Real ≈ 0.11 ; the real inner-race fault yields only ≈ 0.065 at BPFI.

four times the 25 Hz shaft frequency, which independently confirms the four-pole-pair motor construction and validates the decision not to assume a fixed 50 Hz supply. The theoretical sideband locations $f_e \pm k f_{\text{fault}}$ are marked. The fault-related components sit two to three orders of magnitude below the fundamental, exactly the regime described by the MCSA literature [6, 7].

Figure 3.13 presents the complementary and, as the later feature-importance analysis confirms, more sensitive view: the spectrum of the demodulated current envelope. Demodulation converts the fault-induced amplitude modulation into spectral lines located directly at the defect frequencies, where they are easier to measure than the sidebands themselves. This analysis motivates the current-envelope features of Section 3.5.

3.3.4 Time–Frequency and Cepstral Views

Three further analyses complete the signal characterisation and motivate the corresponding feature domains. The short-time Fourier transform of Figure 3.14 tracks the spectral content over the first second of each recording. The faulty states show persistent vertical striations across the excited resonance bands, and their regular spacing corresponds to the defect period. The continuous wavelet transform (Figure 3.15) sharpens this picture by adapting its resolution to each scale [16]: the artificial defects appear as regularly spaced impulse ridges, while naturally developed damage produces a diffuse texture without clear ridges. The real cepstrum (Figure 3.16) collapses each harmonic family of the spectrum into a single peak at its reciprocal spacing, called the quefrency, providing a compact periodicity detector. The artificial outer-race defect shows a clear harmonic at $1/\text{BPFO} \approx 13 \text{ ms}$, whereas the real-damage cepstra remain flat.

3.3.5 Statistical Summary and Dataset-Wide Views

Table 3.9 collects the time-domain statistics of the six representative recordings. The current-channel columns confirm the raw-signal impression numerically: RMS, kurtosis, and crest factor of the current are practically identical across all health states, so any current-based diagnosis must come from spectral fine structure.

The single-recording statistics generalise to the whole corpus. Figure 3.17 shows the class composition and its balance across operating conditions, and Figure 3.18 presents the distributions of RMS and kurtosis per health state over all 2,559 recordings for both the vibration and the current channel. The violin plots make two points that a single recording cannot. For vibration, the artificial sub-classes are clearly

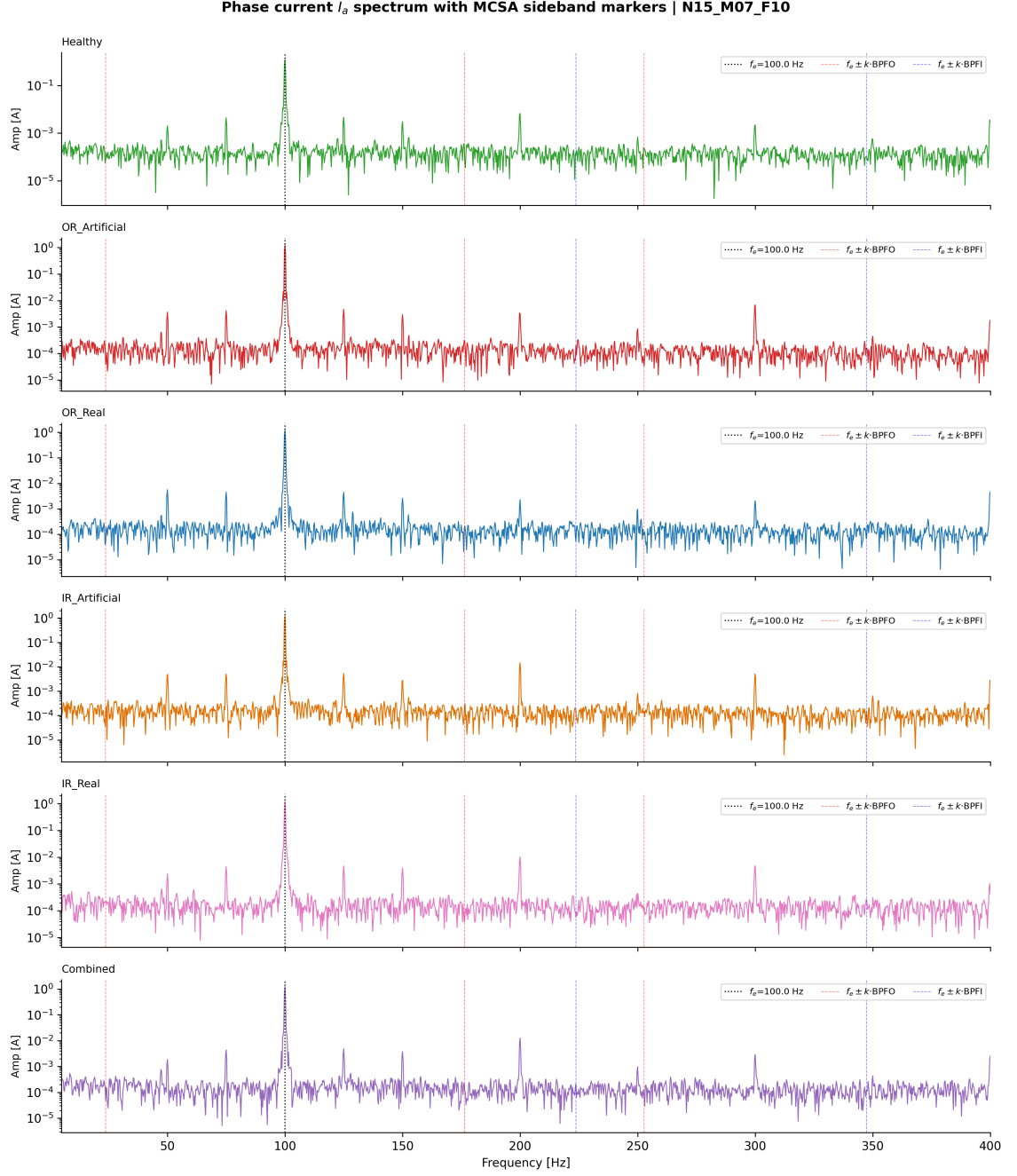


Figure 3.12: Phase-current I_a spectra per health state (log scale) with the data-estimated electrical fundamental f_e and the theoretical sideband positions $f_e \pm k \cdot \text{BPFO/BPFI}$. Fault components are orders of magnitude below the fundamental.

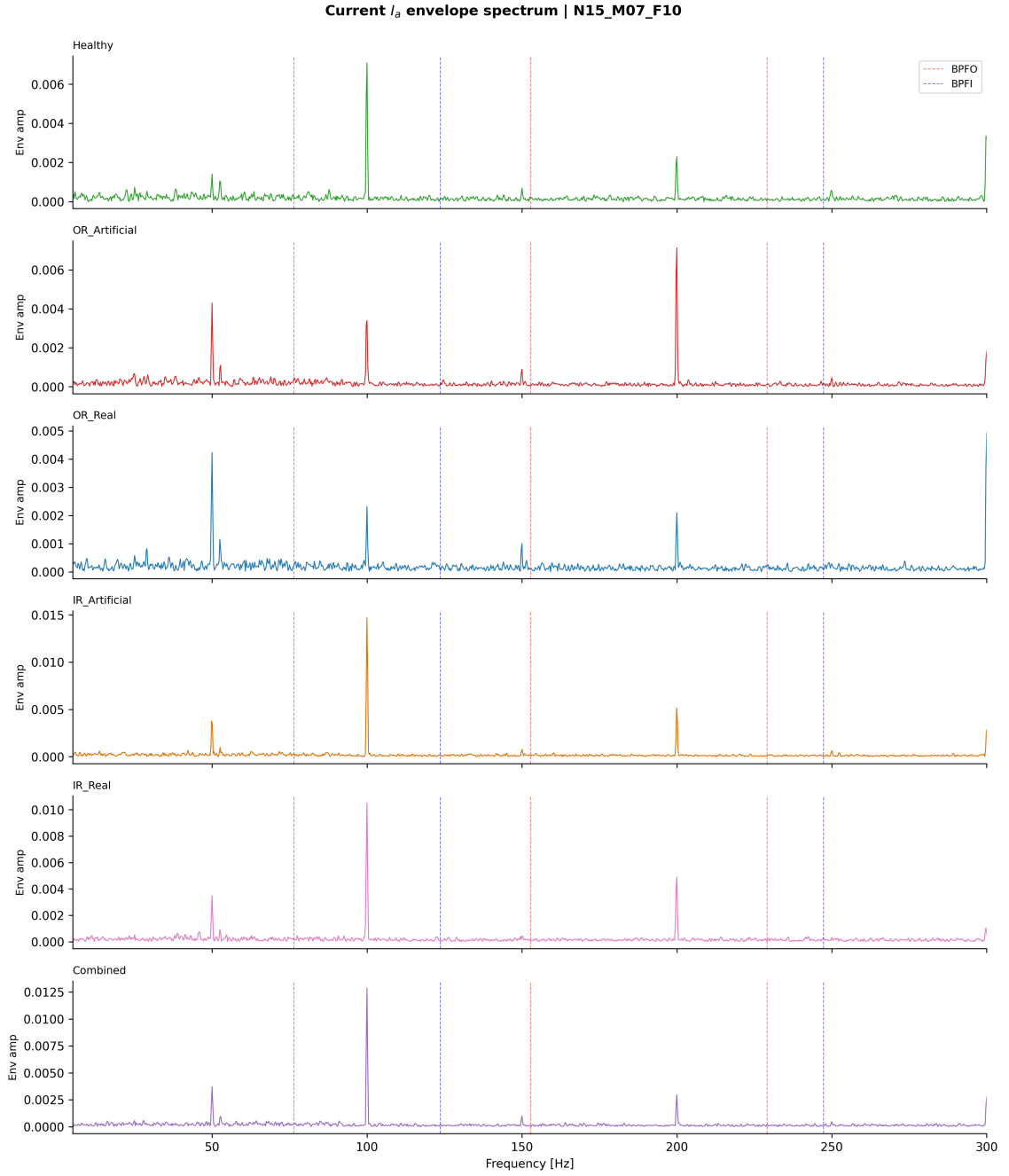


Figure 3.13: Envelope spectra of the demodulated phase current I_a per health state. Bearing-induced modulation appears as lines directly at the defect frequencies (dashed markers).

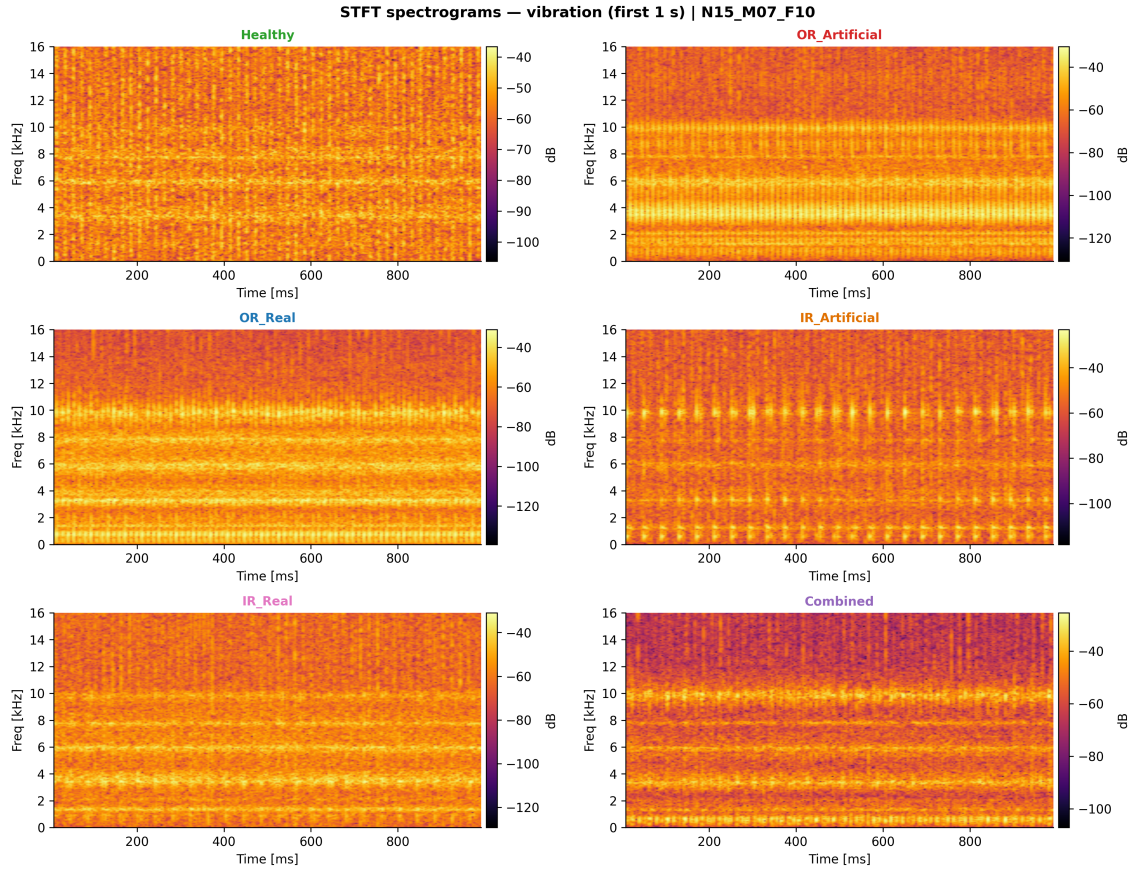


Figure 3.14: Short-time Fourier spectrograms of the vibration signal (first second) per health state. Faulty states show periodic striations across the excited resonance bands.

Table 3.9: Time-domain statistics of the six representative recordings (reference condition; 256,000 samples). CF = crest factor.

Class	Vibration					Current (I_a)			
	RMS	Kurt.	CF	Skew	Std	RMS	Kurt.	CF	Skew
Healthy	0.407	10.10	12.12	+0.026	0.406	1.767	-1.494	1.732	-0.012
OR_Artificial	0.539	11.58	7.64	+0.699	0.539	1.800	-1.503	1.694	+0.006
OR_Real	0.477	2.61	6.75	-0.492	0.475	1.728	-1.498	1.718	+0.005
IR_Artificial	0.687	11.76	11.95	+0.347	0.687	1.772	-1.495	1.773	-0.009
IR_Real	0.352	2.20	9.25	-0.001	0.352	1.785	-1.503	1.763	+0.003
Combined	0.662	2.69	9.06	+0.087	0.661	1.779	-1.500	1.693	+0.009

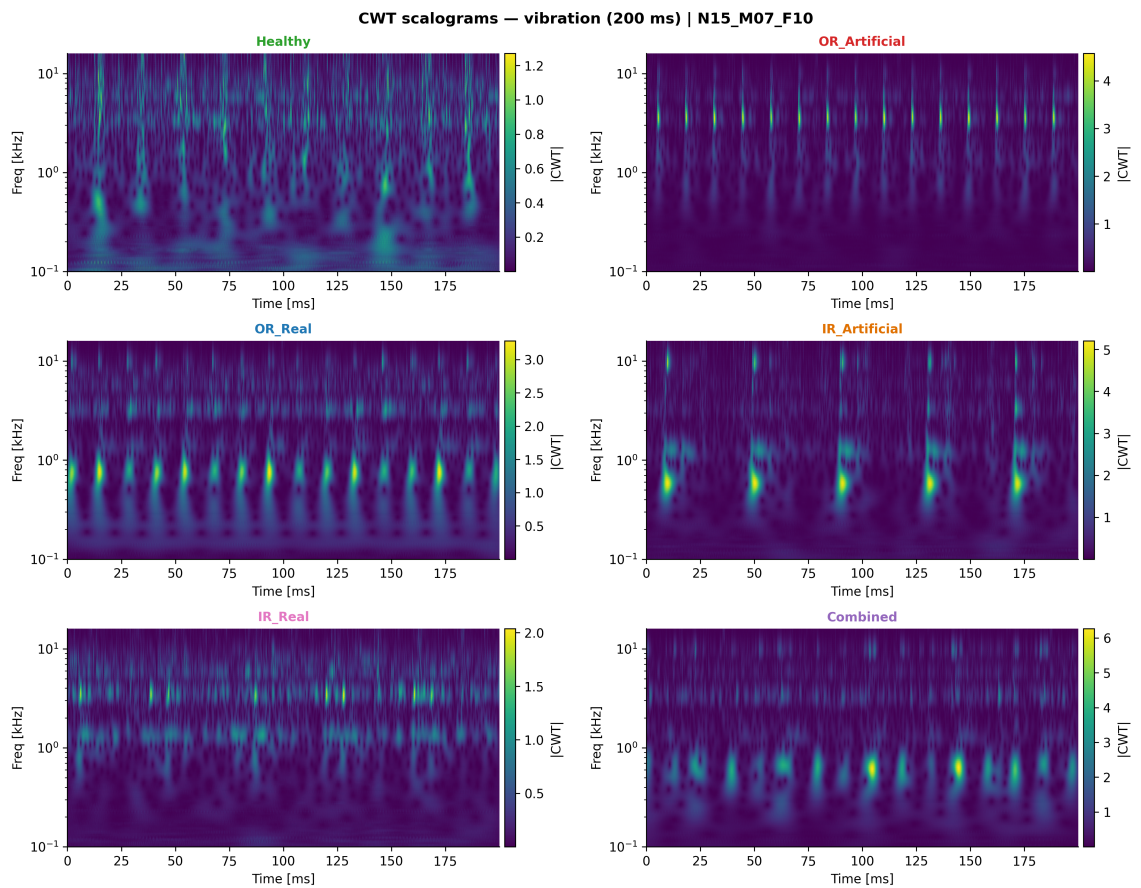


Figure 3.15: Continuous wavelet scalograms of the vibration signal (200 ms) per health state. Artificial defects produce periodic impulse ridges; real damage produces diffuse texture.

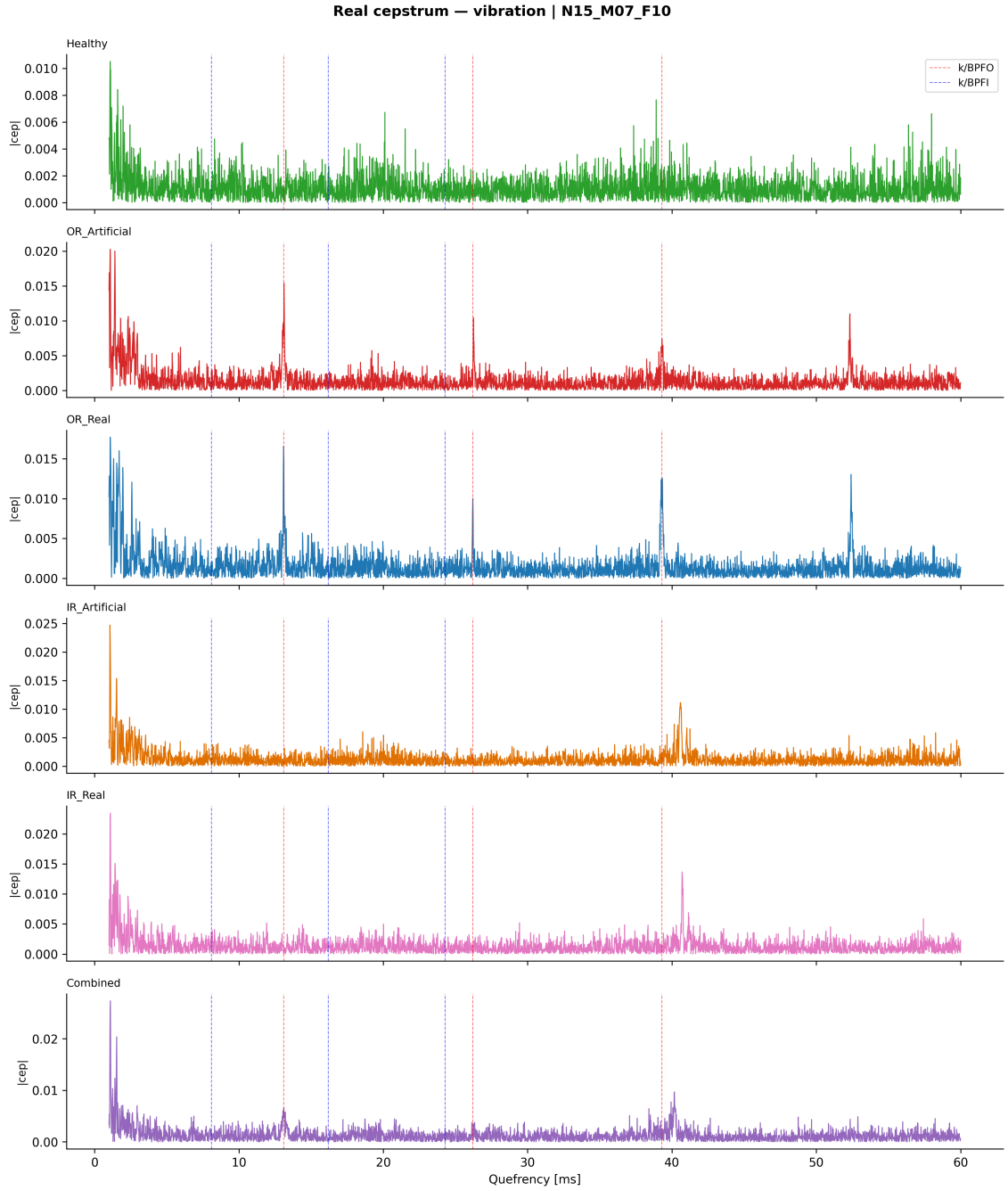


Figure 3.16: Real cepstra of the vibration signal with the expected harmonic positions $k/BPFO$ and $k/BPFI$ marked. Harmonic periodicity is clearly detectable only for the artificial defects.

shifted in kurtosis but the real sub-classes overlap the healthy distribution almost entirely, so the single-recording observations of the previous subsections are not outliers but the norm. For current, all six distributions coincide to within plotting accuracy, which rules out time-domain current statistics as features and is why the current feature set of Section 3.5 is built exclusively on spectral and modulation descriptors.

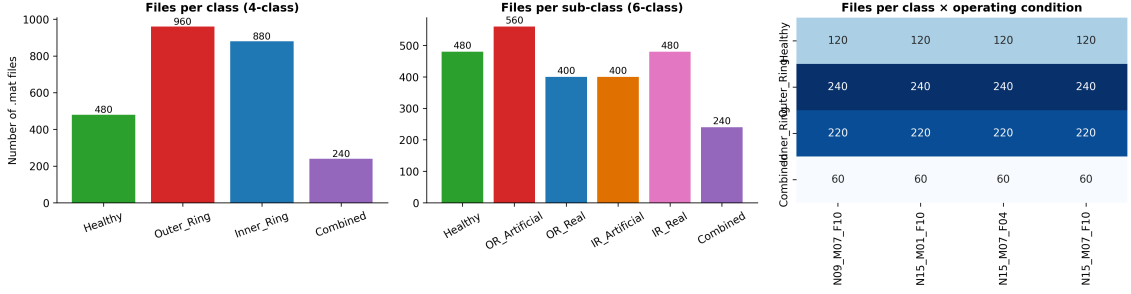


Figure 3.17: Dataset composition: recordings per class (4- and 6-class views) and the balanced class $\times$ operating-condition design.

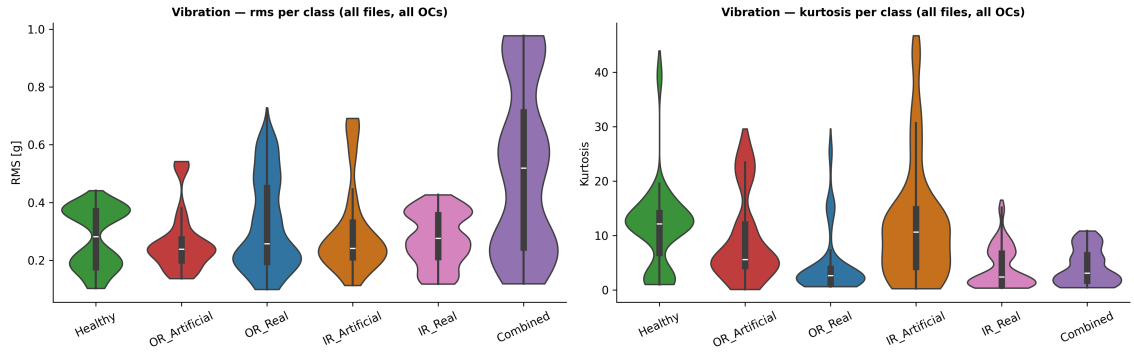
Figure 3.19 condenses the central challenge of the benchmark into a single image by placing every recording in the RMS-kurtosis plane: artificial defects form a separable high-kurtosis cloud, while naturally damaged recordings overlap the healthy population almost completely.

3.3.6 The Operating-Condition Effect

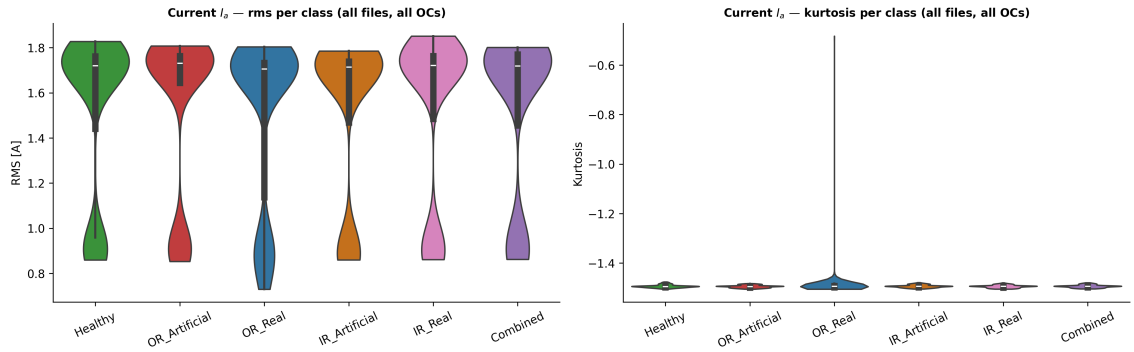
The final exploratory question concerns how strongly the operating condition, rather than the health state, shapes the signals. Figure 3.20 shows one and the same physically damaged bearing measured under the four conditions: the change in signal character between conditions is comparable in magnitude to the change a fault produces at a fixed condition. The dataset-wide view of Figure 3.21 confirms this quantitatively. Mean vibration RMS varies almost as much across the columns (conditions) as across the rows (health states). Two methodological consequences follow: cross-condition evaluation is mandatory for any honest performance claim, and the healthy-only detectors of Phase 2 must be conditioned on the measured operating condition, since otherwise condition-induced variance masks health-induced variance.

3.3.7 Implications for the Methodology

Five operational conclusions follow from the EDA and govern everything that comes after:



(a) Vibration channel



(b) Current channel I_a

Figure 3.18: Distributions of RMS and kurtosis per health state over all recordings and operating conditions. Vibration separates the artificial sub-classes only; the current statistics are indistinguishable across all states.

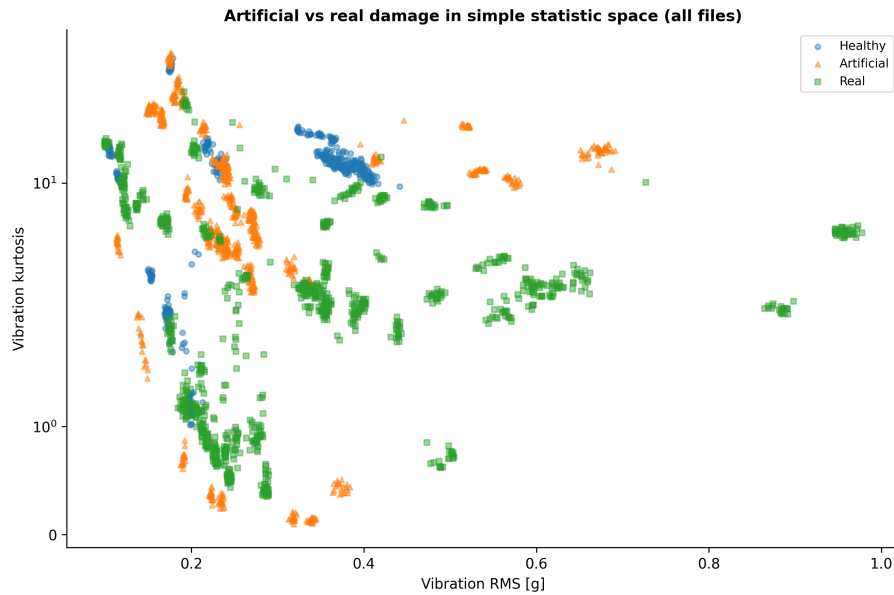


Figure 3.19: All 2,559 recordings in the vibration RMS-kurtosis plane. Artificial damage (triangles) separates cleanly; naturally developed damage (squares) overlaps the healthy cluster (circles).

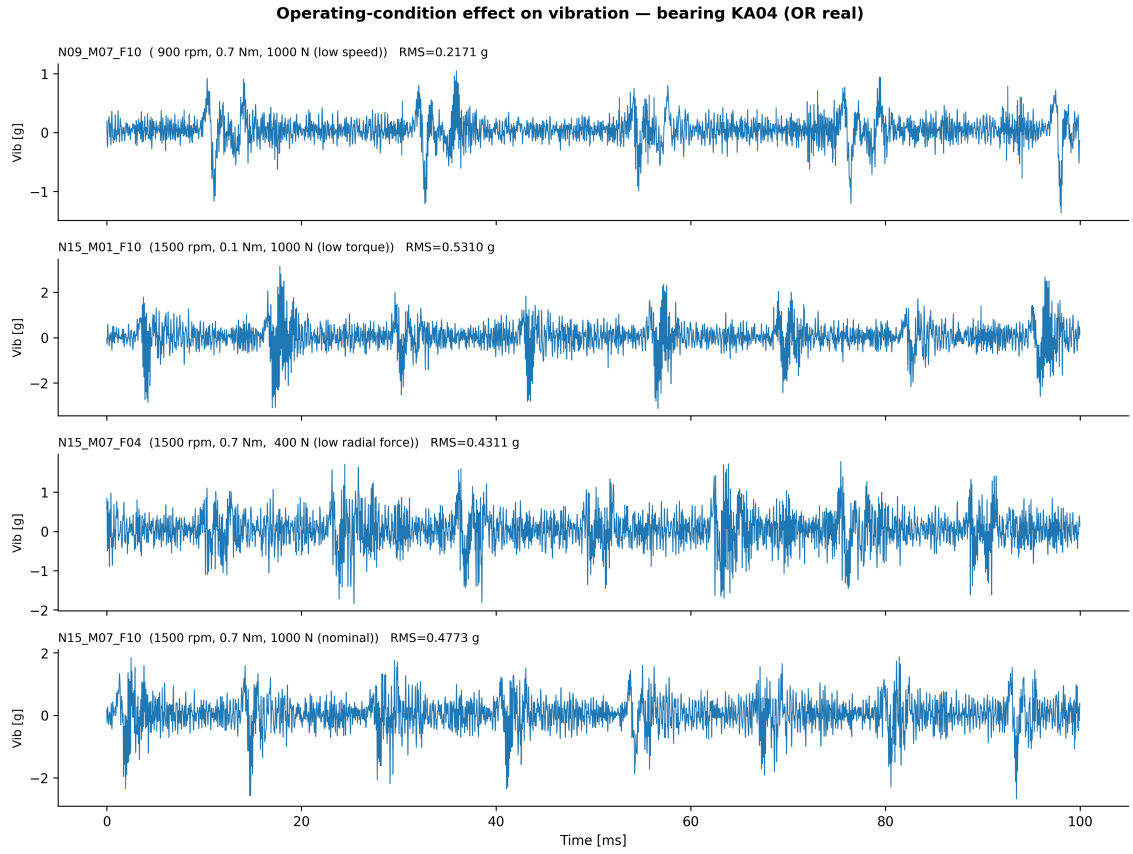


Figure 3.20: One naturally damaged bearing (KA04) measured under the four operating conditions. The condition changes the signal almost as much as the damage does.

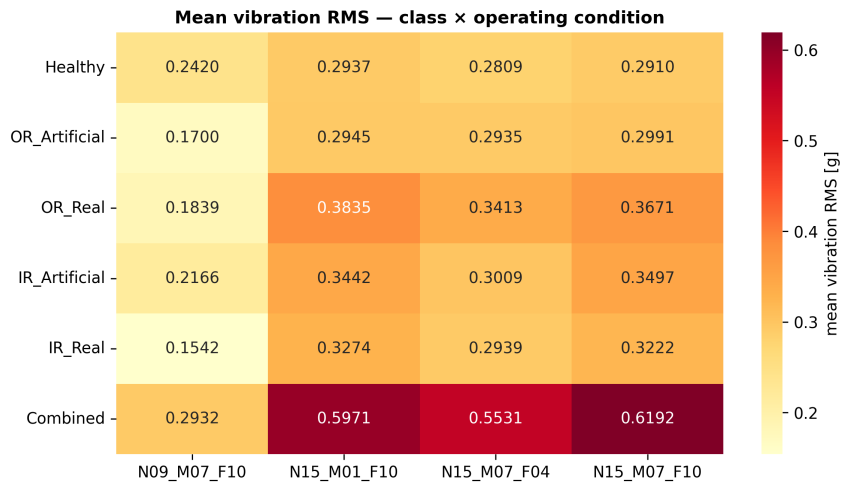


Figure 3.21: Mean vibration RMS per health state (rows) and operating condition (columns) over the whole dataset. Condition-induced variation rivals fault-induced variation.

1. **Windowing is required.** Twenty recordings per bearing and condition are too few training samples; segmenting each 4 s record into overlapping windows multiplies the sample count while preserving within-recording stationarity.
2. **A single-domain feature set cannot work.** Kurtosis and RMS fail outright for real damage; spectral, envelope, and time-frequency descriptors must carry the discrimination.
3. **Current diagnosis is a spectral problem.** No time-domain current statistic varies with health state; the information resides in sidebands and modulation structure, which dictates long windows and dedicated MCSA features.
4. **Artificial and real damage are distinct sub-domains.** A kurtosis gap of roughly nine units separates the two damage origins; any evaluation that pools them without also testing the cross-domain direction will overstate practical performance.
5. **The operating condition is a dominant nuisance variable.** Condition-induced signal variation rivals fault-induced variation (Figures 3.20 and 3.21), so cross-condition evaluation and condition-aware normalisation are built into the methodology from the outset.

3.4 Preprocessing and Segmentation

3.4.1 Windowing

Each 4 s recording is segmented into windows of 32,000 samples (0.5 s) with 50% overlap, producing 15 windows per recording and 38,388 windows in total. The window length is set by the current channel, not the vibration channel. MCSA sidebands sit only tens of hertz from the electrical fundamental, and the envelope-spectrum lines of the defect frequencies are similarly narrow, so a frequency resolution of $\Delta f = 2$ Hz is needed to resolve them. Shorter windows common in vibration-only work (2,048 to 4,096 samples) yield $\Delta f \geq 15$ Hz and smear the sidebands into the fundamental. The same windows are used for both modalities so that the comparison remains strictly like for like. The 50% overlap follows common practice: it doubles the number of training windows at the cost of statistical dependence between neighbours, and that dependence is neutralised by the grouping rule described next.

3.4.2 Leakage Control

Because adjacent windows overlap and windows from one recording share the same signal realisation, any train-test split that separates windows individually leaks information: the model is effectively tested on slightly shifted copies of its training data. Throughout Phase 1, splits are therefore grouped by recording, meaning that all 15 windows of a recording fall entirely on one side of every split. Chapter 4 additionally reports leave-one-bearing-out evaluation, in which the grouping unit is the physical bearing. That is the strictest protocol, and it controls for the possibility that a model recognises the bearing individual rather than the fault.

3.4.3 Per-Recording Physical Parameters

Two parameters are computed per recording rather than assumed globally. The kinematic defect frequencies of Equations (3.1) and (3.2) are scaled by the actual shaft speed encoded in the filename (25 Hz or 15 Hz). The electrical fundamental of the current is estimated as the dominant spectral line below 400 Hz. This data-driven estimate automatically tracks the pole-pair scaling (100 Hz at 1,500 rpm, 60 Hz at 900 rpm) and any small deviation of the drive from its set-point.

3.5 Feature Engineering

The EDA established that no single signal-analysis domain suffices for both damage origins. The feature set therefore spans six domains, summarised in Table 3.10. This section introduces each domain in turn: what the underlying transform computes, why it is expected to carry diagnostic information for this dataset, and which features are extracted from it. All features are computed per window. The vibration channel contributes 68 features and the two current channels 110, for a total of 178. In the notation below, $x[n]$ denotes a window of N samples, $\bar{x}$ its mean, and σ its standard deviation.

3.5.1 Time-Domain Features

Time-domain statistics summarise the amplitude distribution of the window without any transformation. They are computationally trivial, directly interpretable, and historically the first features used in machinery monitoring [17]. Eleven descriptors are computed per channel. The energy-related quantities are the root mean square

Table 3.10: Multi-domain feature set extracted per 0.5 s window (68 vibration + 110 motor-current features).

Domain	Features
Time	33
Frequency	67
Time–frequency (wavelet packet)	27
Envelope	26
MCSA	18
Cepstrum	7
Total	178

and the peak value,

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{n=1}^N x[n]^2}, \quad \text{Peak} = \max_n |x[n]|, \quad (3.3)$$

and the dispersion is captured by the standard deviation and the mean absolute value $\text{MAV} = \frac{1}{N} \sum_n |x[n]|$. The shape-related quantities are the third and fourth standardised moments,

$$\text{Skewness} = \frac{\frac{1}{N} \sum_n (x[n] - \bar{x})^3}{\sigma^3}, \quad \text{Kurtosis} = \frac{\frac{1}{N} \sum_n (x[n] - \bar{x})^4}{\sigma^4} - 3, \quad (3.4)$$

where the subtraction of 3 references the Gaussian value, so that positive kurtosis indicates a heavy-tailed, impulsive amplitude distribution. Three normalised impulsiveness indicators complete the statistical picture: the crest factor Peak/RMS , the shape factor RMS/MAV , and the impulse factor Peak/MAV . Finally, the waveform length $\sum_n |x[n+1] - x[n]|$ measures signal complexity, and the zero-crossing rate acts as a crude frequency-content proxy.

The diagnostic reasoning behind this group is straightforward. A single-point defect produces short high-amplitude impacts, which raise the peak much faster than the RMS, inflate the crest and impulse factors, and drive the kurtosis far above zero. Distributed roughness, in contrast, raises the broadband floor and leaves all impulsiveness indicators essentially unchanged, which is exactly the failure mode the EDA documented. The group is retained despite this known weakness because it is nearly free to compute and remains informative for the artificial sub-domain.

3.5.2 Frequency-Domain Features

The frequency-domain group operates on the one-sided amplitude spectrum of the Hann-windowed signal,

$$X[k] = \frac{2}{N} \left| \sum_{n=0}^{N-1} w[n] x[n] e^{-j2\pi kn/N} \right|, \quad (3.5)$$

where $w[n]$ is the Hann window. Windowing matters here: without it, spectral leakage from the strong tonal components (the shaft harmonics in vibration, the fundamental in current) smears energy across neighbouring bins and buries the small fault lines that the features are supposed to measure.

Five spectral-shape descriptors summarise the distribution of energy over frequency. With the normalised power distribution $p[k] = X[k]^2 / \sum_j X[j]^2$, the spectral centroid and the RMS frequency are the first two moments $\sum_k f_k p[k]$ and $\sqrt{\sum_k f_k^2 p[k]}$, the dominant frequency is the location of the largest bin, the spectral entropy

$$H = - \sum_k p[k] \log_2 p[k] \quad (3.6)$$

measures how uniformly the energy is spread (high for broadband noise, low for tonal content), and the spectral kurtosis of the amplitude values flags spectra dominated by a few strong lines. Relative band powers over four bands (0–1, 1–5, 5–16, and 16–32 kHz) quantify the resonance-band elevation observed in the EDA in a form that is insensitive to overall signal scale.

The physically sharpest members of this group are the defect-frequency amplitudes: the spectrum is evaluated at the first five harmonics of BPFO, BPFI, BSF, and FTF, taking the maximum within a ± 3 Hz tolerance band around each target to absorb small deviations of the actual shaft speed from its set-point. These features encode the kinematics of Equations (3.1) and (3.2) directly and are the reason the defect frequencies are recomputed per recording.

3.5.3 Time–Frequency Features: Wavelet Packet Energies

Fourier analysis assumes stationarity within the window, but impact sequences are inherently transient. The wavelet packet transform (WPT) addresses this by recursively filtering the signal into equal-width frequency sub-bands with a filter pair derived from a mother wavelet, retaining time localisation inside each band [16]. A three-level decomposition with the Daubechies-4 wavelet is used, chosen because its compact support matches short transients and because three levels split the 0 to

32 kHz range into eight bands of 4 kHz, wide enough to be robust against speed variation and narrow enough to separate the resonance regions. With E_i denoting the energy of the i -th terminal node, the features are the relative energies $e_i = E_i / \sum_j E_j$ and their Shannon entropy $-\sum_i e_i \log_2 e_i$. The relative energies describe where in the spectrum the energy sits, and the entropy describes how concentrated the distribution is: a healthy signal spreads its energy broadly, whereas a defect that excites one resonance band concentrates it.

3.5.4 Envelope Features

The envelope group implements the high-frequency resonance technique that the EDA identified as the most discriminative classical analysis [4]. The chain has three steps. First, the vibration window is band-pass filtered to 2 to 10 kHz with a fourth-order Butterworth filter, isolating the structural resonances that the defect impacts excite while rejecting low-frequency shaft harmonics and high-frequency noise. Second, the analytic signal is formed with the Hilbert transform,

$$x_a(t) = x(t) + j \mathcal{H}\{x\}(t), \quad e(t) = |x_a(t)|, \quad (3.7)$$

whose magnitude $e(t)$ is the instantaneous amplitude, that is, the envelope. Third, the spectrum of the mean-removed envelope is computed. If the resonance is amplitude-modulated at a defect frequency, the envelope spectrum shows a line exactly there, independent of which resonance carries the modulation. The extracted features are the envelope amplitudes at the first five harmonics of BPFO and BPFI, again with the ± 3 Hz tolerance, together with the envelope RMS and the envelope kurtosis. These are the most physically interpretable features of the entire set, and the feature-importance analysis of Chapter 4 confirms that the classifiers rely on them heavily.

3.5.5 MCSA Features

The motor-current group translates the same demodulation logic to the electrical domain. A bearing defect modulates the load torque at the defect frequency; the motor responds with a small oscillation of speed and air-gap flux; and the stator current is modulated accordingly, which in the spectrum appears as sidebands at

$$f_{sb} = f_e \pm k f_{\text{fault}}, \quad k = 1, 2, 3, \quad (3.8)$$

around the electrical fundamental f_e [6, 7]. Because f_e on this inverter-fed machine follows the commanded speed, it is estimated per recording as the dominant line below 400 Hz, as described in Section 3.4. For each current phase the extracted features are the fundamental amplitude, the total harmonic distortion

$$\text{THD} = \frac{\sqrt{\sum_{k=2}^5 X(kf_e)^2}}{X(f_e)}, \quad (3.9)$$

and the sideband amplitudes of Equation (3.8) for BPFO and BPF1, normalised by the fundamental amplitude so that they measure modulation depth rather than load level. In addition, the current is band-pass filtered and demodulated exactly as the vibration signal, and the envelope-spectrum amplitudes at the defect frequencies are extracted. This demodulated view proved the most sensitive current indicator in the exploratory analysis (Figure 3.13), because it converts the sideband pattern into direct lines at the fault frequencies where they are easier to measure against the local noise floor.

3.5.6 Cepstral Features

The real cepstrum is the inverse Fourier transform of the logarithmic magnitude spectrum,

$$c[q] = \mathcal{F}^{-1}\{\log |X(f)|\}, \quad (3.10)$$

with the independent variable q measured in seconds and called the quefrency. Its diagnostic value comes from a simple property: a family of harmonics with spacing f_0 in the spectrum appears as a single peak (a rahmonic) at $q = 1/f_0$ in the cepstrum. Instead of measuring many individual harmonic lines, a single cepstral value therefore tests whether the whole family is present. The extracted features are the rahmonic amplitudes at k/BPFO and $k/\text{BPF1}$ for $k \in \{1, 2, 3\}$ and a peak-prominence measure over the physiological quefrency range, defined as the ratio of the tallest cepstral peak to the local mean. The cepstrum is computed for the vibration channel, where harmonic families are the expected fault signature.

3.5.7 Feature Selection

With 178 features, redundancy and noise dilute classifier performance and slow training, so a three-stage funnel reduces each modality's set to at most 30 features before classical model training. The first stage removes near-constant features, whose variance is below a numerical threshold and which therefore cannot discriminate any-

thing. The second stage addresses redundancy: for every feature pair with absolute Pearson correlation

$$r_{ij} = \frac{\text{cov}(x_i, x_j)}{\sigma_i \sigma_j} \quad (3.11)$$

above 0.95, the member with the lower mutual information with the class label is removed, since keeping both adds no information while inflating the dimensionality. The third stage ranks the survivors by mutual information

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x) p(y)}, \quad (3.12)$$

which, unlike correlation, captures arbitrary nonlinear dependencies between a feature and the class label, and retains the thirty highest-ranked features. The funnel is fitted on training data only, so no test information influences the selection.

Figure 3.22 shows the resulting global ranking. Envelope and defect-frequency amplitudes dominate the vibration side and demodulated-current features carry the current side, which is reassuring because these are precisely the features the physics predicts should matter. Figure 3.23 shows the correlation structure among the top-ranked features and explains why the redundancy stage is necessary: harmonics of the same defect frequency are strongly correlated with one another, and without the correlation filter the final selection would consist of near-duplicates. Figure 3.24 displays the class-conditional distributions of the six strongest features, and Figure 3.25 previews the class structure of the three modality subsets before any classifier is trained. Vibration and fused features form visibly separable clusters, while the current-only embedding shows heavier class overlap, an early indication of the modality ranking that the classifiers later confirm.

3.6 Phase 1: Supervised Baselines and Evaluation Protocols

3.6.1 Model Set

Four classifiers spanning distinct learning paradigms are compared. Each is introduced with a short account of how it works and why it belongs in the comparison.

Support vector machine. The SVM seeks the decision boundary that separates the classes with the largest margin. With the radial basis function kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2), \quad (3.13)$$

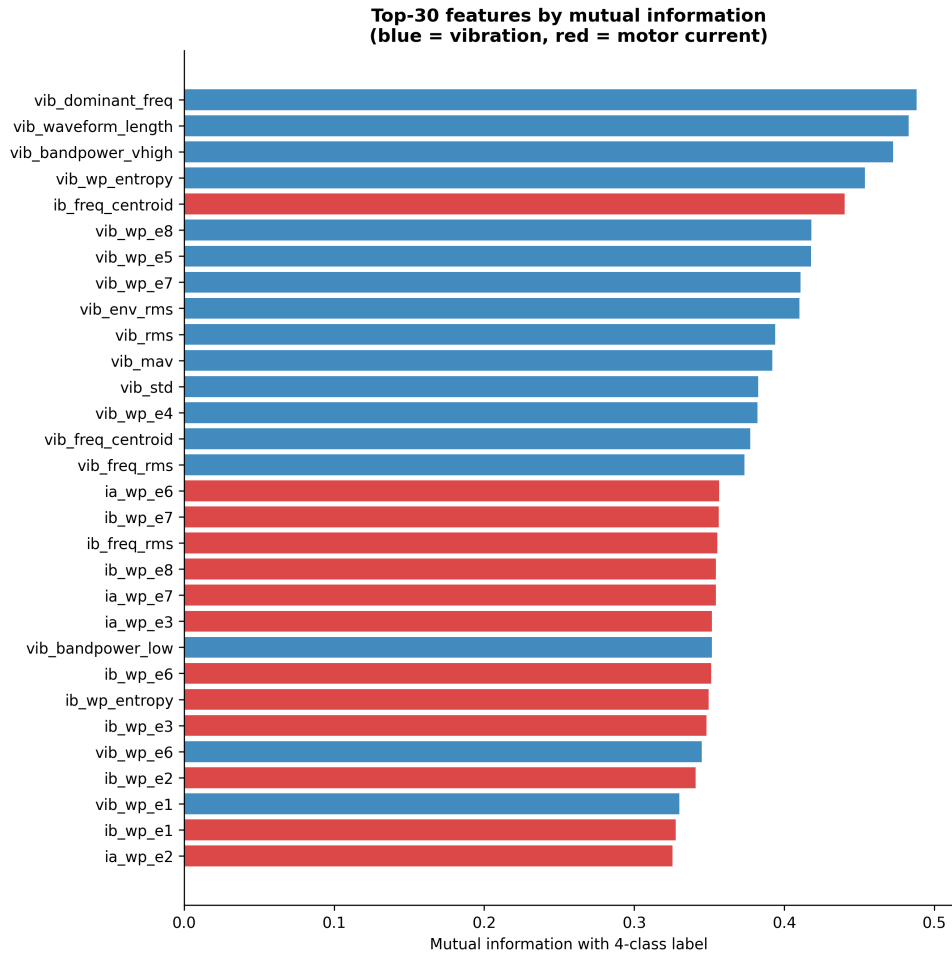


Figure 3.22: Top-30 features by mutual information with the 4-class label. Envelope and defect-frequency amplitudes dominate the vibration side; demodulated-current features carry the current side.

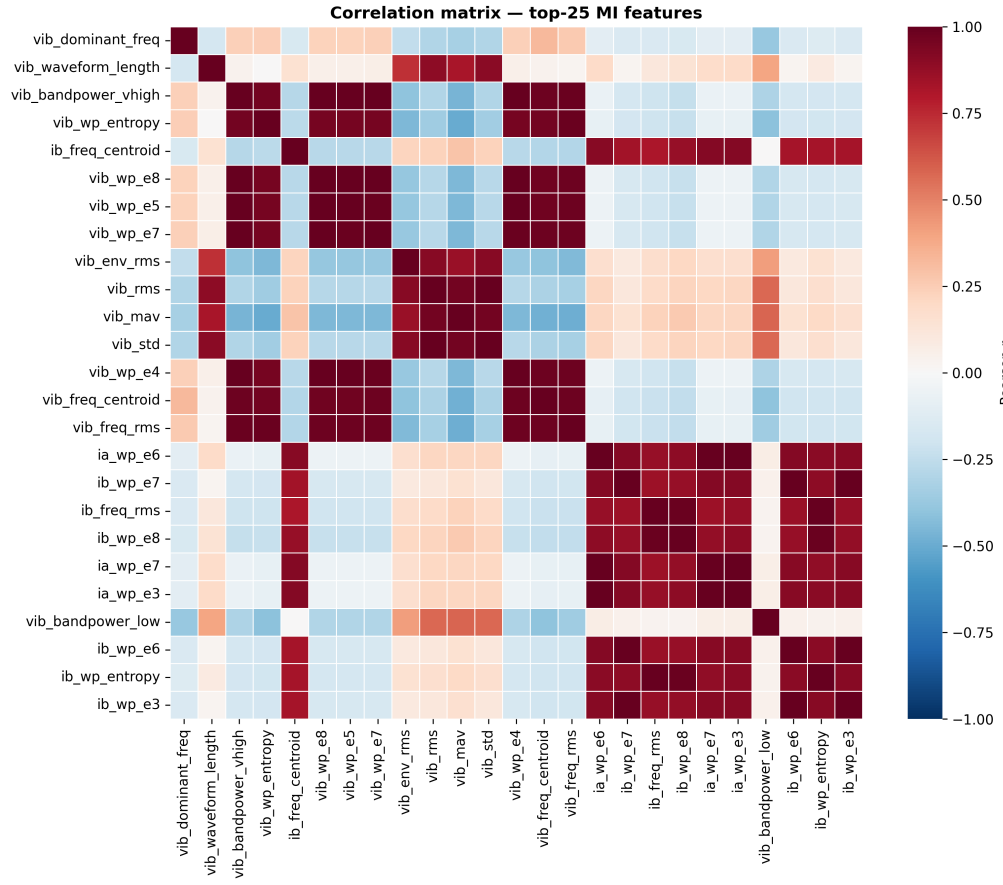


Figure 3.23: Pearson correlation matrix of the top-25 features by mutual information. Harmonic families of the same defect frequency form strongly correlated blocks, motivating the redundancy filter.

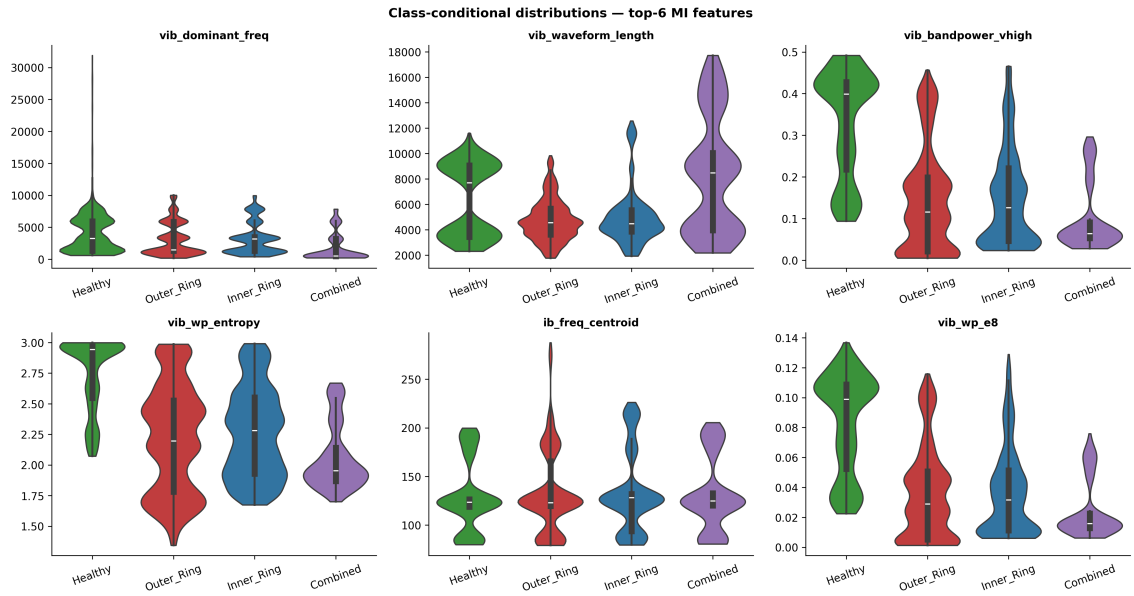


Figure 3.24: Class-conditional distributions of the six features with the highest mutual information. Each feature separates some class pairs but none separates all, which is why the classifiers operate on the combined selection.

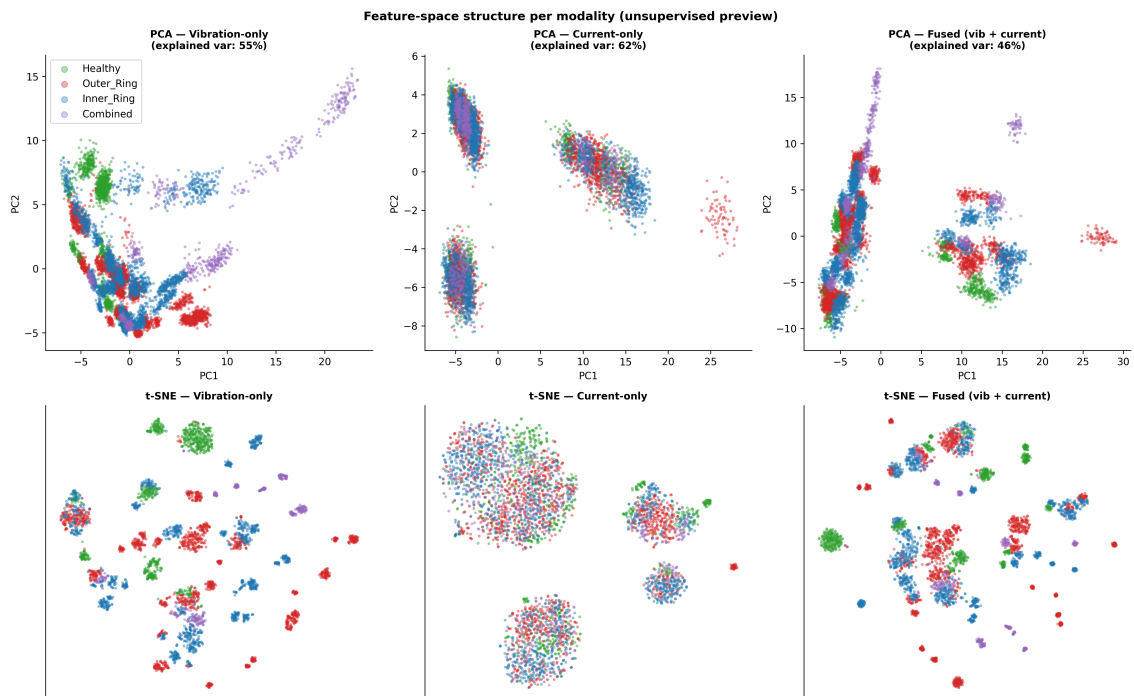


Figure 3.25: PCA (top) and t-SNE (bottom) embeddings of the feature space for vibration-only, current-only, and fused subsets, coloured by class. Class structure is visible before any supervised training.

the boundary becomes an arbitrary smooth surface in the original feature space, and the decision function is a weighted sum of kernel evaluations against the support vectors. The regularisation constant C trades margin width against training errors, and γ sets the kernel length scale. The SVM is the classical reference for moderate-size engineered feature sets [20], and class weighting inversely proportional to class frequency compensates the healthy-to-faulty imbalance of the dataset.

k -nearest neighbours. The k -NN rule assigns a window the majority class among its k closest training windows, with votes weighted by inverse distance and $k = 5$. It involves no training at all, so its performance measures how much of the problem is solvable by raw feature-space geometry. That makes it a useful diagnostic: when k -NN performs well, the classes form compact neighbourhoods; when it fails while other models succeed, the discriminative structure is present but distributed across features in a way that distance in the full space does not capture.

Random Forest. A Random Forest trains several hundred decision trees, each on a bootstrap resample of the training data and each considering only a random feature subset at every split [21]. The ensemble average cancels much of the variance of the individual trees, which makes the model robust to noisy and redundant features and gives it a built-in feature-importance measure through the accumulated impurity reduction per feature. It represents the bagging paradigm in the comparison and needs essentially no tuning to perform well.

XGBoost. Gradient boosting builds trees sequentially, each one fitted to the errors of the current ensemble. XGBoost regularises this process explicitly by minimising

$$\mathcal{L} = \sum_i \ell(y_i, \hat{y}_i) + \sum_t \left[\gamma T_t + \frac{1}{2} \lambda \|\mathbf{w}_t\|^2 \right], \quad (3.14)$$

where ℓ is the classification loss and the second term penalises the number of leaves T_t and the leaf weights $\mathbf{w}_t$ of each tree t [22]. The regularisation keeps the boosted ensemble from overfitting the moderate-sized training sets used here. Balanced sample weights handle class imbalance. XGBoost represents the boosting paradigm and, as Chapter 4 shows, is the best Phase-1 model overall, so it also serves as the reference model for the stricter protocols.

3.6.2 Evaluation Metrics

All models operate on standardised features, with the standardisation statistics fitted on training data only. Reported metrics are accuracy, precision, recall, and F1. For a class c with true positives TP_c , false positives FP_c , and false negatives FN_c ,

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = \frac{2P_cR_c}{P_c + R_c}, \quad (3.15)$$

and the macro F1 is the unweighted mean of the $F1_c$ over classes. The macro average is the headline metric of this thesis because it weights the minority Healthy class equally with the large fault classes; plain accuracy would reward a model that sacrifices the healthy class entirely. For threshold-free comparisons of detectors the area under the receiver operating characteristic curve (AUROC) is used, which equals the probability that a randomly chosen faulty window receives a higher score than a randomly chosen healthy one. Every experiment fixes random seeds, and deep models are additionally repeated over five seeds.

3.6.3 Evaluation Protocol Suite

Rather than a single train-test split, Phase 1 uses a graded suite of protocols whose strictness increases in the order listed. The design follows directly from the review literature’s critique of single-protocol evaluation [10, 11].

1. **Pooled grouped CV.** Five-fold cross-validation over all windows, stratified by class and grouped by recording (Section 3.4.2). This is the community-standard benchmark and the point of comparison with published results.
2. **Damage-domain scenarios.** The same protocol restricted to the artificial-only sub-domain (three classes) and the real-only sub-domain (three classes), quantifying how much harder naturally developed damage is.
3. **Bidirectional transfer.** Training on healthy plus artificial damage and testing on healthy plus real damage, which is the industrially decisive direction and analogous to the cross-domain setting of [1], and the reverse direction.
4. **Cross-condition.** Training on one operating condition and testing on each other condition (the full 4×4 matrix), isolating the effect of speed, torque, and radial-force changes.
5. **Leave-one-bearing-out (LOBO).** Thirty-two folds, each holding out every window of one physical bearing. LOBO answers the question that all other

protocols leave open: does the model generalise to unseen bearing individuals?

3.7 Phase 2: Hybrid Anomaly Detection and Few-Shot Diagnosis

3.7.1 Rationale and Architecture

Industrial reality inverts the data availability of the laboratory. Healthy operating data are abundant, labelled fault data are scarce or absent. Phase 2 therefore adopts a two-stage architecture [30]. Stage 1 learns the healthy signature only and flags deviating windows, answering the question whether something is wrong. Stage 2 is a compact classifier, trained on a deliberately small pool of labelled recordings, that assigns a fault type to flagged windows, answering what is wrong. The division of labour matches the label economy: detection requires no fault labels at all, and diagnosis requires only as many as an operator can realistically document.

3.7.2 Input Representation: A Negative Result First

The first Stage-1 design followed the common deep-learning recipe: reconstruct raw 32 ms waveform windows with LSTM-based autoencoders and score anomalies by reconstruction error. It failed, with AUROC values barely above chance. The diagnosis of this failure is instructive and is reported as a result in its own right (Section 4.10). Raw 64 kHz vibration is a stochastic, noise-like process, so an autoencoder cannot reconstruct even healthy waveforms; its error therefore measures window variance, which is dominated by the operating condition rather than bearing health. Follow-up analysis showed that even simple statistics fail to separate healthy from faulty at the raw-window level, and that dense log-spectra dilute the few informative bins among hundreds of noisy ones.

The successful representation is the engineered feature vector of Section 3.5: 0.5 s windows and 178 features, with two transformations applied. First, a signed logarithm

$$\tilde{x} = \text{sign}(x) \log(1 + |x|) \quad (3.16)$$

compresses the heavy-tailed feature scales so that no single feature dominates the reconstruction loss. Second, each feature is standardised per operating condition,

$$z = \frac{\tilde{x} - \mu_c}{\sigma_c}, \quad (3.17)$$

where μ_c and σ_c are the healthy-training statistics of the window’s own operating condition c . Conditioning on the operating condition is industrially legitimate, because speed, torque, and force are measured continuously on any inverter-fed drive, and it removes the dominant source of healthy-signature variance identified in the EDA.

3.7.3 Stage-1 Detector Set

Five detectors of increasing sophistication are trained exclusively on healthy windows. Comparing them separates the contribution of the representation from that of model capacity.

Mahalanobis distance. On the normalised features of Equation (3.17) the diagonal Mahalanobis distance reduces to the mean squared deviation from the healthy centroid, $A(\mathbf{z}) = \frac{1}{d} \sum_i z_i^2$. It involves no learning beyond the normalisation itself and serves as the floor that any learned detector must beat.

k -NN distance. The mean Euclidean distance to the five nearest healthy training windows. This local density estimate adapts to a multi-modal healthy manifold, which matters here because six different healthy bearings with different run-in histories populate the healthy class.

Dense autoencoder. A fully connected encoder-decoder with layer sizes $178 \rightarrow 128 \rightarrow 32 \rightarrow 128 \rightarrow 178$ trained to minimise the reconstruction error $\|\hat{\mathbf{z}} - \mathbf{z}\|^2$ on healthy windows, with early stopping on a healthy validation split. The 32-dimensional bottleneck forces the network to learn the correlation structure of healthy operation, so a window that violates that structure reconstructs poorly. The anomaly score is the reconstruction mean squared error.

Variational autoencoder. The probabilistic counterpart [28]. The encoder outputs the mean and variance of a latent Gaussian $q(\mathbf{z}_l | \mathbf{z})$, a latent sample is drawn with the reparameterisation trick, and the training loss is the negative evidence lower bound

$$\mathcal{L} = \underbrace{\|\hat{\mathbf{z}} - \mathbf{z}\|^2}_{\text{reconstruction}} + \beta \underbrace{D_{\text{KL}}(q(\mathbf{z}_l | \mathbf{z}) \parallel \mathcal{N}(0, I))}_{\text{regularisation}}, \quad (3.18)$$

with the weighting β following [29]. The KL term keeps the latent space smooth and centred, preventing trivial memorisation, and the anomaly score combines both terms of Equation (3.18). The log-variance output is clamped to a bounded range,

since faulty inputs far outside the healthy manifold otherwise produce numerical overflow in the exponential.

Temporal LSTM-VAE. The fifteen consecutive window-feature vectors of one recording form a sequence that an LSTM-based VAE encodes and reconstructs, scoring each recording rather than each window. The LSTM cell [23] maintains a cell state c_t regulated by forget, input, and output gates,

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad h_t = o_t \odot \tanh(c_t), \quad (3.19)$$

which lets the encoder accumulate evidence across the 4 s measurement. This model tests whether the temporal evolution of the feature vector carries health information beyond what averaging the window scores already captures.

3.7.4 Thresholding and Calibration

A detector becomes a decision rule only through a threshold. The default rule follows [24]: the threshold is the 99th percentile of the anomaly scores on a held-out healthy validation split, targeting a nominal 1% false-alarm rate. Because this choice proves decisive for the end-to-end system, a calibration study sweeps the percentile from 90 to 99.9 and selects the operating point that maximises the F1 of the binary flag decision on a calibration set consisting of the healthy validation windows and the few-shot labelled fault windows. Those fault windows are data that Stage 2 already uses for training, so the selection consumes no additional labels, and the test set is never touched during calibration.

3.7.5 Stage 2: Few-Shot Fault Classification

Stage 2 assumes a realistic label budget of twelve labelled recordings per fault class, drawn across bearings and operating conditions, with smaller budgets examined in the sensitivity analysis. Two classifiers are compared on the same budget. The first is a compact multilayer perceptron with layer sizes $178 \rightarrow 256 \rightarrow 128 \rightarrow 4$, dropout of 0.3 after each hidden layer, and a class-weighted cross-entropy loss

$$\mathcal{L} = - \sum_i w_{y_i} \log \hat{p}_{y_i}(\mathbf{z}_i), \quad (3.20)$$

trained with early stopping on a validation slice of the training pool. The capacity is deliberately small: with roughly one hundred and fifty windows per class, a larger

network would memorise the pool. The second classifier is XGBoost with the same configuration as in Phase 1, as the classical reference.

At inference the two stages compose as follows. A window whose Stage-1 score falls below the threshold is declared Healthy; otherwise the Stage-2 classifier assigns its label, and it may itself return Healthy, acting as a veto on borderline flags. A recording-level decision is obtained by majority vote over the fifteen windows of the recording.

3.7.6 Extended Analyses

Three analyses complete Phase 2. A severity analysis relates the Stage-1 anomaly score and detection rate to the documented extent-of-damage levels of Tables 3.3 and 3.5 and to the damage method, testing whether the score behaves as a condition indicator; the association is measured with Spearman’s rank correlation, which is appropriate because the severity levels are ordinal. A label-budget analysis retrains Stage 2 with $K \in \{1, 2, 4, 8, 12\}$ labelled recordings per class, three random draws each and an identical test set, to chart diagnosis quality against labelling effort. Finally, a seed-robustness study retrains every stochastic model with five random seeds so that all deep results are reported as mean plus or minus one standard deviation.

3.7.7 Implementation

All experiments are implemented in Python using scikit-learn [35] for the classical models, the XGBoost library [22], and PyTorch [36] for the neural models, with fixed random seeds throughout. Deep models were trained on CPU. The feature-space architectures of Phase 2 train in minutes, which is itself a practical argument for the chosen representation, since retraining on a new machine or after a maintenance event carries negligible cost.

Chapter 4

Results and Discussion

This chapter presents the experimental results in order of increasing protocol strictness for Phase 1 (Sections 4.1 to 4.5), followed by the Phase-2 anomaly detection and end-to-end system results (Sections 4.6 and 4.7), the severity and label-budget analyses (Sections 4.8 and 4.9), the representation study (Section 4.10), a robustness check (Section 4.11), and a synthesis with a comparison against published results (Sections 4.12 and 4.13). Unless stated otherwise, deep-model results are means over five random seeds and classical results are means over five grouped cross-validation folds.

4.1 Supervised Baselines under the Standard Protocol

Table 4.1 reports macro F1 for all model and modality combinations under pooled recording-grouped cross-validation, and Figure 4.1 visualises the comparison. Three observations stand out. First, XGBoost is the strongest model in every modality, which matches the general pattern of the tabular-learning literature [18, 22] and justifies its use as the reference model in the stricter protocols that follow. Second, the modality ranking is consistent across all four models: fused features beat vibration alone, which beats current alone. Third, and centrally for the internal-signals question, the current-only configuration is far from useless. XGBoost reaches a macro F1 of 0.961 without any vibration input, and fusion adds only 0.002 over vibration alone, indicating that under this lenient protocol the two modalities are largely redundant rather than complementary.

The differences between models are themselves informative. The instance-based k -NN collapses on current features (0.715): the current feature space is noisier,

Table 4.1: Supervised 4-class diagnosis, all operating conditions pooled: macro F1 (mean $\pm$ std over 5 recording-grouped CV folds).

Model	Vibration	Current	Fused
<i>k</i> -NN	0.993 $\pm$ 0.001	0.715 $\pm$ 0.008	0.996 $\pm$ 0.001
Random Forest	0.996 $\pm$ 0.001	0.946 $\pm$ 0.006	0.997 $\pm$ 0.001
SVM (RBF)	0.982 $\pm$ 0.004	0.846 $\pm$ 0.018	0.991 $\pm$ 0.002
XGBoost	0.996 $\pm$ 0.002	0.961 $\pm$ 0.002	0.998 $\pm$ 0.001

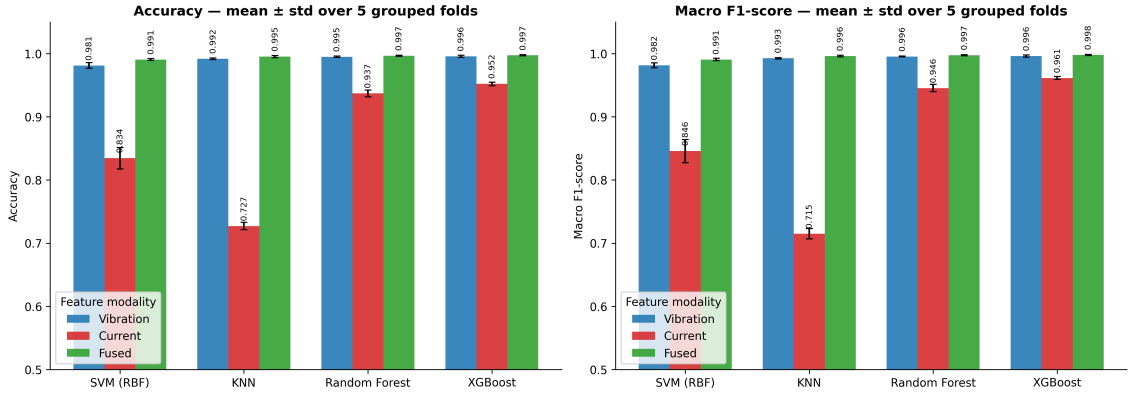


Figure 4.1: Accuracy and macro F1 for all model and modality combinations (mean $\pm$ std over five grouped folds). The modality ranking fused, then vibration, then current holds for every model.

so local neighbourhood geometry is less reliable, and a model that cannot weight features differently pays the full price for that noise. The SVM sits between the instance-based and the ensemble methods, and its larger fold-to-fold spread on current (0.018) reflects sensitivity of the margin to the noisy sideband features. The two tree ensembles are nearly insensitive to the feature noise because every split considers features individually, and boosting gains a further edge over bagging by focusing successive trees on the residual errors, which on this dataset concentrate in the Combined class.

Figure 4.2 shows the row-normalised confusion matrices for the held-out fold, and Figure 4.3 breaks the best model’s performance down by class. Two patterns recur across modalities. The Healthy class is recognised nearly perfectly by every configuration, which matters operationally because healthy misclassifications translate into false alarms. The Combined class is the weakest for the current modality: its windows are drawn from only three bearings whose inner-race and outer-race signatures superimpose, and in the current channel this superposition blurs the sideband pattern that distinguishes the two single-fault classes. The per-class view also reveals where fusion actually helps under this protocol: it repairs the current channel’s Combined-class weakness rather than improving the already saturated majority classes.

4.2 What the Models Learn: Feature Importance and SHAP

Before trusting any of the numbers above, it is worth verifying that the models rely on physically meaningful evidence rather than on artefacts of the rig or the recording process. Two complementary views are used. Figure 4.4 shows the gain-based XGBoost feature importances per modality. Envelope-spectrum amplitudes at the defect frequencies dominate the vibration model, and demodulated-current amplitudes at the same frequencies dominate the current model, exactly as the demodulation theory of Chapter 3 predicts. In the fused model both families appear, with vibration envelope features ranked first.

Gain-based importances can overstate features that appear early in trees, so the SHAP decomposition [19] provides a second, model-agnostic check. Figure 4.5 shows the mean absolute SHAP values of the fused XGBoost model over a test sample. The ranking agrees with the gain-based view in its leading entries, and the agreement between two attribution methods with different biases strengthens the conclusion that the classifier’s evidence is the physical fault signature. This

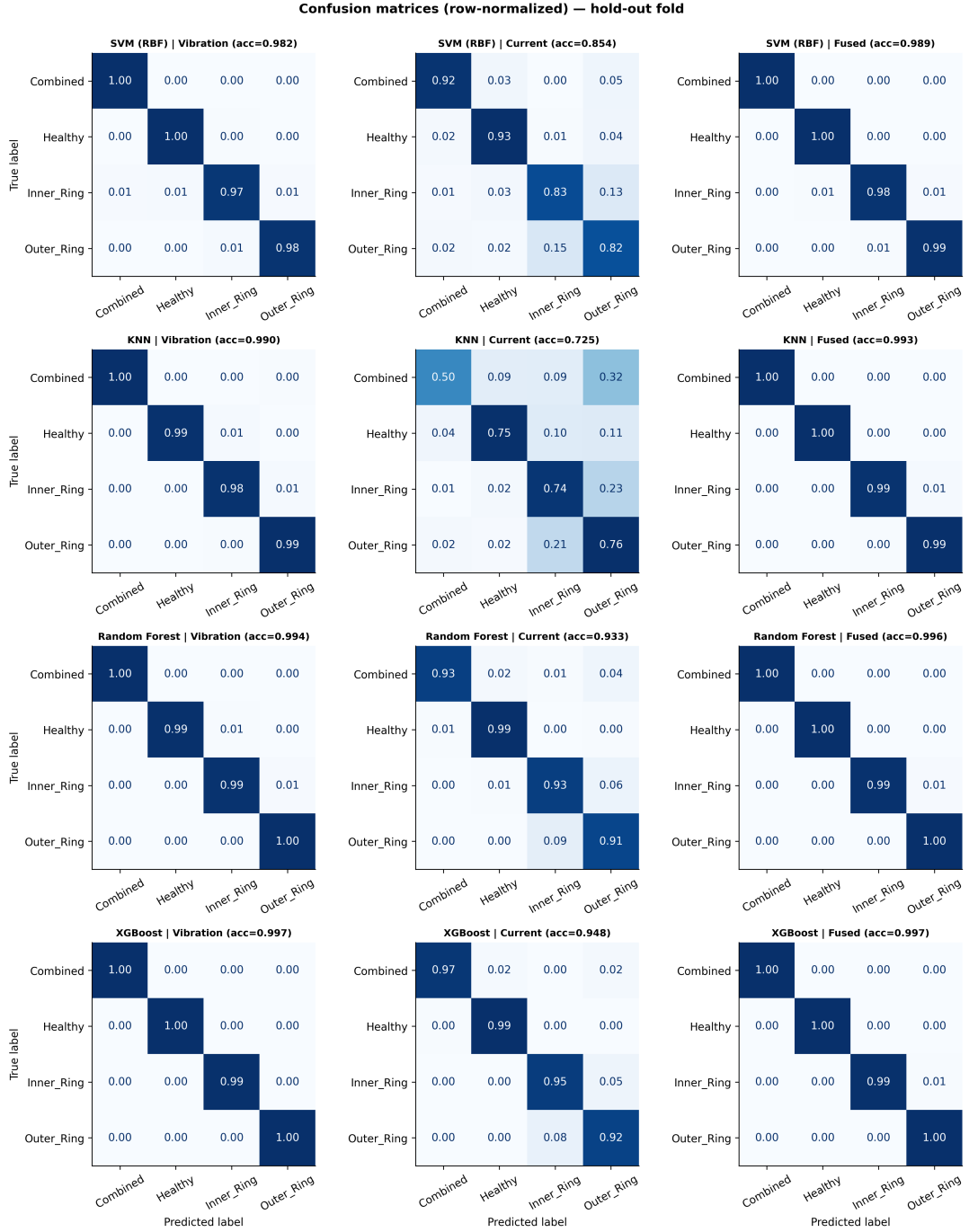


Figure 4.2: Row-normalised confusion matrices on the held-out fold for all model and modality combinations. The Combined class is the dominant error source, most visibly for the current modality.

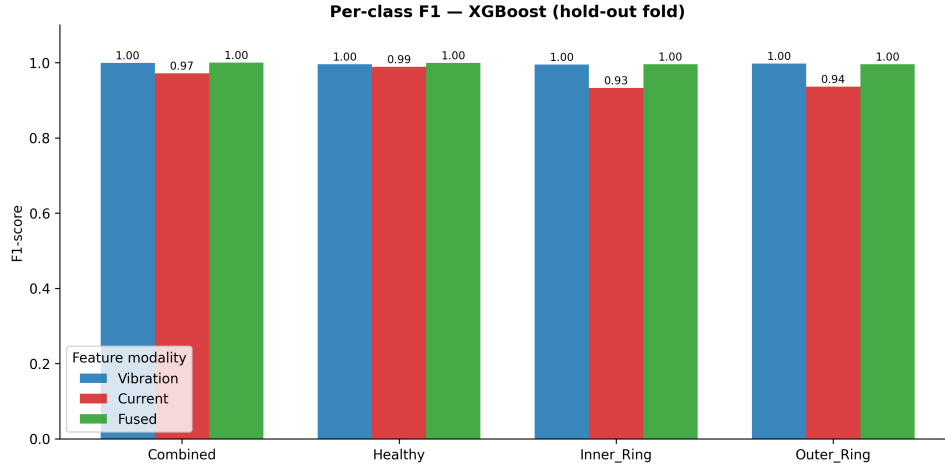


Figure 4.3: Per-class F1 of the best model per modality on the held-out fold. Fusion mainly repairs the Combined class of the current channel.

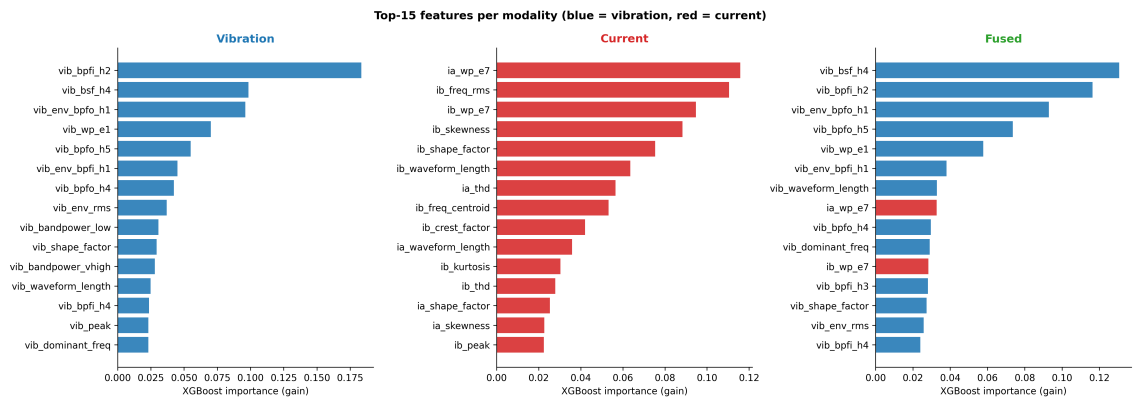


Figure 4.4: Top-15 XGBoost feature importances per modality (blue: vibration features; red: current features). The models select the physically expected envelope and sideband features.

interpretability result also has a practical use: when the model is deployed on a machine with a different bearing geometry, the important features indicate exactly which frequencies must be recomputed.

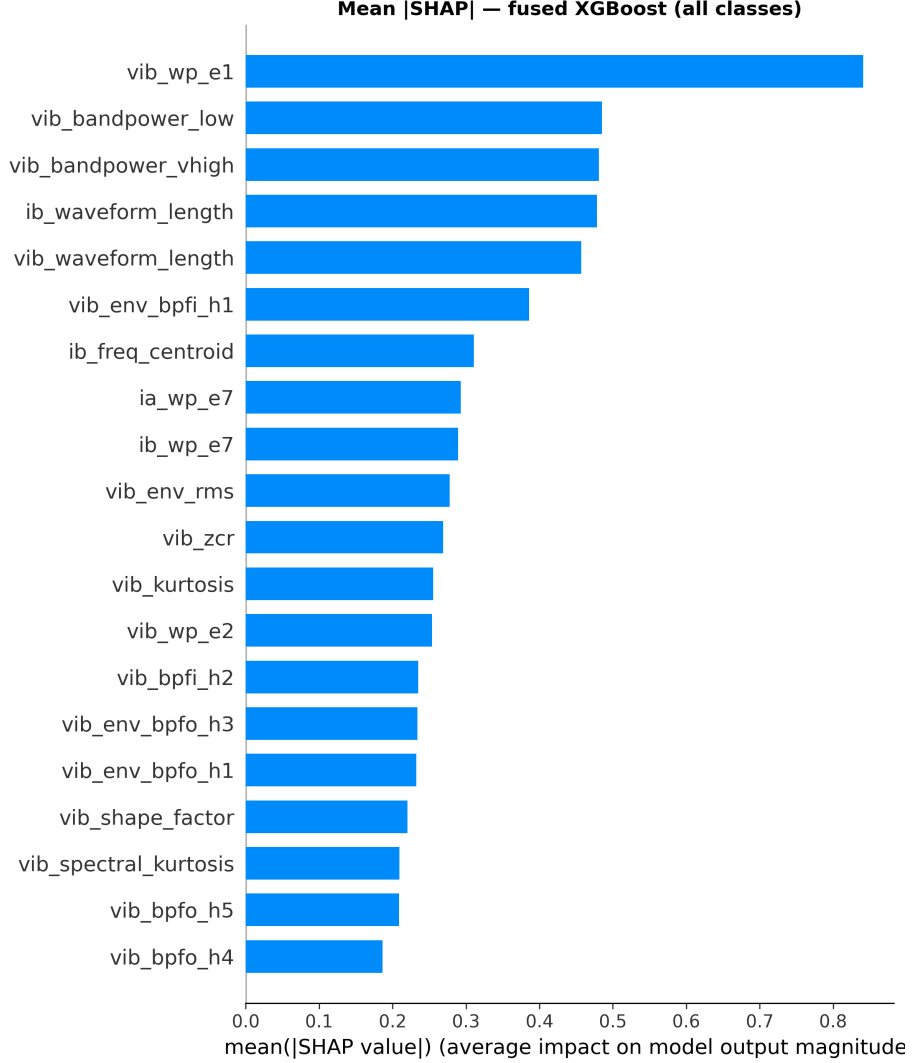


Figure 4.5: Mean absolute SHAP values of the fused XGBoost model. The attribution confirms the gain-based ranking: envelope and defect-frequency features carry the decisions.

4.3 Damage-Domain Scenarios and Transfer

Table 4.2 and Figure 4.6 decompose the pooled result along the damage-origin axis. Within either sub-domain the task is almost solved, even for current-only features, which reach 0.993 on artificial and 0.979 on naturally developed damage. The comparison of the two sub-domains also quantifies a claim made qualitatively in the

EDA: real damage is harder for the current channel (a drop of 0.014 from artificial to real for the best model) but not for vibration, whose envelope features resolve even the weak real-damage modulation once the training data contain examples of it.

Table 4.2: Damage-domain scenario study: best-model macro F1 per scenario and modality (recording-grouped splits).

Scenario	Vibration	Current	Fused
Artificial only (3-class CV)	0.996	0.993	1.000
Real only (3-class CV)	0.999	0.979	0.999
All pooled (4-class CV)	0.996	0.961	0.998
Transfer artificial $\rightarrow$ real	0.627	0.547	0.556
Transfer real $\rightarrow$ artificial	0.480	0.480	0.470

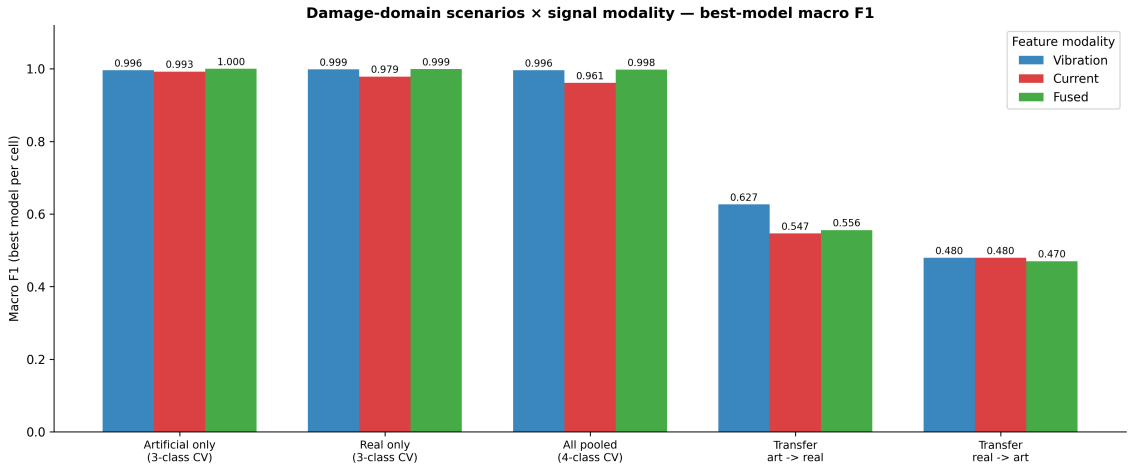


Figure 4.6: Best-model macro F1 across the five damage-domain scenarios. Within-domain diagnosis is nearly saturated; cross-domain transfer collapses for every modality.

The picture changes completely when the domains are crossed. Training on artificial damage and testing on real damage drops the best macro F1 to 0.627 for vibration, 0.556 for fusion, and 0.547 for current, and the reverse direction is harder still, between 0.47 and 0.48 for all modalities. Figure 4.7 shows that this collapse is not an artefact of one model: all four classifiers fall into the same range under transfer, and the model ranking that held within domains partially dissolves, which is typical when the failure is caused by a distribution shift rather than by insufficient model capacity.

Three aspects of the transfer result deserve discussion. First, it quantifies for the full modality grid what the original benchmark paper demonstrated for selected

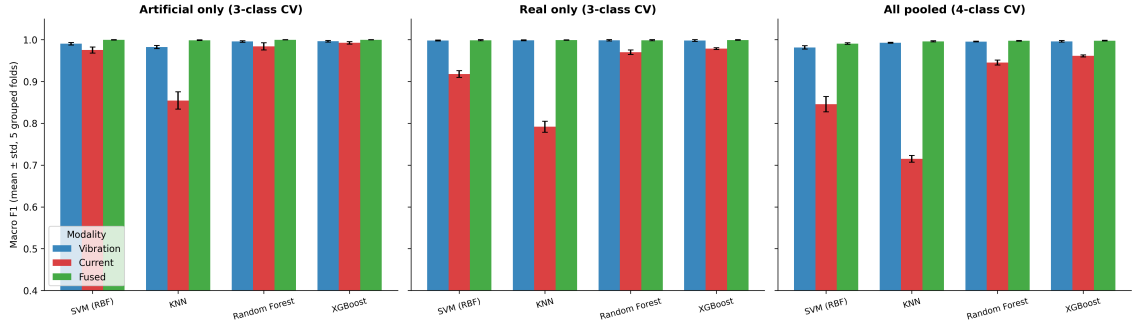


Figure 4.7: Macro F1 of all four models and three modalities in the artificial-only, real-only, and pooled scenarios (mean $\pm$ std over grouped folds). The within-domain conclusions do not depend on the choice of model.

classifiers [1]: models trained only on manufactured defects do not learn representations that transfer to genuine fatigue damage. Second, the failure is specific to diagnosis rather than detection. For a binary healthy-versus-faulty task, transfer is considerably more benign: a vibration SVM trained on artificial faults only reaches 96% accuracy and 0.957 recall on real faults. What breaks under transfer is the assignment of the fault location, not the recognition that something is wrong, and this distinction directly motivates the two-stage architecture of Phase 2, which places the transferable task (detection) in the label-free stage. Third, the transfer is asymmetric: the artificial-to-real direction outperforms the reverse by roughly 0.08 to 0.15 macro F1. The artificial sub-domain, with its three distinct machining methods, evidently spans a wider signature variety than the fatigue-dominated real sub-domain, so training on it provides broader coverage. The transfer confusion matrices of Figure 4.8 locate the failure precisely: real inner-race and combined damage are systematically mislabelled, consistent with their weak envelope signatures documented in the EDA.

4.4 Cross-Condition Generalisation

Table 4.3 presents the full train-condition by test-condition accuracy matrices for XGBoost, and Figure 4.9 the corresponding heatmaps. Averaged over the off-diagonal cells, accuracy falls from effectively perfect within-condition values to 0.613 for vibration, 0.548 for fusion, and 0.471 for current.

The structure of the matrices is physically interpretable, and reading it cell by cell is more informative than the averages. The most damaging shift is the speed change, visible as the uniformly low first row and first column of every matrix: defect frequencies and the electrical fundamental both scale with speed, so a model trained

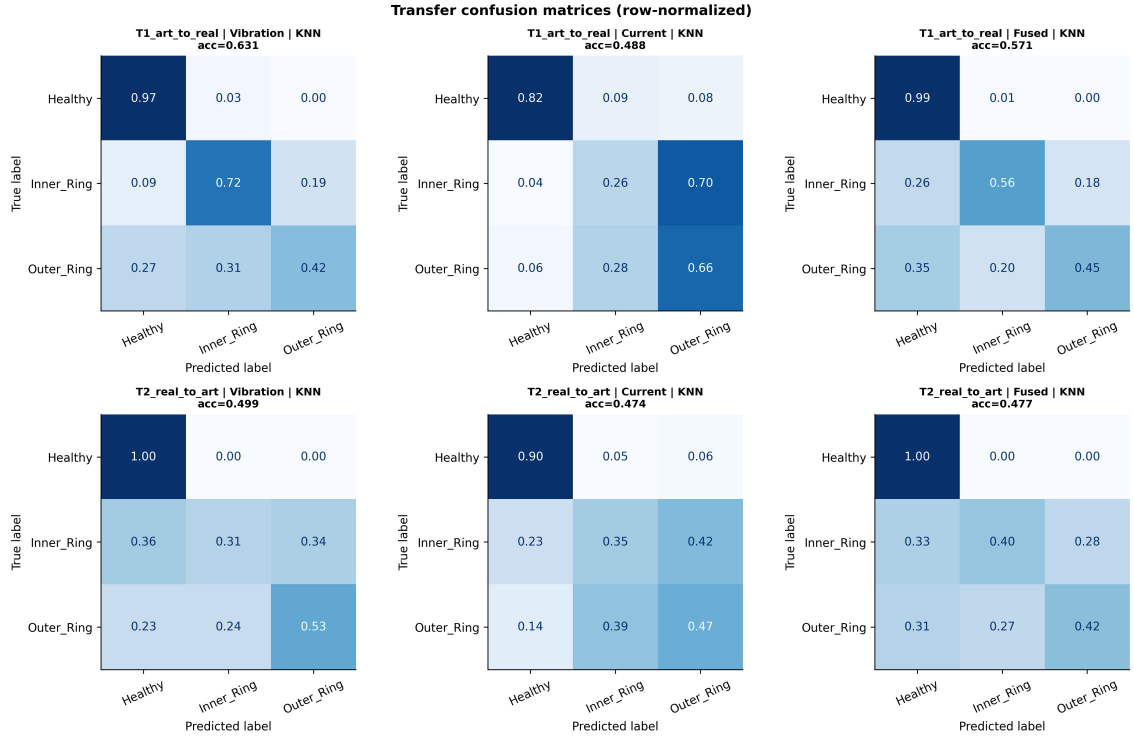


Figure 4.8: Row-normalised confusion matrices of the cross-domain transfer experiments (best model per direction). The diagnosis of naturally developed inner-race and combined damage breaks down after artificial-only training.

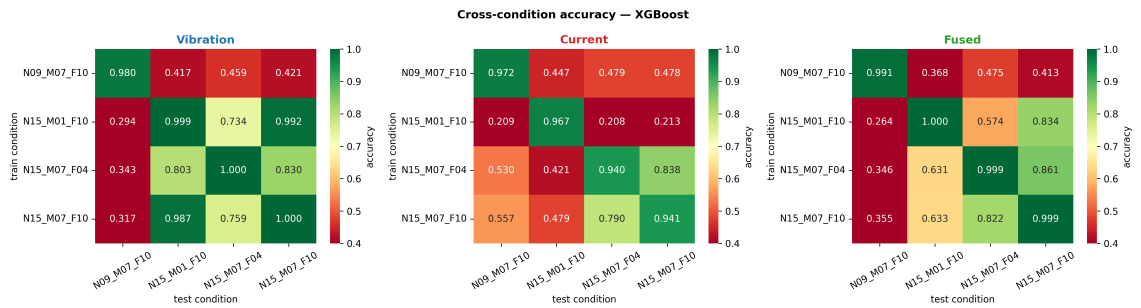


Figure 4.9: Cross-condition accuracy heatmaps per modality (XGBoost). Speed changes (first row and column) are the hardest shift for both modalities; torque changes additionally cripple the current channel.

Table 4.3: Cross-condition generalisation (XGBoost): accuracy when training on one operating condition (rows) and testing on another (columns).

Vibration (mean cross-condition accuracy: 0.613)				
Train \ Test	N09_M07_F10	N15_M01_F10	N15_M07_F04	N15_M07_F10
N09_M07_F10	0.980	0.417	0.459	0.421
N15_M01_F10	0.294	0.999	0.734	0.992
N15_M07_F04	0.343	0.803	1.000	0.830
N15_M07_F10	0.317	0.987	0.759	1.000
Current (mean cross-condition accuracy: 0.471)				
Train \ Test	N09_M07_F10	N15_M01_F10	N15_M07_F04	N15_M07_F10
N09_M07_F10	0.972	0.447	0.479	0.478
N15_M01_F10	0.209	0.967	0.208	0.213
N15_M07_F04	0.530	0.421	0.940	0.838
N15_M07_F10	0.557	0.479	0.790	0.941
Fused (mean cross-condition accuracy: 0.548)				
Train \ Test	N09_M07_F10	N15_M01_F10	N15_M07_F04	N15_M07_F10
N09_M07_F10	0.991	0.368	0.475	0.413
N15_M01_F10	0.264	1.000	0.574	0.834
N15_M07_F04	0.346	0.631	0.999	0.861
N15_M07_F10	0.355	0.633	0.822	0.999

at 1,500 rpm looks for spectral lines in the wrong places at 900 rpm, and the tolerance bands of the feature extraction cannot absorb a 40% frequency shift. Between the three equal-speed conditions the vibration model transfers well (0.73 to 0.99), confirming that its features are primarily kinematic. For the current modality the torque change is almost equally destructive: training on the low-torque condition and testing anywhere else yields barely 0.21, because current amplitude and modulation depth depend directly on load, and a model trained at 0.1 Nm has never seen the feature ranges that 0.7 Nm produces. The practical conclusion is that current-based diagnosis requires either training data spanning the load range or an explicit load normalisation, and the per-condition standardisation of Phase 2 implements exactly the latter.

4.5 Leave-One-Bearing-Out: The Evaluation-Protocol Effect

The protocols above group by recording but still allow different recordings of the same physical bearing on both sides of a split. Table 4.4 shows what happens when this last channel of leakage is closed: macro F1 of the same XGBoost model falls from the 0.96 to 0.998 range to between 0.43 and 0.57, a gap of 43 to 53 percentage points.

Table 4.4: Effect of the evaluation protocol on supervised diagnosis (XGBoost, macro F1): recording-grouped cross-validation versus leave-one-bearing-out (LOBO). The gap quantifies bearing-identity inflation.

Modality	Grouped-CV F1	LOBO F1	Gap
Vibration	0.996	0.509	0.488
Current	0.961	0.430	0.532
Fused	0.998	0.569	0.429

This is arguably the most consequential single result of the thesis. The evaluation protocol influences the reported number far more than the choice of model, feature set, or modality. Under conventional protocols the classifiers evidently learn, in large part, to recognise bearing individuals: each physical bearing carries a stable signature arising from manufacturing tolerances, mounting, and run-in history, and the model can exploit that signature whenever other recordings of the same bearing appear in training. Importantly, the modality ranking survives the protocol change, with fusion still ahead of vibration and vibration ahead of current, so the comparative conclusions of this thesis do not rest on the inflated numbers.

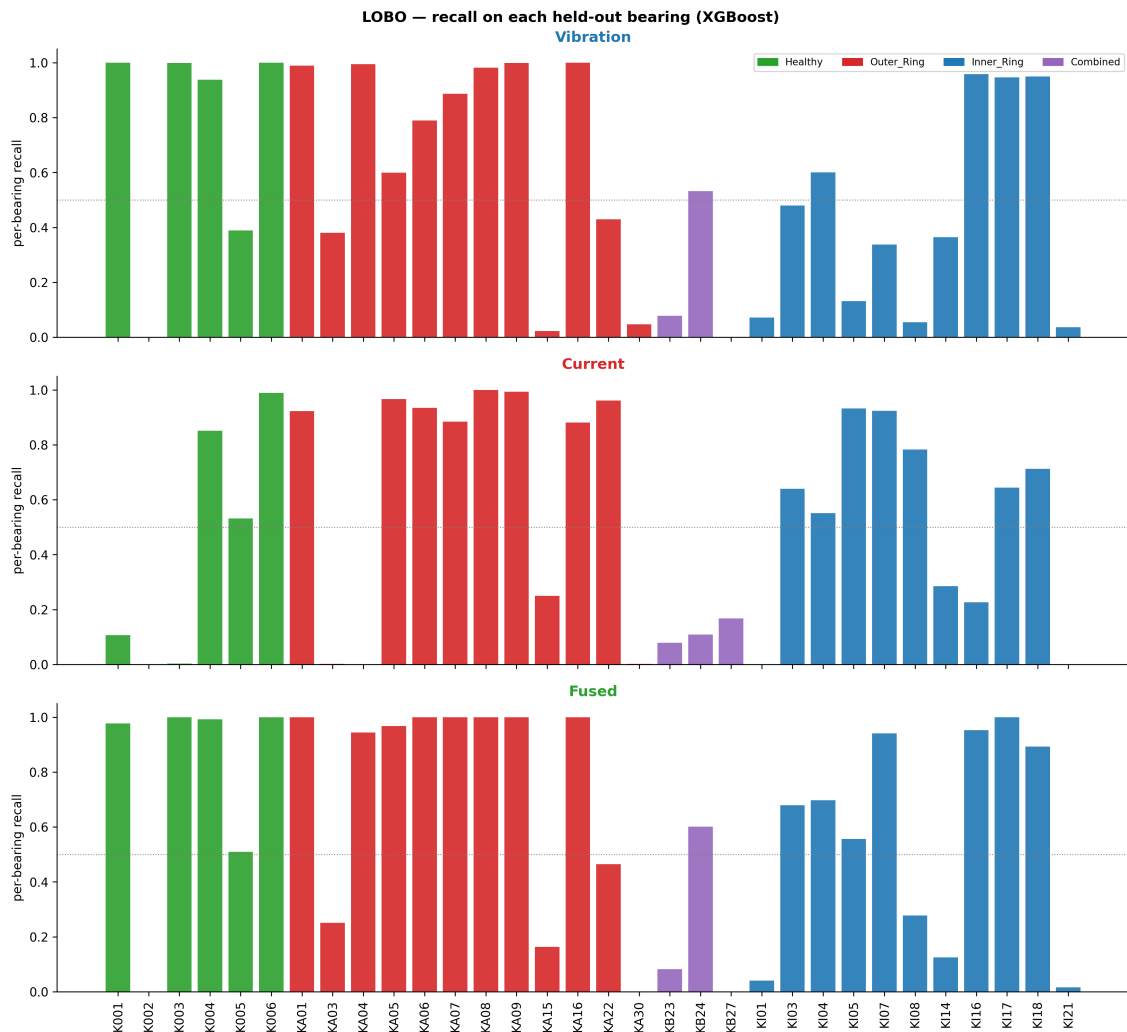


Figure 4.10: LOBO recall for each held-out bearing (colour = true class). Failures concentrate on specific individuals: KA30, KB27, and KA15, the plastic-deformation indentation bearings, and the healthy bearing K002 are diagnosed at or near zero recall.

The per-bearing breakdown of Figure 4.10 shows that LOBO failures are not uniform but concentrate on particular bearings, and the pattern is systematic. The three bearings damaged by plastic deformation (KA30, KB27, KA15) are diagnosed at recalls of 0.00, 0.00, and 0.16 respectively, while every EDM- and drilling-damaged bearing exceeds 0.94. Among the healthy references, bearing K002 is classified at zero recall, an early sign of the healthy-individuality effect that reappears in the Stage-1 analysis. Averaged per class, the fused model retains a recall of 0.73 on outer-race and 0.75 on healthy bearings but only 0.56 on inner-race and 0.23 on Combined bearings; the latter is expected, since the Combined class contains just three bearings and holding one out removes a third of the class’s training evidence. The pooled LOBO confusion matrices of Figure 4.11 summarise where the surviving accuracy comes from: outer-race faults remain identifiable, while Combined windows scatter into both single-fault classes. Section 4.8 returns to the failure pattern and shows that damage morphology rather than damage severity is the decisive factor.

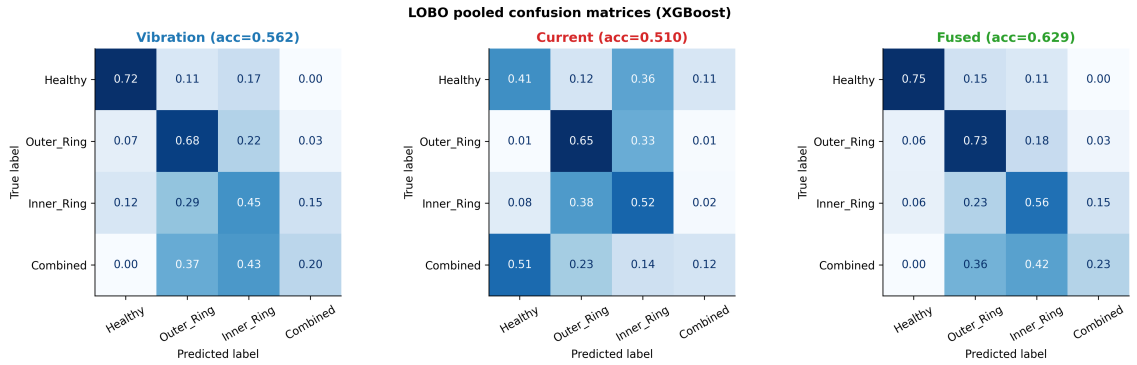


Figure 4.11: Pooled confusion matrices over all 32 LOBO folds per modality (XGBoost, row-normalised). The Combined class scatters into the single-fault classes once its bearings are unseen.

4.6 Stage 1: Healthy-Only Anomaly Detection

Tables 4.5 and 4.6 report the Stage-1 results at the window and recording level, and Figures 4.12 and 4.13 show the receiver operating characteristics and the score distributions.

Several conclusions follow. The representation carries most of the performance: even the learning-free Mahalanobis baseline reaches 0.899 on fused features, and the learned detectors add roughly nine AUROC points on top, mainly by modelling correlations that the diagonal baseline ignores. Among the learned detectors the dense autoencoder is the best overall, and the variational variant offers no accuracy

Table 4.5: Stage-1 anomaly detection, window level (0.5s): AUROC per detector and modality with healthy-only training. Deep models are reported as mean $\pm$ std over 5 seeds.

Detector	Vibration	Current	Fused
Mahalanobis	0.855	0.606	0.899
k -NN-5	0.931	0.923	0.988
Dense-AE	0.945 ± 0.005	0.917 ± 0.007	0.989 ± 0.002
Dense-VAE	0.937 ± 0.001	0.813 ± 0.006	0.981 ± 0.002

Table 4.6: Stage-1 anomaly detection, recording level (4s decision): AUROC per detector and modality.

Detector	Vibration	Current	Fused
Mahalanobis	0.857	0.610	0.908
k -NN-5	0.936	0.957	0.994
Dense-AE	0.953	0.984	0.998
Dense-VAE	0.954	0.899	0.996
LSTM-VAE (sequence)	0.950 ± 0.006	0.857 ± 0.026	0.984 ± 0.009

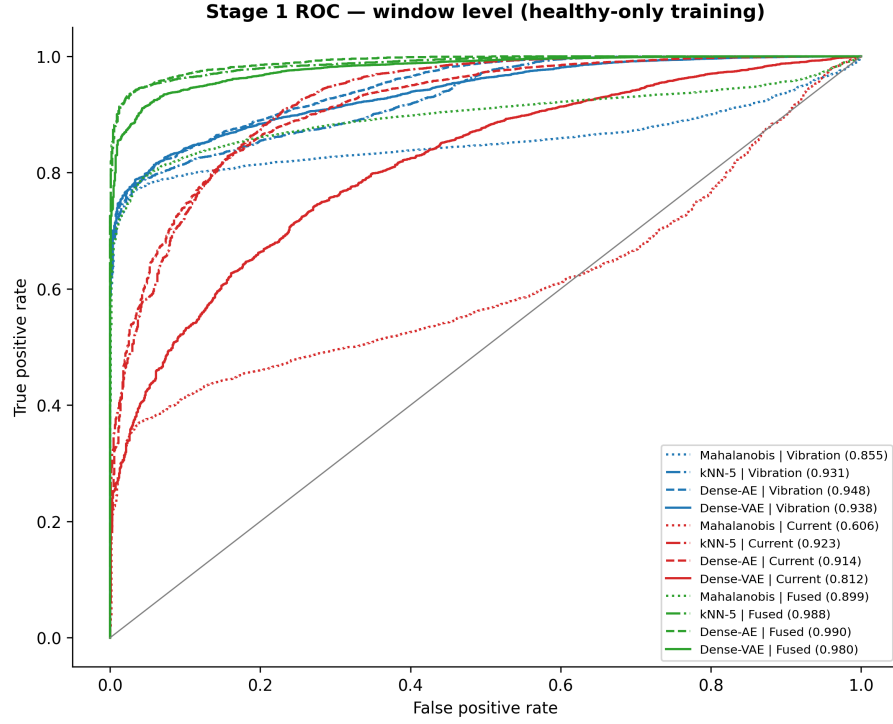


Figure 4.12: Receiver operating characteristics of all Stage-1 detectors and modalities at window level. Line style encodes the detector, colour the modality.

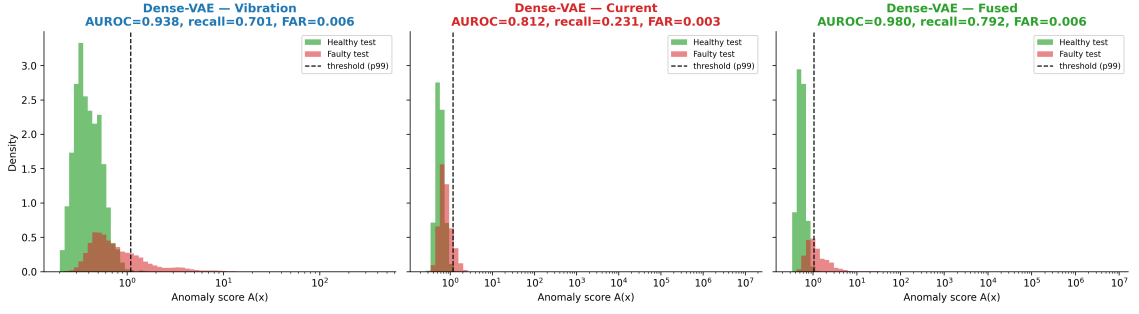


Figure 4.13: Anomaly-score distributions of the healthy test set (green) and faulty test set (red) per modality, with the 99th-percentile threshold. Fusion produces the cleanest separation.

advantage here. Its value lies elsewhere: the latent space of Figure 4.14 shows that fault classes organise themselves along coherent directions despite the model never having seen a fault label, which supports the interpretation that the learned healthy manifold captures the physically relevant structure of the feature space.

The strongest internal-signals result of the thesis appears at recording level: the current-only autoencoder separates faulty from healthy recordings with an AUROC of 0.984. A software-only monitoring layer on an unmodified drive can therefore detect bearing degradation with high reliability, and the breakdown of Figure 4.16 locates the residual misses in the naturally damaged inner-race bearings and the reduced-speed condition. Aggregating the fifteen window scores of a recording improves every detector (Figure 4.15), which is expected since averaging suppresses the score variance of individual windows. The temporal LSTM-VAE matches but does not exceed the aggregated dense models, indicating that the temporal evolution across a 4 s stationary measurement adds little beyond averaging; on non-stationary duty cycles this conclusion could change, which is noted as future work.

The LOBO stress test of Table 4.7 closes the Stage-1 analysis on an honest note. When the held-out healthy negatives come from a bearing the detector has never seen, fused AUROC drops from 0.989 to 0.862 with a standard deviation of 0.102 across folds, and the fold spread of Figure 4.17 reveals the cause. Healthy bearings themselves carry individual signatures, with bearing K002 again the pronounced outlier, so a threshold calibrated on five healthy individuals does not transfer cleanly to a sixth. The current modality suffers most, with fold AUROCs ranging from 0.20 to 0.75, because the electrical signature is more sensitive to the individual run-in history than the mechanical one. For deployment this implies that healthy-baseline data must be collected on the monitored machine itself. That is a mild requirement in practice, since healthy data are abundant by definition, but one that laboratory evaluations silently assume away.

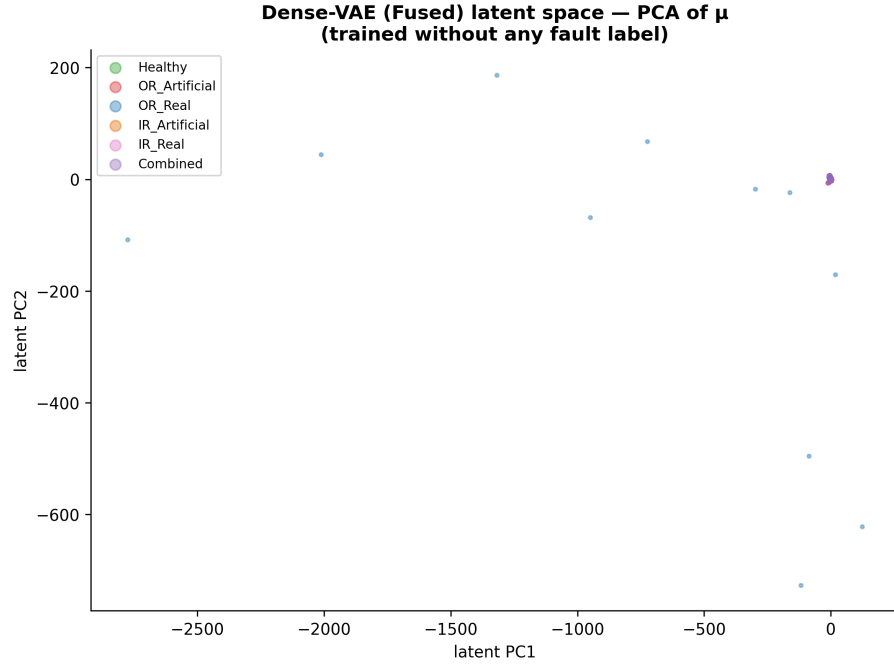


Figure 4.14: PCA projection of the fused VAE latent means, coloured by the (never seen) 6-class label. The healthy cluster separates, and fault sub-types organise without supervision.

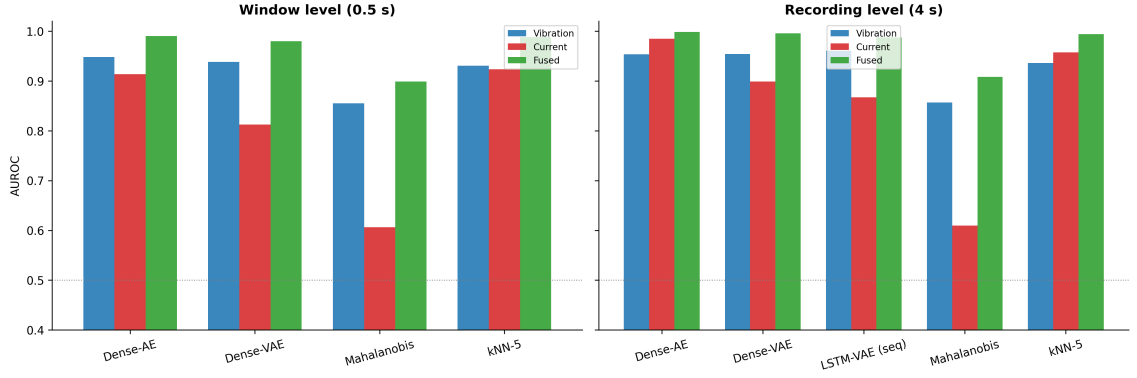


Figure 4.15: Window-level versus recording-level AUROC for all detectors and modalities. The 4s aggregated decision is uniformly stronger.

Table 4.7: Stage-1 detectors under leave-one-healthy-bearing-out: AUROC (mean $\pm$ std over 6 folds).

Detector	Vibration	Current	Fused
Mahalanobis	0.792 ± 0.041	0.477 ± 0.201	0.772 ± 0.091
k -NN-5	0.841 ± 0.083	0.622 ± 0.202	0.880 ± 0.122
Dense-AE	0.835 ± 0.077	0.461 ± 0.190	0.862 ± 0.102
Dense-VAE	0.861 ± 0.050	0.554 ± 0.181	0.875 ± 0.115

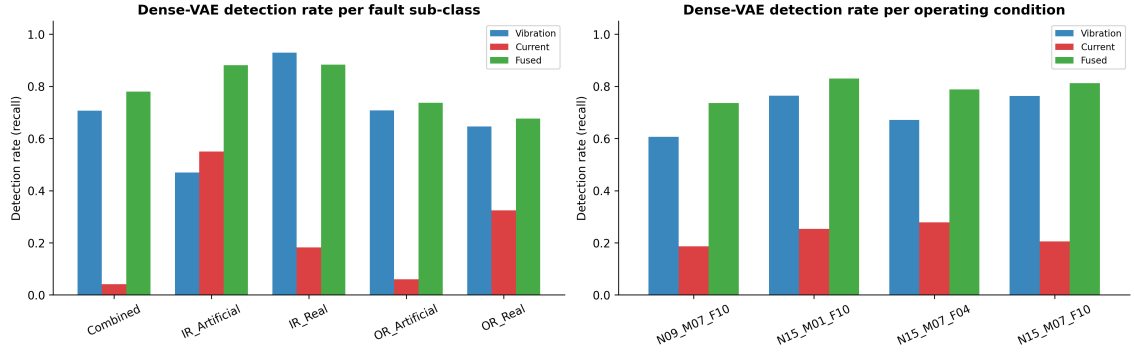


Figure 4.16: Detection rate of the fixed-threshold detector per fault sub-class and per operating condition. Naturally damaged inner-race bearings and the reduced-speed condition are the weakest spots.

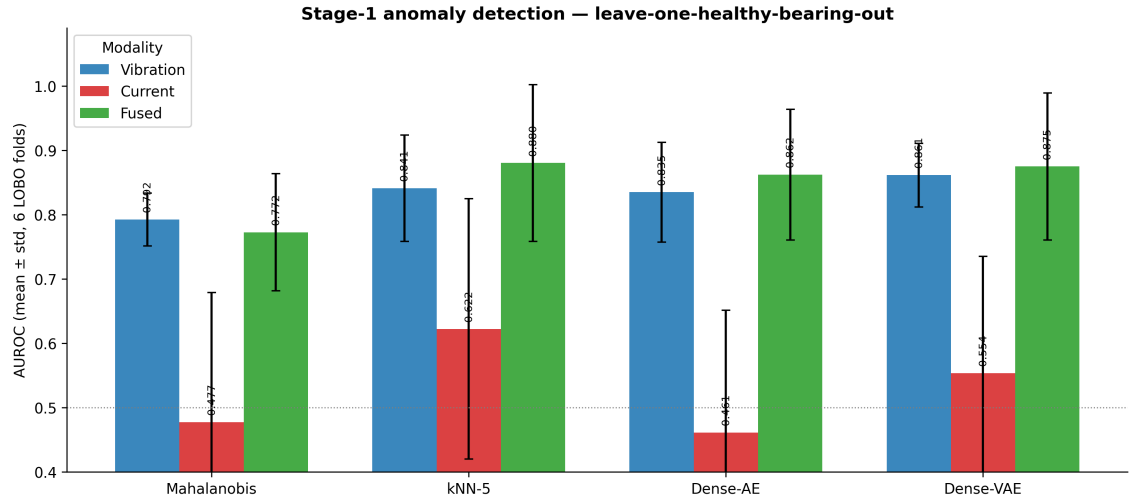


Figure 4.17: Stage-1 AUROC under leave-one-healthy-bearing-out (mean $\pm$ std over six folds). Healthy-bearing individuality widens the uncertainty substantially, especially for the current modality.

4.7 The Two-Stage System and Threshold Calibration

Table 4.8 evaluates the composed system at the default 99th-percentile threshold, Table 4.9 shows the effect of calibrating the threshold on the label-free calibration set of Section 3.7.4, and Figure 4.18 traces the full sweep.

Table 4.8: Few-shot Stage-2 diagnosis and end-to-end two-stage system: macro F1 (mean $\pm$ std over 5 seeds; 12 labelled recordings per fault class; default p99 threshold).

System	Vibration	Current	Fused
Stage-2 classifier alone (MLP)	0.714 ± 0.012	0.534 ± 0.010	0.747 ± 0.014
Two-stage system (p99)	0.544 ± 0.019	0.171 ± 0.027	0.592 ± 0.009

Table 4.9: Stage-1 threshold calibration: operating points selected on the calibration set (healthy-validation plus few-shot windows; no test labels used) and the resulting end-to-end macro F1.

Modality	Chosen pct.	Detection	FAR	E2E F1 (p99)	E2E F1 (calib.)	E2E F1 (recording)
Vibration	p97.5	0.757	0.024	0.524	0.534	0.520
Current	p90	0.721	0.085	0.180	0.430	0.446
Fused	p97	0.896	0.010	0.602	0.678	0.688

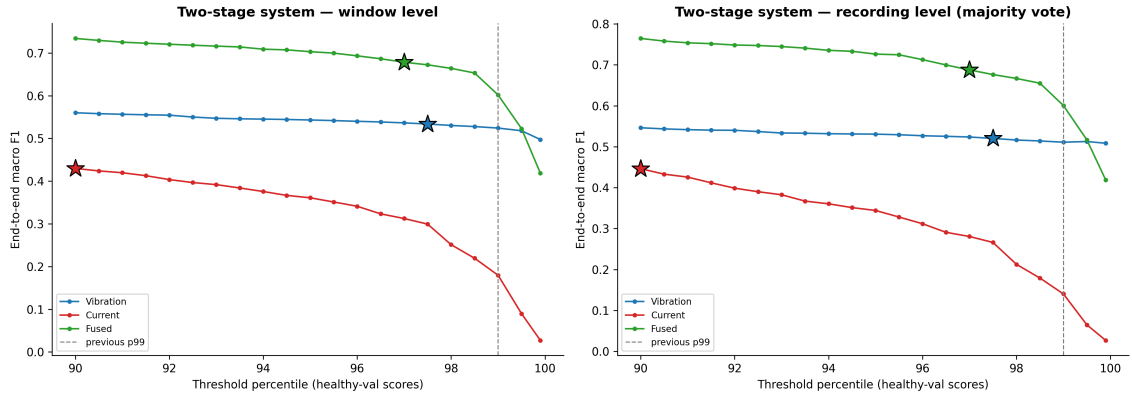


Figure 4.18: End-to-end macro F1 versus the Stage-1 threshold percentile, at window level (left) and recording level with majority voting (right). Stars mark the calibrated operating points; the dashed line marks the conventional p99 rule.

The default p99 rule, standard in the anomaly detection literature, turns out to be the bottleneck of the composed system. It suppresses false alarms so aggressively that a fifth of the genuinely faulty windows are forced to a Healthy verdict before

Stage 2 ever sees them, and for the current modality, whose score distributions overlap more, the effect is catastrophic (end-to-end F1 of 0.171). Calibrating the percentile on the label-free calibration set raises the fused end-to-end macro F1 from 0.602 to 0.678, and 0.688 at recording level with majority voting, and rescues the current-only system from 0.171 to 0.430. The calibration procedure is worth emphasising methodologically: it selects the operating point using only data the system already possesses (the healthy validation split and the Stage-2 training windows), so the gain is achievable in deployment without any additional labelling effort. Figure 4.19 shows the same sweep in detection-versus-false-alarm coordinates, which is the representation an operator would use to choose a point matching a specific alarm budget: at the calibrated operating point the fused Stage 1 flags 89.6% of faulty windows at a 1.0% false-alarm rate, and the composed confusion matrices of Figure 4.20 show where the remaining errors sit.

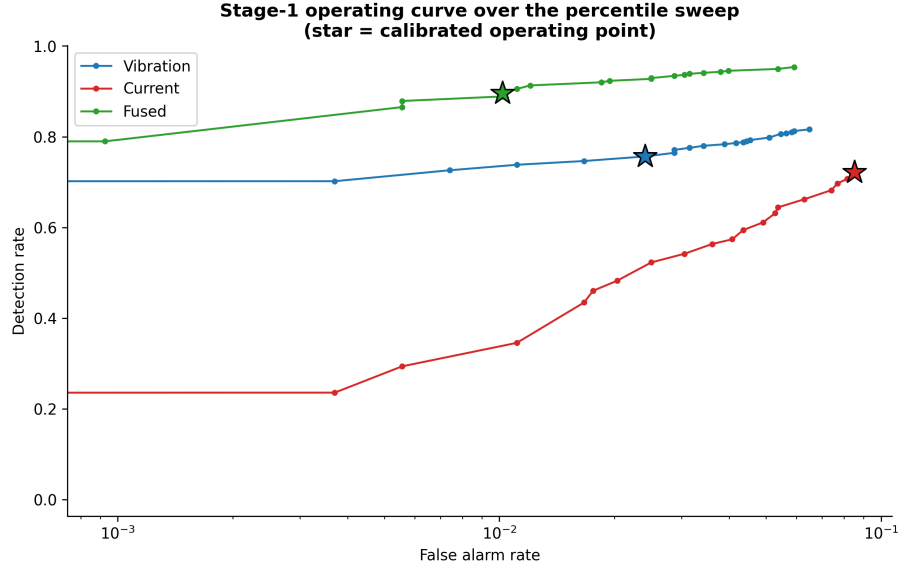


Figure 4.19: Detection rate versus false-alarm rate along the threshold sweep, per modality. Stars mark the calibrated operating points.

Two honest observations complete this section. First, the composed system does not exceed the Stage-2 classifier evaluated alone (0.747 fused). The value of the two-stage design lies not in higher F1 but in its label economy: Stage 1 provides a detection guarantee with zero fault labels, and Stage 2 adds diagnosis on top of a twelve-recording budget, whereas a single-stage supervised system with the same diagnostic accuracy would provide no detection capability at all until labels exist. Second, the choice of threshold is not a technical detail but an operating-point decision that deserves explicit treatment in any deployment, which is why the full sweep is reported rather than a single number.

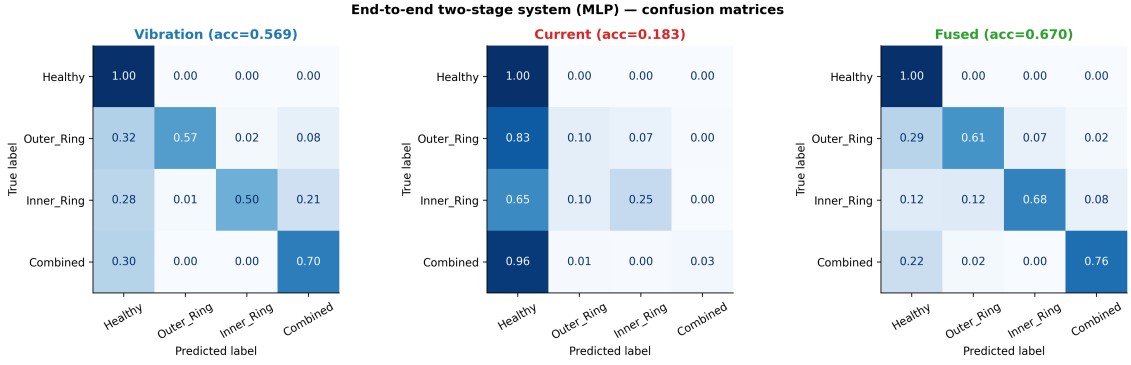


Figure 4.20: Row-normalised confusion matrices of the end-to-end two-stage system per modality (MLP Stage 2). Residual errors concentrate in the Combined class and in faulty windows suppressed by Stage 1.

4.8 Severity-Aware Evaluation

Table 4.10 relates the Stage-1 detector to the documented extent-of-damage levels, and Figure 4.21 shows the underlying score distributions. For the vibration and fused detectors the detection rate rises monotonically with the damage level, reaching 0.859, 0.936, and 1.000 for the fused detector at levels 1, 2, and 3, and the per-bearing median anomaly score correlates positively with the level (Spearman $\rho = 0.41$, $p = 0.038$ for vibration). Within the limits of a dataset that contains only two level-3 bearings, the anomaly score thus behaves as a rudimentary condition indicator. It does not merely flag damage, it grows with damage extent, which is the property required of a trending quantity in a prognostics setting.

Table 4.10: Severity-aware evaluation of the Stage-1 detector (Dense-AE, calibrated threshold): detection rate per extent-of-damage level and rank correlation between anomaly score and severity.

Extent of damage	Vibration	Current	Fused
Level 1	0.688	0.749	0.859
Level 2	0.817	0.731	0.936
Level 3	1.000	0.453	1.000
Spearman ρ (score vs level)	0.41 ($p=0.038$)	-0.21 ($p=0.296$)	0.37 ($p=0.064$)

The current modality breaks the monotone pattern at level 3, but the apparent anomaly dissolves on inspection. Level 3 contains only bearings KB24 and KI16, and KB24 carries distributed damage, whose weak modulation signature the current channel struggles with regardless of extent. This points to the deeper finding of Figure 4.22: joining the supervised LOBO recalls of Section 4.5 with the severity labels shows essentially no severity effect, and the catastrophic failures cluster by damage

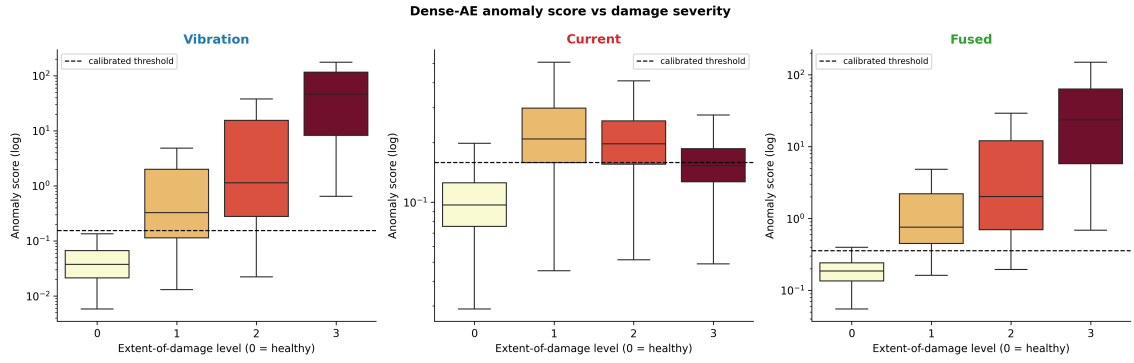


Figure 4.21: Anomaly-score distributions per extent-of-damage level (level 0 = healthy) and modality. The vibration and fused scores shift upward monotonically with the level.

method instead. All three plastic-deformation bearings are diagnosed at near-zero recall while EDM and drilled bearings are near-perfect at every level, and within the engraver-damaged bearings the level-2 specimens are actually diagnosed better than several level-1 pitting bearings. The practical lesson for dataset builders and model developers alike is that generalisation is limited by the morphological coverage of the training population, not by damage size. With only three indentation bearings in the dataset, holding one out removes a third of that morphology’s training evidence, and no amount of model capacity can substitute for the missing coverage.

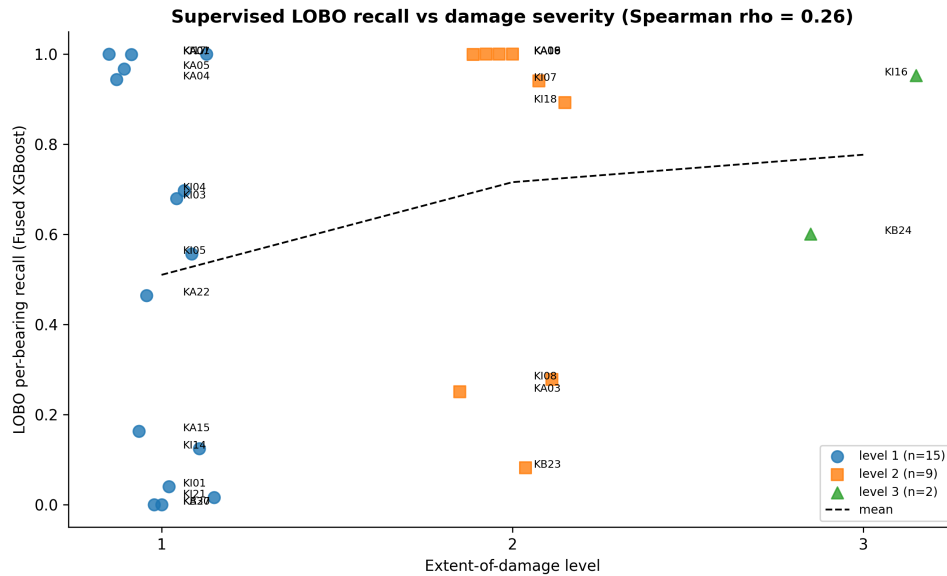


Figure 4.22: Supervised LOBO recall versus extent-of-damage level (fused XGBoost, one marker per bearing). Failures do not follow severity; they cluster on the plastic-deformation bearings KA30, KB27, and KA15.

4.9 Label-Budget Analysis

Table 4.11 and Figure 4.23 chart the Stage-2 macro F1 against the number of labelled recordings per fault class, with an identical test set at every budget so that the curves are directly comparable. Diagnosis quality degrades gracefully. A single labelled recording per class already yields 0.32 on fused features, four recordings 0.53, and twelve 0.75, with the curve still rising, so twelve recordings do not saturate the task and an operator who can document additional failures should expect further gains.

Table 4.11: Label budget: Stage-2 MLP macro F1 versus the number of labelled recordings per fault class (mean over 3 random draws; identical test set for every K).

Modality	$K=1$	$K=2$	$K=4$	$K=8$	$K=12$
Vibration	0.344	0.400	0.497	0.632	0.706
Current	0.224	0.346	0.368	0.489	0.532
Fused	0.322	0.392	0.525	0.677	0.749

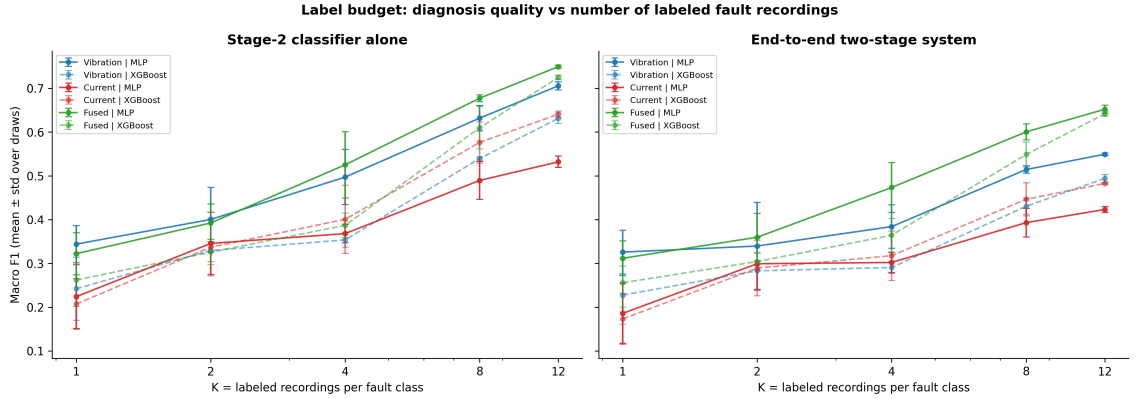


Figure 4.23: Diagnosis quality versus label budget: macro F1 of the Stage-2 classifier alone (left) and of the end-to-end system (right) for $K \in \{1, 2, 4, 8, 12\}$ labelled recordings per class (mean $\pm$ std over three draws).

Two secondary observations are of practical interest. The MLP is more label-efficient than XGBoost for budgets of four recordings or fewer, whereas the ensemble catches up and partially overtakes as data grow; regularised neural networks evidently extract more from very small pools, while boosting needs enough data to estimate its residuals reliably. The modality ranking of Phase 1 persists at every budget, and the draw-to-draw standard deviation shrinks from roughly 0.05 at a single recording to below 0.01 at twelve, so the case for signal fusion does not depend on label abundance and the small-budget numbers carry honest error bars.

4.10 Representation Study: Two Negative Results

Table 4.12 contrasts the raw-waveform deep models of the initial Phase-2 design with their feature-space successors. Both raw-waveform attempts failed decisively. Reconstruction-based detection on 32 ms waveform windows peaked at an AUROC of 0.685, barely above the chance level of 0.5 for most configurations, and a few-shot 1D-CNN on the same windows reached only 0.48 macro F1 where the feature-space MLP attains 0.75 on the identical label budget.

Table 4.12: Representation study (fused modality, best configuration of each approach): raw-waveform deep models versus their feature-space equivalents.

Approach	Metric	Result
Raw-waveform reconstruction (LSTM-VAE/AE, 32 ms)	AUROC	0.685
Feature-space reconstruction (Dense-AE, 0.5 s)	AUROC	0.990
Raw-waveform 1D-CNN, few-shot (32 ms)	macro F1	0.480
Feature-space MLP, few-shot (0.5 s)	macro F1	0.747

The failure mechanism was traced experimentally rather than assumed, in three steps. First, the score distributions of the raw-waveform autoencoders showed multi-modal structure aligned with the operating conditions rather than with health state, indicating that the reconstruction error was measuring the condition. Second, simple statistics confirmed the deeper problem: at the raw-window level neither RMS nor kurtosis separates healthy from faulty on this dataset, with AUROC near 0.53, so any score that reduces to signal energy is bound to fail. Third, replacing the waveform with a dense log-spectrum improved matters only marginally, because the handful of informative bins drowns among hundreds of noisy ones under a uniform reconstruction loss. The information that distinguishes a damaged bearing lives in narrow, physically specific spectral structures, namely envelope harmonics and sidebands, and extracting them requires both sufficient window length for frequency resolution and a representation that isolates them. This is precisely what the engineered feature vector provides.

These negative results are reported deliberately. High within-domain accuracies of raw-waveform networks, such as the 99% figures of [14], are real but describe a different and easier problem than the label-free, condition-robust setting studied here. The findings align with the growing critique of raw-waveform deep learning evaluated only under lenient protocols [9, 11] and document a methodological dead end that other practitioners can avoid.

4.11 Robustness to Random Initialisation

All deep-learning results reported above are means over five random seeds, and Figure 4.24 shows the underlying spread. The seed-to-seed variation is small throughout, with AUROC standard deviations below 0.01 for the dense models and below 0.03 for the LSTM-VAE, and macro-F1 standard deviations of 0.01 to 0.03 for the few-shot classifiers. None of the comparative conclusions of this chapter, whether the modality ranking, the detector ranking, or the protocol gaps, lies within initialisation noise. As a working rule, differences smaller than about twice the larger standard deviation of the configurations being compared are not interpreted in this thesis.

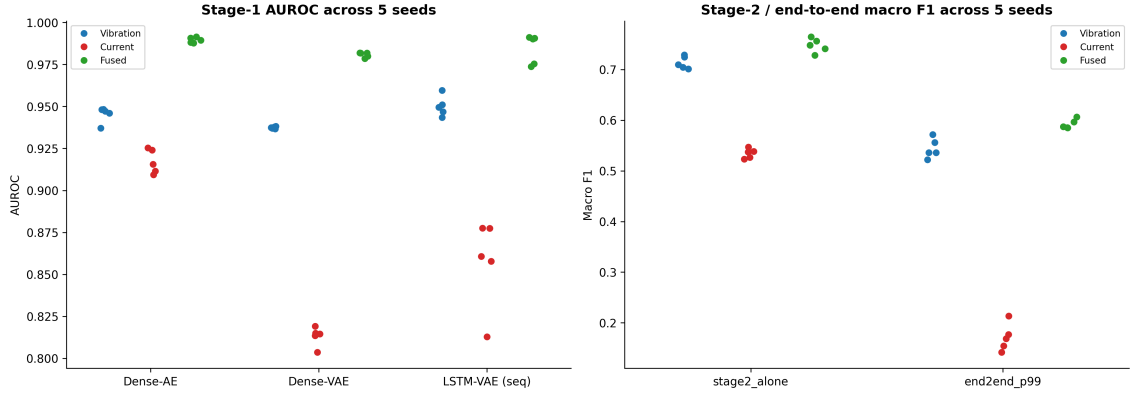


Figure 4.24: Seed-to-seed spread of the Stage-1 detectors (left, AUROC) and of the Stage-2 and end-to-end systems (right, macro F1) over five random seeds.

4.12 Comparison with Published Results

Table 4.13 places the main figures of this thesis next to the published results reviewed in Chapter 2. Direct numerical comparison must be made cautiously, because class definitions, window lengths, and split rules differ between studies; the table is therefore organised by evaluation setting rather than by metric alone.

Three points emerge. The within-domain and pooled results of this thesis are consistent with the high end of the literature, which confirms that the feature pipeline is competitive. The transfer results sit below the best published multi-branch and adaptation methods, which is expected since no adaptation mechanism was used here; the contribution of this thesis on that axis is the controlled measurement across modalities and the binary-versus-multiclass decomposition, which shows that detection transfers while diagnosis does not. Finally, the LOBO and healthy-only figures have no published counterparts on this benchmark known to

Table 4.13: Published results on the Paderborn benchmark and the corresponding figures of this thesis. Settings differ in class definitions and split rules; comparisons are indicative.

Study	Setting	Modality	Result
Lessmeier et al. [1]	artificial→real	current	60.9% acc.
Lessmeier et al. [1]	artificial→real	vibration	75.0% acc.
Tahan et al. [13]	artificial→real	vibration	72.7% acc.
Zhu et al. [33]	cross-domain, with adaptation	vibration	91.1% acc.
Firdausi, Ahmad [14]	within-domain	vibration	>99% acc.
This thesis	pooled grouped CV	fused	0.998 macro F1
This thesis	artificial→real (4-class)	vibration	0.627 macro F1
This thesis	artificial→real (binary)	vibration	96.0% acc.
This thesis	LOBO (bearing-level)	fused	0.569 macro F1
This thesis	healthy-only detection	current	0.984 AUROC

the author, and they are arguably the most deployment-relevant numbers in the table.

4.13 Synthesis

Read together, the results support four statements. First, the evaluation protocol, not the model, is the first-order determinant of reported performance on this benchmark; results quoted without a bearing-level protocol should be treated as upper bounds. Second, the internal-signals hypothesis holds with qualifications: motor current alone supports near-parity supervised diagnosis within a damage domain (0.961 against 0.996 macro F1) and strong label-free detection (recording-level AUROC of 0.984), but it is more sensitive than vibration to operating-condition changes and to distributed damage morphologies. Third, fusion of the two internal channels is uniformly the best configuration and costs nothing at inference time, since both signals are available in the drive cabinet. Fourth, the binding constraint on generalisation is the morphological diversity of the training population. That is a data problem rather than a model problem, and it argues for benchmark extensions with more distributed-damage specimens.

Chapter 5

Conclusions and Future Work

5.1 Summary of the Work

This thesis set out to determine how far the signals already present in a standard drive cabinet, namely the motor phase currents optionally complemented by a housing accelerometer, can carry bearing fault detection and diagnosis, and to do so under evaluation protocols strict enough that the answers survive contact with industrial reality. A 178-dimensional multi-domain feature representation was engineered for half-second windows of the Paderborn benchmark signals. Four classical classifiers were benchmarked across a five-protocol evaluation suite, and a two-stage hybrid architecture was developed in which healthy-only anomaly detectors gate a few-shot fault classifier. The evaluation was extended along two axes rarely reported for this benchmark, damage severity and label budget. All experiments are reproducible from fixed seeds, and results are reported with fold- or seed-level uncertainty.

5.2 Principal Findings

1. **Protocol dominates model.** Pooled recording-grouped cross-validation yields macro F1 up to 0.998; the same model under leave-one-bearing-out evaluation achieves 0.569. The gap of 43 to 53 points measures how much conventional evaluation rewards recognition of bearing individuals rather than fault physics, and it exceeds every model-to-model and modality-to-modality difference observed in this work.
2. **Current-only monitoring is viable, with stated limits.** Within a damage domain, current-only supervised diagnosis is competitive (macro F1 of 0.961), and a healthy-only autoencoder detects faulty recordings from current

features at an AUROC of 0.984. The current channel degrades disproportionately under load changes and for distributed damage morphologies, and its healthy-bearing individuality is stronger than that of vibration.

3. **Fusion is uniformly best and operationally free.** The fused configuration wins under every protocol, reaching 0.998 pooled macro F1, 0.998 recording-level detection AUROC, and the best transfer and end-to-end figures. Since both signals are available in the drive cabinet, fusion carries no additional hardware cost.
4. **The artificial-to-real gap is a diagnosis problem, not a detection problem.** Binary detection transfers well from artificial training to real faults (96% accuracy), but four-class diagnosis collapses to between 0.55 and 0.63 macro F1. Failures concentrate on plastic-deformation damage, and severity plays essentially no role.
5. **The anomaly score carries severity information.** Detection rate rises monotonically with the documented extent-of-damage level, and the per-bearing score correlates with severity ($\rho = 0.41$ for vibration), a first step from detection towards condition indication.
6. **Feature-space learning beats raw-waveform learning under realistic constraints.** Two documented negative results show that raw 64 kHz waveforms defeat both reconstruction-based detection and few-shot convolutional classification, while the same model classes succeed on engineered features.

5.3 Contributions

The thesis contributes, first, a fully controlled vibration-versus-current comparison on the Paderborn benchmark, with one feature pipeline, one model set, and one protocol suite for both modalities. Second, a quantification of evaluation-protocol inflation, recording-grouped versus bearing-level, for both the supervised and the one-class paradigm. Third, evidence that label-free, current-only bearing-fault detection is feasible at recording level. Fourth, a severity- and morphology-resolved error analysis identifying distributed plastic-deformation damage as the binding generalisation constraint. Fifth, a label-budget characterisation of the two-stage architecture, including a threshold calibration procedure that uses no test labels. Sixth, two documented negative results on raw-waveform deep learning that delimit where such models are appropriate.

5.4 Limitations

The study is bounded by the dataset: one bearing type, one test rig, four stationary operating conditions, and only three distributed-damage specimens, which is the very category identified as hardest. Level-3 severity conclusions rest on two bearings. All deep models were sized for CPU training, so larger architectures and longer training schedules were not explored. The two-stage system was evaluated on stationary 4s recordings, and transient regimes such as run-up and load steps remain untested. Finally, the threshold calibration relies on the few-shot fault windows, so a deployment with truly zero fault labels would have to fall back to the percentile rule and accept the corresponding loss.

5.5 Future Work

Four directions follow naturally from the findings. First, domain-adaptation methods, for example conditional adversarial alignment [32, 33], applied to the artificial-to-real and cross-condition gaps quantified here, with the present numbers as the baselines to beat. Second, severity regression and trending: the monotone score-severity relation suggests training explicitly for extent estimation and validating score trending on run-to-failure data. Third, morphology-aware data strategies, such as targeted augmentation or physics-informed simulation of distributed damage, to close the indentation gap identified in the LOBO analysis. Fourth, online deployment: porting the feature pipeline and the calibrated two-stage system to an embedded platform in a drive cabinet, where the current-only configuration evaluated here would require no additional hardware whatsoever.

Bibliography

- [1] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, “Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: a benchmark dataset for data-driven classification,” in *Proceedings of the European Conference of the PHM Society*, vol. 3, 2016.
- [2] A. H. Bonnett and C. Yung, “Increased efficiency versus increased reliability,” *IEEE Industry Applications Magazine*, vol. 14, no. 1, pp. 29–36, 2008.
- [3] McKinsey Global Institute, “The internet of things: Mapping the value beyond the hype,” <https://www.mckinsey.com/mgi>, 2015.
- [4] R. B. Randall, *Vibration-based Condition Monitoring*. Chichester, UK: Wiley, 2011.
- [5] J. Antoni, “The spectral kurtosis: a useful tool for characterising non-stationary signals,” *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 282–307, 2006.
- [6] M. E. H. Benbouzid, “A review of induction motors signature analysis as a medium for faults detection,” *IEEE Transactions on Industrial Electronics*, vol. 47, no. 5, pp. 984–993, 2000.
- [7] R. R. Schoen, T. G. Habetler, F. Kamran, and R. G. Bartheld, “Motor bearing damage detection using stator current monitoring,” *IEEE Transactions on Industry Applications*, vol. 31, no. 6, pp. 1274–1279, 1995.
- [8] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, “Applications of machine learning to machine fault diagnosis: a review and roadmap,” *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, 2020.
- [9] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, “Deep learning algorithms for bearing fault diagnostics—a comprehensive review,” *IEEE Access*, vol. 8, pp. 29 857–29 881, 2020.

- [10] J. Singh, M. Azamfar, F. Li, and J. Lee, “A systematic review of machine learning algorithms for prognostics and health management of rolling element bearings,” *Measurement Science and Technology*, vol. 32, p. 012001, 2021.
- [11] W. A. Smith and R. B. Randall, “Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study,” *Mechanical Systems and Signal Processing*, vol. 64–65, pp. 100–131, 2015.
- [12] S. Ayankoso, A. Dutta, Y. He, F. Gu, A. Ball, and S. K. Pal, “Performance of vibration and current signals in the fault diagnosis of induction motors using deep learning and machine learning techniques,” *Structural Health Monitoring*, 2024.
- [13] M. Tahan, D. C. Nguyen, T. H. Nguyen *et al.*, “Multiple learning strategies for motor bearing fault detection,” *arXiv preprint arXiv:2310.11477*, 2024.
- [14] M. Firdausi and S. Ahmad, “Concise convolutional neural network model for fault detection,” *Communications in Science and Technology*, vol. 7, no. 1, pp. 62–72, 2022.
- [15] R. B. Randall, J. Antoni, and S. Chobsaard, “The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults,” *Mechanical Systems and Signal Processing*, vol. 15, no. 5, pp. 945–962, 2001.
- [16] Z. K. Peng and F. L. Chu, “Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography,” *Mechanical Systems and Signal Processing*, vol. 18, no. 2, pp. 199–221, 2004.
- [17] D. Goyal, Vanraj, B. S. Pabla, and S. S. Dhimi, “Condition monitoring parameters for fault diagnosis of fixed axis gearbox: a review,” *Archives of Computational Methods in Engineering*, vol. 24, pp. 543–556, 2017.
- [18] L. Wang and M. Wu, “Multi-domain feature system with SHAP interpretability for bearing fault diagnosis,” *Scientific Reports*, vol. 15, p. 41242, 2025.
- [19] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [20] L. B. Jack and A. K. Nandi, “Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms,” *Mechanical Systems and Signal Processing*, vol. 16, no. 2–3, pp. 373–390, 2002.

- [21] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [22] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD*, 2016, pp. 785–794.
- [23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [24] A. B. Sulaimon, S. J. Oladebo, M. K. Idris, and F. A. Enesi, “Development of anomaly detector for motor bearing condition monitoring using FFT and LSTM-Autoencoder,” *i-manager’s Journal on Pattern Recognition*, vol. 10, no. 1, 2023.
- [25] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: a survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.
- [26] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
- [27] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” in *Proceedings of the 8th IEEE ICDM*, 2008, pp. 413–422.
- [28] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *International Conference on Learning Representations (ICLR)*, 2014.
- [29] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-VAE: learning basic visual concepts with a constrained variational framework,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [30] G. Nicchiotti, L. Etienne, and L. Fromaigeat, “Anomaly detection and classification for rotating machinery,” *PHM Society European Conference*, vol. 3, 2016.
- [31] Y. Ganin *et al.*, “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.
- [32] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional adversarial domain adaptation,” in *NeurIPS*, vol. 31, 2018.

- [33] C. Zhu, W. Luo, L. Fu, R. He *et al.*, “CASSA: Conditional-adversarial alignment with structured pseudo-labeling,” *Scientific Reports*, vol. 15, p. 42146, 2025.
- [34] P. D. Welch, “The use of fast Fourier transform for the estimation of power spectra,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.
- [35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga *et al.*, “PyTorch: an imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.