University of Pavia

Faculty of Engineering
Department of Electrical, Computer and Biomedical Engineering

Master's Thesis in Computer Engineering

Design and Development of a 1D Conditional Diffusion Framework Tailored for High-Fidelity Vibration Synthesis Targeting Zero-Shot Fault Diagnostics

Supervisor:
Prof. Tullio Facchinetti

Candidate:
Giulia Trespiolli

Academic Year 2025/2026

Abstract

In industrial diagnostics, the performance of deep learning models is frequently limited by the scarcity of mechanical fault data and severe class imbalance.

To overcome the intrinsic vulnerabilities of traditional generative architectures like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), such as mode collapse and oversmoothing, this thesis presents a novel 1D conditional generative framework developed from scratch based on Denoising Diffusion Probabilistic Models (DDPMs).

The proposed framework synthesises high-fidelity vibration signals corresponding to various types and severities of Rolling Element Bearing (REB) faults, as well as dynamic eccentricity faults, thereby demonstrating the framework's versatility across heterogeneous fault signatures. A major achievement of this work is the accurate zero-shot synthesis of operating conditions never encountered during training. The continuous embedding of physical parameters enables generalisation to unseen combinations, while Classifier-Free Guidance (CFG) is employed to increase adherence to the target conditioning during inference.

The model's effectiveness relies on key architectural choices, primarily the integration of local and global self-attention mechanisms across the U-Net, coupled with a dual-stage Adaptive Group Normalisation (AdaGN) within each residual convolutional block to ensure robust conditioning. Furthermore, generation quality is enhanced by replacing the standard Mean Squared Error (MSE) loss with a novel hybrid loss function that includes a spectral component, forcing the model to respect frequency-domain characteristics while synthesising time-domain signals.

The framework is validated on two datasets: CWRU, comprising signals from Direct-on-Line (DOL) induction motors, and EECSL, featuring inverter-driven motors. The practical utility of the synthetic data is evaluated through a Train on Synthetic – Test on Real (TSTR) scenario, where a downstream Convolutional Neural Network (CNN) classifier trained exclusively on synthetic data achieves diagnostic performance comparable to that of a classifier trained on real signals.

Contents

List of Figures

List of Tables

Acronyms

1-NN 1-Nearest Neighbour

AC Alternating Current

ACGAN Auxiliary Classifier Generative Adversarial Network

AdaBN Adaptive Batch Normalisation

AdaGN Adaptive Group Normalisation

AI Artificial Intelligence

BN Batch Normalisation

BPFI Ball Pass Frequency Inner race

BPFO Ball Pass Frequency Outer race

BSF Ball Spin Frequency

CFG Classifier-Free Guidance

CNN Convolutional Neural Network

CPU Central Processing Unit

CWRU Case Western Reserve University

CWT Continuous Wavelet Transform

D2CGAN Dual Discriminator Conditional Generative Adversarial Network

DDPM Denoising Diffusion Probabilistic Model

DE Drive End

DOL Direct-on-Line

EDM Electric Discharge Machining

EECSL Electrical Energy Conversion System Lab

ELBO Evidence Lower Bound

EMI Electromagnetic Interference

FC Frequency Centroid

FE Fan End

FFT Fast Fourier Transform

FIR Finite Impulse Response

FPGA Field-Programmable Gate Array

FTF Fundamental Train Frequency

GAN Generative Adversarial Network

GN Group Normalisation

GPU Graphics Processing Unit

HP Horsepower

IIR Infinite Impulse Response

KDE Kernel Density Estimation

KL Kullback-Leibler

LSD Log-Spectral Distance

MLP Multi-Layer Perceptron

MAE Mean Absolute Error

MSE Mean Squared Error

OOD Out-of-Distribution

PIDM Physics-Informed Diffusion Model

PWM Pulse Width Modulation

RAM Random Access Memory

REB Rolling Element Bearing

ReLU Rectified Linear Unit

RF Receptive Field

RMS Root Mean Square

RMSF Root Mean Square Frequency

RVF Root Variance Frequency

SCS Spectral Cosine Similarity

SiLU Sigmoid Linear Unit

SNR Signal-to-Noise Ratio

SSH Secure Shell

t-SNE t-Distributed Stochastic Neighbour Embedding

TPE Tree-structured Parzen Estimator

TSTR Train on Synthetic – Test on Real

VAE Variational Autoencoder

VFD Variable Frequency Drive

VPN Virtual Private Network

VRAM Video Random Access Memory

WDCNN Deep Convolutional Neural Networks with Wide First-layer Kernels

Chapter 1

Introduction

In the current industrial context, predictive maintenance is an essential practice to ensure production continuity. It comprises all strategies aimed at assessing machine conditions to schedule targeted interventions before critical failures occur.

Applying these strategies is particularly crucial for electric motors, which constitute the vast majority of actuators in industrial applications, with motor-driven systems accounting for 72% of global industrial electricity consumption [24]. Within this context, unexpected mechanical failures, particularly in REBs (some of which are exemplified in Figure 1.1), remain a leading cause of unscheduled machine downtime [55, 48]. In addition to their serious impact on plant safety, these faults entail a significant economic burden linked to production interruption, component replacement, and increased energy consumption caused by abnormal friction. These critical issues highlight the need for automated and efficient diagnostic tools capable of promptly detecting incipient faults.

Among the various monitoring techniques, vibration signature analysis has established itself as the standard diagnostic method. This approach enables the assessment of machinery integrity by recording and analysing the vibration signals generated by moving components. However, traditional fault diagnosis typically requires a significant amount of prior knowledge and appropriate techniques to achieve accurate results, remaining a complex challenge especially when the weak characteristics of incipient REB faults are masked by background noise [59, 65].

These limitations have driven a transition towards diagnostic approaches based on artificial intelligence. Although traditional machine learning methods are frequently employed for condition assessment, they heavily depend on the manual extraction of damage-sensitive features [29]. This requirement becomes a significant challenge as mechanical complexity grows and the number of simultaneous defects increases,

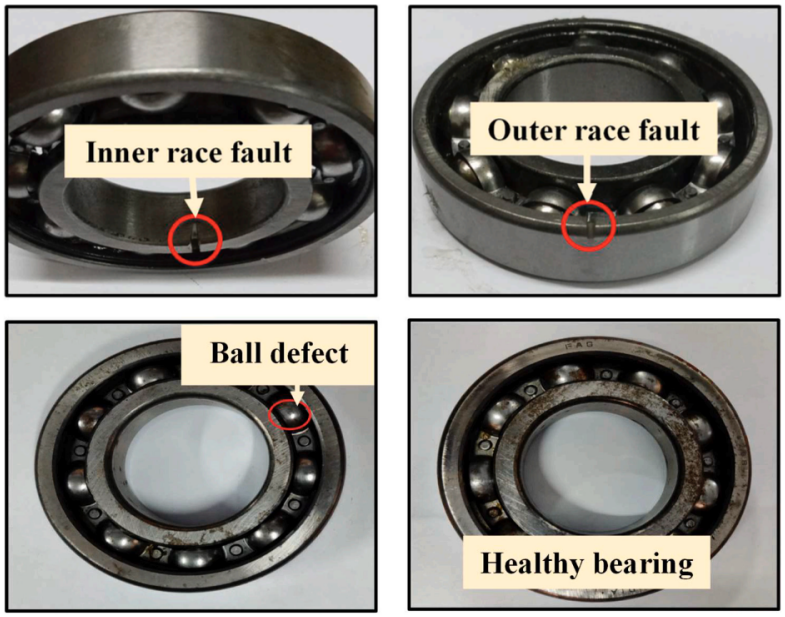


Figure 1.1: Bearing fault types: inner race, outer race, and ball defects [19]

often rendering hand-crafted features sub-optimal for accurately characterising the vibration signals [29]. To overcome these dependencies, 1D CNNs have emerged as a highly effective solution, primarily due to their unique ability to fuse feature extraction and classification into a single learning process [70, 69]. Furthermore, many studies have confirmed that 1D CNNs enable real-time condition monitoring and nearly instantaneous anomaly detection, consistently achieving state-of-the-art accuracy levels [25, 30, 13].

However, the effective implementation of these supervised models is inherently restricted by data availability. Specifically, the generalisation capability of deep learning architectures is severely challenged by the scarcity of labelled fault data, the significant class imbalance, and the high variability of vibration signatures caused by changing operating conditions. This data scarcity arises because, while healthy operational data are abundant, recording complete run-to-failure trajectories in the field represents a rare, costly, and often impractical operation due to industrial safety constraints [74].

To overcome this limitation, models are typically trained using datasets acquired from controlled laboratory test stands. Consequently, the transition of these deep learning architectures from such controlled environments to real-world industrial plants is severely constrained by the phenomenon of domain shift. In the field of deep learning, this problem occurs when a model optimised for a source domain, such as

an isolated test stand, fails when applied to a target domain due to a divergence in their data distributions [1]. In the context of mechanical diagnostics, this discrepancy stems from the highly dynamic nature of real-world operating environments. Here, sudden load variations, electromagnetic and structural noise induced by inverter Pulse Width Modulation (PWM) switching, and previously unobserved fault topologies profoundly alter the vibration signals, thus compromising the generalisation ability of classifiers.

While conventional generative architectures, such as GANs and VAEs, have been widely employed to augment empirical datasets by synthesising new signals, these solutions present structural limitations well documented in the literature [60, 53]. These primarily include mode collapse, severe instability in convergence dynamics, and the tendency to generate over-smoothed, low-fidelity samples. In response to these issues, DDPMs have emerged as the current state-of-the-art in deep generative modelling, effectively mitigating these drawbacks through their probabilistic framework [22, 18].

1.1 Objectives of the thesis

Motivated by these recent advancements, the primary objective of this thesis is to design and evaluate a novel conditional diffusion framework specifically tailored for the generation of mechanical vibration signals in the time domain. The proposed architecture aims to overcome the critical bottlenecks of data scarcity and class imbalance in mechanical diagnostics by acting as an advanced data augmentation solution. It is designed to synthesise high-fidelity vibration signals representing a wide range of REB fault types and severities, accurately capturing how fault signatures evolve with mechanical load and rotational speed, and faithfully replicating the harmonic distortions introduced by Variable Frequency Drives (VFDs).

To preserve the sequential and cyclostationary structure of vibration signals, the proposed model builds on a custom-designed 1D U-Net backbone, incorporating residual connections alongside a combination of local and global self-attention mechanisms. Training is guided by a tailored hybrid loss function jointly penalising the noise prediction error and the spectral discrepancy between the reconstructed and target signals. Precise control over the generative process is achieved through a conditioning mechanism in which the timestep embedding and the continuous physical parameters, encoded via Gaussian Fourier projections and dedicated Multi-Layer Perceptrons (MLPs), are injected into each residual block via AdaGN.

Furthermore, to demonstrate that the proposed diffusion framework goes beyond empirical memorisation to act as a physics-aware surrogate model, a core objective of this work is to evaluate its zero-shot generative capabilities: its ability to synthesise reliable vibration signatures for operating conditions absent from the training set. To this end, the architecture is trained with stochastic conditioning dropout, enabling CFG to increase adherence to the target conditioning during inference.

To demonstrate its effectiveness, the framework is validated across two experimental datasets. The CWRU dataset [9] is employed as a controlled baseline for localised bearing defects under varying fault severities, mechanical loads, and rotational speeds. In contrast, the EECSL dataset from Korea University [32] is utilised to assess the model's robustness against the harmonic distortions induced by VFDs under variable driving frequencies, while extending the generative domain to electromechanical faults, in particular dynamic rotor eccentricity.

Finally, the diagnostic utility and physical consistency of the generated signals are assessed through an extensive multi-domain evaluation framework. This involves extracting temporal and spectral diagnostic indicators, whose high-dimensional distributional similarity is subsequently visualised through t-Distributed Stochastic Neighbour Embedding (t-SNE). The practical applicability of the synthetic data is further demonstrated in a Train on Synthetic – Test on Real (TSTR) scenario, where a tailored CNN classifier trained exclusively on synthetic signals achieves diagnostic performance comparable to that of a classifier trained on real signals.

1.2 Organisation of the document

The thesis is organised as follows:

- Chapter 2 presents a critical review of the state of the art. Specifically, it demonstrates how the generalisation capabilities of deep learning diagnostic frameworks are inherently constrained by data scarcity and class imbalance. Consequently, it examines conventional generative architectures (GANs and VAEs) and analyses their structural vulnerabilities, thereby justifying the paradigm shift towards DDPMs.

- Chapter 3 details the hardware and software solutions adopted in this work, addressing the high computational demands of training diffusion models.

- Chapter 4 introduces the reference datasets and outlines the selection strategy employed to construct the comprehensive training dataset.

- Chapter 5 presents the algorithmic core of the thesis, illustrating the preprocessing pipeline for vibration signals, the design from scratch of the conditional diffusion framework, and the rigorous evaluation methodology adopted.
- Chapter 6 systematically presents the empirical results, quantifying the fidelity of the generated signals through temporal and spectral diagnostic features, evaluating their practical utility on diagnostic tasks, and demonstrating the model's zero-shot generation capabilities.
- Finally, chapter 7 summarises the scientific contributions of this thesis, critically discusses current architectural limitations, and outlines future research directions.

1.3 Partnership

This thesis was carried out in collaboration with the Institute for Industrial Information Technology (inIT) in Lemgo, Germany, during my six-month research stay.

Chapter 2

State of the art

"The wireless telegraph is not difficult to understand. The ordinary telegraph is like a very long cat. You pull the tail in New York, and it meows in Los Angeles. The wireless is the same, only without the cat."

Albert Einstein

2.1 The Evolution of Fault Diagnostics

In modern industrial environments, it is crucial to ensure the continuous and reliable operation of rotating machinery, particularly induction motors, which are the primary source of motive power in the manufacturing sector. Although these machines are inherently robust, their internal components are highly susceptible to mechanical degradation induced by constant operational stress. Statistical reliability assessments consistently identify REBs as the most vulnerable mechanical components. This issue is confirmed by a comprehensive survey of 2596 induction motors operating in harsh environments, demonstrating that bearing faults alone account for approximately 51% of all recorded failures [55].

The susceptibility of REBs to failure is significantly amplified by the widespread industrial adoption of VFDs. Unlike traditional Direct-on-Line (DOL) configurations, inverter-driven systems rely on PWM, which inherently generates a common-mode voltage that acts as the primary source of parasitic currents in the bearings [7]. Specifically, small Alternating Current (AC) motors (up to 20 kW) are highly vulnerable to these parasitic phenomena, which manifest as destructive Electric Discharge Machining (EDM) currents that break the lubricant's dielectric film, melting the metal surfaces and causing rapid structural degradation such as fluting or frosting on

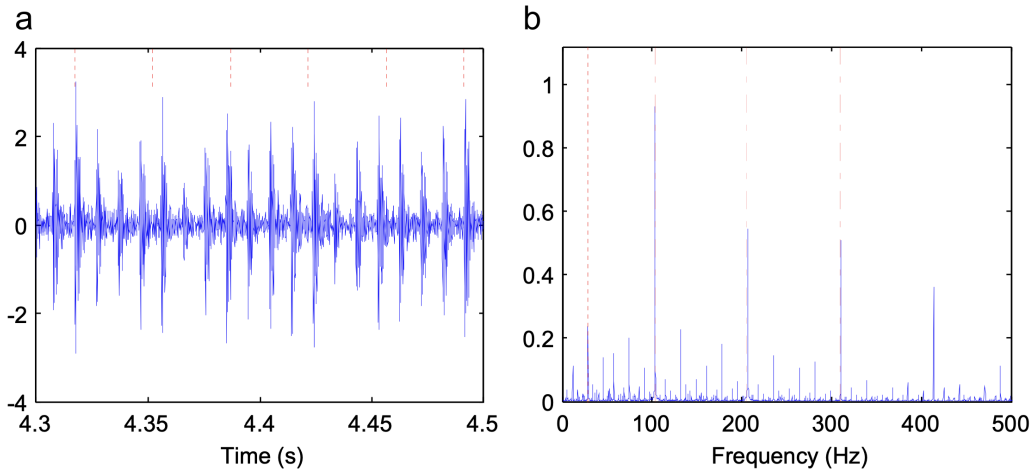


Figure 2.1: Comparison between the raw time-domain signal (a) and the corresponding envelope spectrum (b) for an outer race fault (CWRU dataset). Reprinted from [48].

the bearing raceways [7]. Consequently, VFDs pose a twofold challenge to modern machinery: not only do they accelerate mechanical wear, but they also significantly hinder the detection of incipient faults by introducing harmonic distortions and high-frequency switching noise into the acquired vibration signals.

2.1.1 Limitations of Traditional Signal Processing

Historically, identifying the localised REB defects, such as inner raceway, outer raceway, or rolling element faults, relied exclusively on traditional signal processing techniques. However, conventional spectral methods like the standard Fast Fourier Transform (FFT) have proven inadequate for incipient REB defects, as the generated high-frequency, low-amplitude transient impacts are easily masked by the complex noise environment characteristic of realistic industrial applications [49, 70].

To overcome this limitation, *Envelope Analysis* became the standard technique for demodulating the signal and extracting hidden periodic patterns [49]. Figure 2.1 illustrates how this demodulation process enables the clear isolation of the characteristic fault frequencies of an outer raceway defect, namely the Ball Pass Frequency Outer race (BPFO) and its harmonics, which would otherwise remain obscured in the raw vibration spectrum.

Nevertheless, its effectiveness depends entirely on the accurate identification of the optimal resonance frequency band, where the impulse energy of the fault predominates over the background noise [43, 3].

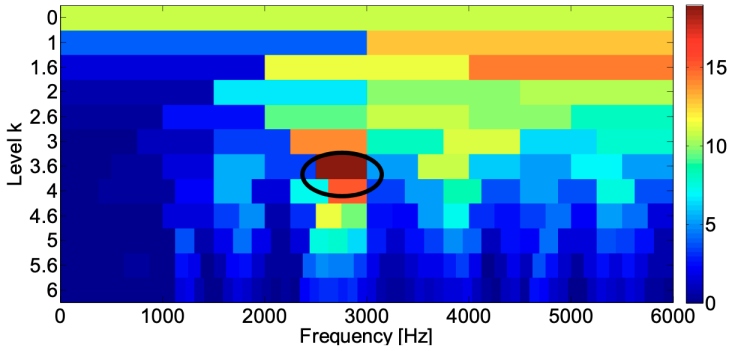


Figure 2.2: Fast Kurtogram of an inner race fault (defect diameter: 0.014 inches, CWRU dataset) [73]. The y-axis represents the step-discretised bandwidth resolution (Δf), while the x-axis represents the frequency (f).

To address this challenge, *spectral kurtosis* has been introduced as an ideal statistical metric for quantifying the impulsiveness of the vibration signal and detecting hidden transients generated by incipient bearing defects [43]. Based on this metric, the Kurtogram was developed as a two-dimensional mapping algorithm which, by evaluating spectral kurtosis across different centre frequencies and bandwidths, systematically identifies the exact spectral region of maximum impulsivity, thereby enabling the precise isolation of the defect signature [43].

Although highly effective, the original Kurtogram algorithm was computationally expensive for online industrial applications, as it required an exhaustive search across the entire frequency-resolution plane $(f, \Delta f)$ [3]. To solve this computational burden, the Fast Kurtogram was introduced as highly efficient alternative based on a pyramidal structure of multirate filter-banks. By leveraging quasi-analytic filters and a binary tree decomposition, this algorithm reduces the computational complexity to $O(N \log N)$, making it as fast as the traditional FFT [3].

Figure 2.2 shows the Fast Kurtogram applied to an inner race fault vibration signal, where the optimal demodulation band, identified by the black circle, is located at decomposition level 3.6 with a centre frequency of 2750 Hz [73]. The band exhibiting the highest kurtosis value, which concentrates more information about the fault signature, is subsequently isolated via a bandpass filter prior to envelope analysis. To identify the defect, traditional diagnostic pipelines require the analytical calculation of characteristic fault frequencies, such as BPFO and Ball Pass Frequency Inner race (BPFI), derived from the exact bearing geometry and shaft rotational speed [25, 13].

A slight discrepancy typically exists between the theoretically computed fault

frequencies and those effectively observed in the spectrum, due to speed fluctuations and manufacturing tolerances. A limitation of kurtosis-based band selection is its sensitivity to impulsive background noise: when random impacts dominate the signal, the Kurtogram may select the noise-driven frequency band rather than the one carrying the bearing fault signature [73].

Furthermore, the large-scale industrial deployment of these traditional signal processing techniques is constrained by several structural limitations. While the Kurtogram automates the selection of the optimal demodulation band, the subsequent identification of fault type still requires prior knowledge of the bearing geometry to compute the characteristic fault frequencies. Moreover, the filters derived from this process are inherently static and do not adapt dynamically to changes in load or speed.

These limitations have driven the research community towards diagnostic approaches based on artificial intelligence. Although traditional machine learning methods are frequently employed for condition monitoring, they heavily depend on the manual extraction of damage-sensitive features [29]. This requirement becomes a significant challenge as mechanical complexity grows and the number of simultaneous defects increases, often rendering hand-crafted features sub-optimal for accurately characterising the vibration signals [29]. To overcome these dependencies, 1D CNNs have emerged as a highly effective solution, primarily due to their unique ability to fuse feature extraction and classification into a single learning process, enabling fully autonomous operation even within highly noisy industrial environments [29, 25, 21].

By learning to extract discriminative representations directly from raw vibration signals, deep architectures free condition monitoring from its dependence on static filters and prior knowledge of theoretical kinematics. This end-to-end learning paradigm enables the model to automatically and adaptively identify fault-relevant features without any handcrafted pre-processing, while substantially improving generalisation across heterogeneous operating conditions [25, 70].

2.1.2 Compact CNNs for Real-Time Condition Monitoring

Among the deep learning architectures proposed for rolling bearing fault diagnosis, CNNs have emerged as the most widely adopted approach.

Although classical two-dimensional CNNs are standard models in computer vision, their application to time series analysis requires the transformation of one-dimensional signals into two-dimensional matrices, such as spectrograms. This reliance on complex matrix calculations imposes a heavy computational burden,

typically precluding real-time execution without the support of high-end Graphics Processing Units (GPUs) [29].

To overcome this limitation, compact one-dimensional CNNs have been specifically designed to operate directly on raw vibration signals, effectively bypassing the need for computationally expensive 2D transformations. By processing the input through linear convolutions along the time axis, these networks autonomously learn to extract short-duration transients caused by kinematic defects, thereby completely eliminating the need for manual feature extraction [25, 29].

The computational efficiency of these models stems from the intrinsic mathematical simplicity of 1D convolutions, which are executed as elementary scalar additions and multiplications [29]. By employing shallow configurations, often comprising one to three hidden layers and fewer than 10,000 trainable parameters, these lightweight architectures drastically minimise the computational load, enabling straightforward deployment on standard Central Processing Units (CPUs) or on resource-constrained embedded systems, such as Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) [13, 25, 29].

The feasibility of real-time bearing condition monitoring was initially demonstrated by applying a one-dimensional CNN to motor current signatures, achieving a diagnostic accuracy (healthy versus faulty condition) of over 97% [25]. Subsequently, the same topology was successfully applied to the analysis of vibration signals, demonstrating that high diagnostic performance does not necessarily require deep architectures. In particular, an ultra-lightweight model, consisting of just three convolutional layers and a single hidden layer of 20 neurons, achieved over 93.2% accuracy on the benchmark CWRU dataset [13].

The most significant operational advantage demonstrated by these practical implementations is their inference speed. Specifically, the time required to classify a new signal segment and identify a potential fault is less than 1 millisecond on a standard CPU [13, 25]. Given this negligible computational overhead, embedded fault diagnosis emerges as a practical and highly scalable solution for modern industrial condition monitoring [13, 25].

2.1.3 Industrial Challenges: Domain Shift and Signal Noise

Despite the remarkable diagnostic efficiency demonstrated in controlled laboratory environments, standard one-dimensional CNNs often suffer a significant decline in accuracy when used in real-world production plants. This issue is the result of two structural challenges.

First, conventional one-dimensional CNNs typically employ small receptive fields in the initial layers, making them highly susceptible to high-frequency interference [70], such as the switching noise generated by VFDs, which disturbs the useful signal and compromises the feature extraction process. Second, operational fluctuations, including variations in workload or rotational speed, continuously alter the underlying signal distribution, triggering a phenomenon known as domain shift. Consequently, a model trained exclusively on a specific load regime typically fails to generalise when exposed to new, previously unobserved operating conditions [70].

To overcome these practical constraints, the Deep Convolutional Neural Networks with Wide First-layer Kernels (WDCNN) [70] was introduced, leveraging the Adaptive Batch Normalisation (AdaBN) algorithm [33]. This architecture replaces the conventional small filters of the first convolutional layer with wide kernels, which effectively act as adaptive low-pass filters. This structural modification enables the network to capture essential macroscopic kinematic features, while simultaneously suppressing high-frequency stochastic noise.

Furthermore, to counteract the impact of load variations, the AdaBN algorithm is employed to dynamically align the feature distributions across different operational domains. Instead of relying on fixed statistical parameters computed during the training phase, the algorithm recalculates the mean (μ_t) and variance (σ_t^2) directly from samples of the new target domain [33]. Only the scale (γ) and shift (β) parameters, previously learned from the source domain, remain unchanged. This mathematical strategy provides the model with robust domain adaptation capabilities without requiring complete retraining [70].

Remarkably, the combination of the wide kernels for noise suppression and the AdaBN for distributional alignment allows the framework to maintain a classification accuracy of over 90% in a 10-class diagnostic task on the CWRU dataset, even when the Signal-to-Noise Ratio (SNR) is -4 dB [70].

2.1.4 Data Scarcity and Generative Solutions

Although the WDCNN architecture, together with AdaBN, theoretically allows the network to adapt to new operating conditions, it presents a significant limitation for practical implementation. For statistical adaptation to function optimally, the system requires a robust estimate of the new target domain. Consequently, when faced with a sudden change in operating regime, the network is unable to reliably classify the very first acquired signals, since it must first accumulate a sufficient batch of new vibration samples to accurately recalculate the mean (μ_t) and variance

(σ_t^2). This adaptation latency, combined with the constant need for new training samples, highlights the fundamental limitation of modern data-driven diagnostics: a critical dependence on substantial volumes of labelled training data [68].

This dependency becomes particularly problematic if we consider the intrinsic class imbalance that characterises industrial settings. In real-world applications, data representing machinery in optimal working condition is extremely abundant. Conversely, the acquisition of complete run-to-failure trajectories is rare, costly, and often not feasible in most real-world applications due to strict plant safety constraints [45, 68].

This intrinsic data scarcity is further compounded by the difficulty in isolating clear fault signatures in the field. During the early stages of degradation, particularly at low operating speeds, the mechanical impacts generated by the defect are extremely weak and easily masked by background noise [45].

To mitigate the lack of fault data needed to train robust classifiers, the scientific community initially focused on traditional data augmentation techniques, such as Gaussian noise injection or time stretching. However, the application of these superficial transformations to vibration signals introduces physical limitations. Specifically, arbitrarily manipulating the time scale of a vibration signal inevitably leads to an irreversible alteration of the relationship between the characteristic fault frequencies (e.g., BPFI, BPFO) and the structural resonance frequencies, thus generating augmented samples that do not reflect the actual physics of the fault.

In light of this dual challenge, namely the scarcity of real fault data and the inadequacy of classical data augmentation techniques, recent literature has identified deep generative models as a definitive methodological solution. The deployment of adversarial learning frameworks, particularly GANs, initially emerged as the most promising strategy for synthesising high-fidelity raw mechanical sensor signals [46].

2.2 Deep Generative Networks: GANs and VAEs

The transition from discriminative to generative modelling represents the modern response to the challenge of *data scarcity*. While the architectures discussed so far are designed to map an input signal to a specific fault class, an approach known as discriminative learning, the focus is now shifting towards understanding and replicating the probability distribution underlying the data itself.

2.2.1 The Adversarial Framework: Principles and Advantages

A significant advancement in deep generative modelling was achieved through the introduction of the adversarial learning framework [16]. Specifically, GANs are formalised as a system that estimates probability distributions through an adversarial process, which involves the simultaneous and competitive training of two distinct neural networks [16]:

- **The Generative Model (G)** is designed to capture the statistical distribution of the real data. Starting from a latent space of random noise (z), G generates synthetic samples with the aim of making them statistically indistinguishable from the real ones.
- **The Discriminative Model (D)** acts as a traditional classifier, estimating the probability that an input sample originates from the real training dataset rather than from the synthetic data produced by the generator G.

From a mathematical and algorithmic perspective, this architecture takes the form of a two-player minimax game [16]. The generative model is trained to minimise the probability that the discriminator correctly identifies synthetic samples, while the discriminator is concurrently optimised to maximise its accuracy in distinguishing real data from forgeries. This zero-sum competition drives both models to continuously refine their weights until a theoretical Nash equilibrium is reached.

As formally demonstrated, in the space of arbitrary functions, this state coincides with a unique mathematical saddle point: a minimum for one player's strategy and a maximum for the opponent's [16]. In this global optimum, the distribution of the generated data (p_g) perfectly replicates the underlying probability distribution of the real data (p_{data}). Consequently, the forgeries become indistinguishable from the original data, leading the discriminator's classification probability to converge to 50% [16].

The adoption of this adversarial paradigm offers significant advantages for REB diagnostics, effectively overcoming the generalisation limits and severe class imbalance commonly present in empirical datasets [46, 74]. Specifically, by excelling at modelling sharp distributions [16], this framework proves highly effective in accurately replicating the high-frequency impulsive transients caused by kinematic defects.

Furthermore, the generation of these synthetic signals must rigorously respect the physics of the fault without falling into mere data replication. A crucial architectural

advantage is that the generator never directly observes the real sensor data; instead, it updates its parameters exclusively through the error gradients backpropagated by the discriminator [16].

2.2.2 Evolution of GANs in Industrial Diagnostics

Once the theoretical framework for adversarial networks had been established, the next challenge involved adapting this methodology, originally designed for two-dimensional image data, to the domain of raw time series. A fundamental contribution in this direction was provided by the adaptation of the Auxiliary Classifier Generative Adversarial Network (ACGAN) framework, originally introduced for image generation [38], to raw mechanical signals [46]. This approach was specifically conceived to synthesise targeted artificial samples capable of effectively rebalancing training sets.

The innovation of the adapted ACGAN architecture is twofold. First, in both the generator and the discriminator, traditional fully-connected layers are replaced by 1D convolutions, which allow the networks to effectively learn the local features and hierarchical representations directly from vibration data [46]. Second, the discriminator is integrated with an auxiliary classifier, enabling it to yield two distinct outputs for each input sample. Specifically, the discriminator not only predicts whether the sample is real or generated but also identifies the specific fault category [46].

Within this framework, the category label serves a dual purpose: while the generator uses the label as a conditioning input to synthesise a targeted fault signature, the discriminator uses the real labels as a supervisory input to accurately classify actual fault conditions, thereby forcing the generator to produce highly realistic samples.

Although the first GAN architectures demonstrated their effectiveness in synthesising realistic raw vibration signals, their application to industrial machinery soon exposed the theoretical limitations of the original framework. A fundamental structural issue of these models is their susceptibility to *mode collapse*, a pathological training condition in which the generator maps the latent space to a limited set of modes, thereby producing highly repetitive samples that fail to capture the intrinsic diversity of the empirical data [74]. This phenomenon becomes a severe bottleneck when tackling complex multimodal classification tasks, such as distinguishing between defects located across different bearing components or varying severity levels [74].

To directly address this critical issue, the Dual Discriminator Conditional Gen-

erative Adversarial Network (D2CGAN) architecture was specifically developed to generate multimodal fault signals [74]. The structural peculiarity of this framework lies in the employment of two parallel discriminators: the first is designed to reward samples drawn from the real data distribution, whereas the second assigns high scores exclusively to the synthetic samples produced by the generator. From a mathematical perspective, this three-player game explicitly combines two opposing objective metrics: the forward Kullback-Leibler (KL) divergence and the reverse KL divergence. The generator leverages the complementary properties of both metrics by simultaneously minimising them within a unified objective function. This mathematical constraint forces the generator to explore and cover all the failure modes, effectively preventing the collapse towards a single, repetitive type of anomaly and ensuring the generation of highly diverse multimodal fault samples [74].

Another typical problem concerning the training stability of GAN architectures is the *vanishing gradient* phenomenon. The standard Conditional GAN framework relies on the Jensen-Shannon (JS) divergence metric. As formally demonstrated, traditional divergence metrics require a degree of initial overlap between the real and generated data distributions to provide meaningful feedback [41]. When this overlap is lacking, the delicate adversarial equilibrium required for successful training rapidly breaks down: the discriminator becomes excessively proficient at distinguishing real from synthetic samples, which in turn causes the loss function to saturate and makes the backpropagated error gradients infinitesimal [41, 26]. Consequently, the earliest layers of the generative architecture receive weight updates too negligible to drive optimisation, effectively stalling the entire learning process [4].

To overcome the vanishing gradient and the subsequent mode collapse, thus ensuring stable model convergence, research shifted towards advanced architectures based on the Wasserstein distance, such as the Wasserstein Conditional GAN with Hierarchical Feature Matching (WCGAN-HFM) [41]. This advanced model attempts to mitigate the structural defects of the adversarial paradigm by intervening on two distinct methodological fronts [41]. First, the problematic JS divergence is replaced by the Wasserstein distance, which mathematically ensures a smooth and continuous gradient flow even when the real and generated distributions are entirely disjoint. Second, to counteract the network's tendency to align distributions exclusively at the global level, a dynamic that would favour the generation of signals corresponding to healthy machinery (statistically dominant), the architecture incorporates an additional penalty function known as Hierarchical Feature Matching (HFM) [41]. This constraint forces the features extracted in the deeper layers of the network to

approximate the microscopic details of the real signal for each specific fault subclass, ensuring superior fidelity in the reproduction of mechanical transients [41].

2.2.3 Fundamental Issues of the Adversarial Paradigm

Although the integration of sophisticated metrics and complex architectures has attempted to overcome the limitations of GANs, such solutions have ultimately highlighted the inherent rigidity and fragility of the adversarial framework. A fundamental cause of these persistent convergence failures, which is intrinsically linked to mode collapse, is the phenomenon of *catastrophic forgetting* [53].

From a mathematical perspective, training a GAN inherently constitutes a continual learning problem [53]. During the adversarial process, the generator continuously alters the probability distribution of the synthetic data it produces. Consequently, the discriminator is forced to solve a continuous sequence of novel and distinct classification tasks. As the network parameters are stochastically updated to adapt to the features of the current task, this perpetual alteration irreparably overwrites and erases the representations learned in previous iterations [53]. This pathological dynamic, where the knowledge of past distributions is abruptly destroyed, prevents the system from reaching a stable Nash equilibrium, rendering the adversarial framework structurally inadequate for handling complex and multimodal industrial datasets.

Attempts to stabilise this paradigm are hindered by severe design rigidity. Unlike traditional neural architectures, in which the cost function follows a monotonic and predictable trend, the minimax optimisation of GANs is inherently unstable. Even a minimal imbalance in critical hyperparameters, such as the learning rate, can cause one network to overpower the other. Moreover, because this fragile adversarial equilibrium is deeply influenced by the specific nature of the underlying data distribution, the optimal hyperparameter configuration cannot be generalised, thereby necessitating a highly specific and time-consuming recalibration for each individual dataset [4]. Furthermore, since the loss value in a GAN does not directly reflect the physical fidelity of the generated signal, the search for the optimal configuration is reduced to a costly, heuristic trial-and-error procedure. In an industrial context, this operational overhead represents a critical barrier: as operating conditions change, the entire tuning process must be repeated almost from scratch, drastically limiting the practical scalability of such models.

2.2.4 Multivariate Conditioning and OOD Issues

Although the intrinsic instability of minimax optimisation already makes training a challenging process in its own right, the most significant limitation of the adversarial framework used in industrial settings lies in multivariate conditioning. For the creation of dynamic simulators and true industrial *Digital Twins* [62], it is not sufficient to generate a generic anomaly; rather, the network must be capable of synthesising the defect at an exact operating regime.

However, forcing a GAN to map continuous, complex, and overlapping distributions, such as the precise vibrational signatures of a bearing under varying rotational speeds and mechanical loads, exacerbates the already precarious dynamics of adversarial training [4]. The introduction of continuous conditional vectors alongside discrete fault labels excessively complicates the minimax game, almost inevitably triggering mode collapse and catastrophic forgetting [34, 53].

It is precisely to circumvent this issue that advanced diagnostic frameworks based on GANs (e.g., ACGAN, D2CGAN) are forced into a methodological compromise: deliberately ignoring the machine's kinematics. Consequently, the established practice consists of merging samples recorded across different rotational speeds and loads into a single macro-set for each discrete fault class [46, 74]. While this strategy succeeds in training a domain-invariant classifier, from a pure simulation perspective it represents an insurmountable limitation, as the physical fidelity of specific operational regimes is entirely sacrificed to preserve the stability of the optimisation process.

Beyond this methodological compromise, adversarial architectures suffer from a definitive structural issue regarding Out-of-Distribution (OOD) scenarios, as they are strictly restricted to reproducing data from previously observed conditions rather than generating unseen operational states [18]. This limitation is problematic in industrial settings, where acquiring empirical samples for every possible combination of rotational speed, load, and defect severity is both practically unfeasible and cost-prohibitive. Since the adversarial framework learns solely to replicate the input data distributions without internalising the underlying physical dynamics, it remains incapable of modelling physical effects outside its training distribution [34, 18].

It is precisely to overcome these structural limitations that a branch of the diagnostic literature has shifted towards VAEs.

2.2.5 Mathematical Foundations of VAEs

The foundational core of the VAE framework lies in the rigorous optimisation of a probabilistic lower bound on the marginal likelihood [27]. This variational formulation replaces the unstable dynamics of adversarial optimisation with a well-defined objective function, ensuring mathematically stable training and mitigating the risk of mode collapse.

Structurally, this framework adopts a probabilistic *encoder-decoder* topology. The probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$ approximates the intractable true posterior by parameterising the conditional distribution of the latent variables given the input signal $\mathbf{x}$ [27]. Instead of yielding a single deterministic vector, it maps the data into a continuous latent space by estimating the mean μ and covariance matrix $\mathbf{\Sigma}$ of a multivariate Gaussian distribution. Subsequently, the generative counterpart, the probabilistic decoder $p_\theta(\mathbf{x}|\mathbf{z})$, models the conditional distribution of the data, reconstructing the original signal from a latent vector $\mathbf{z}$ sampled stochastically from the approximate posterior [27].

The main analytical obstacle in this architecture lies in the sampling operation $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$. As a stochastic process, it introduces a non-deterministic node into the network's computational graph, thereby blocking the calculation of derivatives and preventing the update of the encoder's weights via classical backpropagation. To overcome this constraint and make the sampling fully differentiable, the architecture employs a mathematical formulation known as the *reparameterisation trick* [27], through which the latent variable $\mathbf{z}$ is rewritten as a continuous and deterministic transformation:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2.1}$$

By decoupling the stochasticity and confining it to an auxiliary, independent noise variable $\boldsymbol{\epsilon}$, the model creates a deterministic path for the gradient flow, which can now propagate freely through the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Consequently, the network is optimised by maximising the variational lower bound, referred as Evidence Lower Bound (ELBO) [27], an objective function that balances two opposing analytical forces:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \tag{2.2}$$

The first term, the Kullback–Leibler divergence (D_{KL}), acts as a strict mathematical regulariser, forcing the estimated posterior distribution to conform to a

prior distribution $p(\mathbf{z})$, typically assumed to be a standard multivariate Gaussian. The second term represents the expected log-likelihood, acting as a reconstruction term that guides the probabilistic decoder to faithfully reproduce the original input from the latent vector $\mathbf{z}$ sampled from $q_\phi(\mathbf{z}|\mathbf{x})$ [27].

It is this delicate mathematical balance that ensures the latent space remains continuous and structured, while simultaneously guaranteeing the high fidelity of the reconstructed data. This property proves crucial during the subsequent inference phase, where the encoder is entirely discarded, and the generation of new physically consistent samples occurs by feeding standard Gaussian noise directly into the decoder.

2.2.6 Hybrid Generative Frameworks in Industrial Diagnostics

In the field of industrial diagnostics, the tendency of VAEs to generate blurry samples and suffer from oversmoothing has driven research towards the development of hybrid architectures, capable of combining the stability typical of VAEs with the high-fidelity generative performance of GANs.

Architectures based on this hybrid paradigm are specifically designed to address data scarcity in complex mechanical systems. For instance, VAE-enhanced adversarial networks have been successfully validated for the automated fault diagnosis of industrial chillers [64]. By projecting rare failure samples into a continuous and structured latent space, the VAE guides and stabilises the adversarial generation process. This allows the model to synthesise realistic surrogate data, providing a balanced training set for downstream classifiers [64].

However, while these models perform well under controlled conditions, extracting diagnostic features from vibration signals contaminated by noise interference remains a significant challenge. When noise is combined with highly imbalanced datasets, Variational Autoencoding Generative Adversarial Network (VAEGAN) often suffer from training instability and gradient collapse, failing to produce reliable surrogate data. To overcome this limitation, advanced frameworks such as the Adaptive Variational Autoencoding Generative Adversarial Network (AVAEGAN) have been introduced [59]. This framework addresses the dual challenge posed by noise and data scarcity through three targeted innovations: an adaptive network equipped with self-attention for robust feature extraction in noisy environments, an adaptive loss calculation method that balances the model loss and function gradients to stabilise training, and an adaptive optimal data seeker that evaluates and selects only the synthetic samples with the highest diagnostic fidelity [59].

Beyond noise and data scarcity, industrial diagnostics must also address the problem of *domain shift* across varying operational states. In the context of transfer learning, the attempt to overcome this phenomenon through a direct alignment of divergent feature distributions proves ineffective when the discrepancy between the source and target domains is substantial [71]. To overcome this limitation, recent advancements propose an indirect latent alignment process: rather than enforcing a direct match, the source and target domains are projected into a shared latent space and aligned by constraining them to a common Gaussian prior distribution [71]. To achieve this indirect alignment effectively, a novel conditional weighting transfer Wasserstein auto-encoder (CWTWAE) is proposed for rolling bearing fault diagnosis with multi-source domains [71].

This model not only maps features extracted from multiple source domains to a single prior Gaussian distribution, but also integrates a conditional weighting strategy to dynamically quantify the similarity of different source domains to the target domain [71]. By assigning greater weights to the source domains that exhibit higher similarity with the target domain, the framework minimises the discrepancy in conditional distribution, achieving accurate cross-domain diagnoses [71].

Although combining VAEs and GANs helps balance out their individual weaknesses, these hybrid models still suffer from their core structural issues. Because autoencoders rely on data compression within a low-dimensional latent space and adversarial networks remain unstable during training, there is a strict limit to the quality and reliability they can achieve in complex industrial applications. To overcome these limitations, we need a completely different approach that moves away from both compression-based and adversarial methods.

By replacing the continuous latent bottleneck with an iterative denoising mechanism, the DDPMs introduce a novel generative paradigm capable of preserving the finest high-frequency details without sacrificing training stability.

2.3 Theoretical Foundations of Denoising Diffusion Probabilistic Models

To overcome the structural constraints of prior architectures, research has shifted toward a novel class of deep networks originally inspired by non-equilibrium thermodynamics: Denoising Diffusion Probabilistic Models (DDPMs) [50, 22]. These models rely on an iterative, dimension-preserving Markovian process, which perfectly fulfils the analytical requirements for high-fidelity signal generation. By circumventing

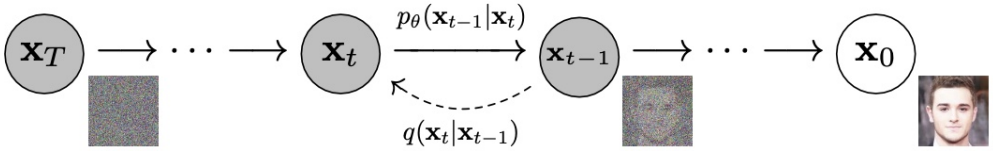


Figure 2.3: Directed graphical model of the Markov chains within the DDPM framework. Reproduced from Ho et al. [22].

both the mathematical instability of GANs and the oversmoothing induced by VAEs, diffusion architectures have elevated data generation to unprecedented levels of structural quality and diversity [11].

The success of this paradigm is justified by a radical topological and methodological divergence from its predecessors. Unlike VAEs, which compress information by forcing data through a low-dimensional bottleneck, thereby losing high-frequency details, diffusion models maintain the original dimensionality of the input data intact. Instead of compressing features, they systematically corrupt the data structure through a controlled noise injection mechanism, which is mathematically stable [22].

Furthermore, rather than relying on the competitive minimax training of two adversarial networks, DDPMs are optimised using a likelihood-based objective [22]. This formulation facilitates stable optimisation dynamics, effectively eliminating two critical issues typical of adversarial settings: the risk of vanishing gradients, which stalls network learning [4], and the phenomenon of catastrophic forgetting, which triggers mode collapse and can make the training of GANs non-convergent [53].

2.3.1 Forward and Reverse Processes

From a mathematical perspective, the progressive corruption of the data structure is formalised through a sequence of latent states governed by a fixed Markov chain. Within this framework, these representations correspond precisely to the intermediate noisy variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ generated along the chain from the original data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ [22].

To formalise this generative framework, the architecture is defined as a class of latent variable models driven by two Markovian chains: a *forward process* (or diffusion process), and a *reverse process* (or generative process) (see Figure 2.3) [22].

The Forward Process

The *forward process*, mathematically denoted as $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$, is a fixed Markov chain which gradually adds Gaussian noise to the data according to a variance schedule $\beta_1, \ldots, \beta_T$. The transition for a single timestep is modelled analytically as follows [22]:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2.3}$$

To obtain the state $\mathbf{x}_t$, it is not necessary to calculate iteratively all the intermediate states from $\mathbf{x}_1$ to $\mathbf{x}_{t-1}$. Given the additive properties of the Gaussian distribution, the forward process allows for sampling the noisy state $\mathbf{x}_t$ at an arbitrary timestep t in closed form [22]. Defining $\alpha_t := 1 - \beta_t$ and the corresponding cumulative product $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2.4}$$

For a sufficiently large number of diffusion steps T, the cumulative product $\bar{\alpha}_T$ asymptotically approaches zero, causing the final state $\mathbf{x}_T$ to lose correlation with the initial state and converge to standard isotropic Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

The Reverse Process

The generative phase, known as the *reverse process*, is defined by a Markov chain with learned Gaussian transitions, mathematically formalised by the joint distribution $p_\theta(\mathbf{x}_{0:T})$ starting at $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ [22].

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{2.5}$$

The neural network is trained to progressively reverse the diffusion trajectory. Because the variance schedule restricts the noise injection to sufficiently small amounts at each individual timestep (via small β_t values), the true reverse transitions $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ can be approximated by conditional Gaussian distributions. Accordingly, each reverse transition, estimated by the neural model parameterised by θ, is defined as follows:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{2.6}$$

The discretisation into many steps is mathematically necessary. The Gaussian approximation of $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ relies on a small-noise assumption: $\beta_t \ll 1$. Attempting single-step generation would violate this assumption: with a single large transition, the true posterior would deviate substantially from a Gaussian form, potentially exhibiting complex multimodal structure.

2.3.2 Objective Function Formulation

The ultimate goal of training a generative model is to maximise the probability that the network generates the true data, mathematically expressed as the data log-likelihood. Since directly computing this exact probability is analytically intractable in DDPMs, diffusion models are trained by minimising a variational upper bound (L) on the negative log-likelihood of the data [22].

This formulation uses the KL divergence to compare the reverse transitions estimated by the neural network $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ against the forward process posteriors $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, which act as the tractable analytical target when conditioned on the uncorrupted data $\mathbf{x}_0$ [22].

Since both $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ and $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ are parameterised as Gaussian distributions with identical, untrained diagonal covariances ($\sigma_t^2\mathbf{I}$), this KL divergence simplifies to the weighted MSE between their respective mean vectors:

$$L_{t-1} - C = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2 \right] \tag{2.7}$$

where C is a constant independent of θ [22].

Although the network could theoretically be trained to predict the clean signal $\mathbf{x}_0$ directly, doing so leads to worse sample quality [22]. Instead, the DDPM framework parameterises the reverse process mean $\boldsymbol{\mu}_\theta$ to predict the residual noise vector $\boldsymbol{\epsilon}$ injected at timestep t:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) \tag{2.8}$$

where $\boldsymbol{\epsilon}_\theta$ is a neural function approximator trained to estimate $\boldsymbol{\epsilon}$ from the noisy input $\mathbf{x}_t$ [22]. This parameterisation simplifies the variational bound into a weighted objective resembling denoising score matching:

$$L_{t-1} - C = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\|^2 \right] \tag{2.9}$$

Although optimising this exact variational bound is mathematically rigorous, Ho et al. demonstrated that a simplified, unweighted variant of this objective function, denoted as L_{simple}, is easier to implement and yields significantly superior sample quality in practice [22]:

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2 \right] \tag{2.10}$$

where the timestep t is sampled uniformly between 1 and T [22]. By discarding the original weighting coefficients, L_{simple} effectively down-weights the loss at small values

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Figure 2.4: Training (Algorithm 1) and sampling (Algorithm 2) procedures for probabilistic diffusion models. Reproduced from Ho et al. [22].

of t. This forces the network to focus less on easily denoising slight perturbations, and more on solving the difficult denoising tasks at larger timesteps, which empirically improves the final generative quality [22].

2.3.3 Training and Inference Dynamics

To fully comprehend how the architecture operates and the exact nature of the predictive target assigned to the neural network, it is essential to draw a clear algorithmic distinction between the training phase and the inference (sampling) phase, whose formal differences are summarised in Figure 2.4.

While the inference phase is intrinsically bound to the sequential and recursive nature of the Markov chain, the training is designed as a stochastic and highly parallelisable process [22]. As formalised in *Algorithm 1* (Figure 2.4), the training algorithm operates by evaluating arbitrary timesteps independently, without requiring the sequential calculation of intermediate states. During this phase, the system samples an original data point $\mathbf{x}_0$ and selects a random timestep t from a discrete uniform distribution, ensuring that each timestep has an equal probability of being considered. Simultaneously, a noise tensor $\boldsymbol{\epsilon}$ is sampled from a standard isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Finally, by exploiting the analytical properties of the latter, the original data is corrupted by directly calculating the state $\mathbf{x}_t$ in closed form.

At this point, the network is not trained to approximate the microscopic fraction of noise added during the single local transition between $t-1$ and t; instead, the model is assigned the task of estimating the entire cumulative noise tensor $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ introduced starting from the original data, relying exclusively on the observation of the corrupted state $\mathbf{x}_t$ and the degradation level encoded by the timestep t [22].

Moving to the inference phase, which is detailed in *Algorithm 2* in Figure 2.4, the generation process implements an iterative reverse sampling mechanism. Within

this procedure, although the network estimates the global noise $\boldsymbol{\epsilon}_\theta$, the update law utilises this tensor exclusively to define a local directional gradient to execute a single reverse step. Analytically, the total noise estimate $\boldsymbol{\epsilon}_\theta$ is attenuated by multiplying it by the following scaling factor:

$$\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \tag{2.11}$$

Because this coefficient strictly depends on $\beta_t = 1 - \alpha_t$ (small value by design), the system effectively subtracts only a small fraction of the predicted noise, extracting a quantity that is necessary and sufficient to step back solely to the adjacent timestep $\mathbf{x}_{t-1}$.

2.3.4 The U-Net Backbone

The implementation of a diffusion model requires designing a neural network capable of estimating the residual noise $\boldsymbol{\epsilon_\theta(\mathbf{x}_t, t)}$ [22]. This architecture must ensure dimensional invariance, as the latent states $\mathbf{x}_1, \ldots, \mathbf{x}_T$ share the same dimensionality of the original input $\mathbf{x}_0$ [22].

In the literature, this prediction task is conventionally delegated to architectures based on the U-Net topology [22, 12]. Originally conceived for biomedical semantic segmentation [44], whose foundational 2D topology is illustrated in Figure 2.5, the U-Net has rapidly emerged as the standard backbone for diffusion models [22].

The crucial feature of this network resides in its encoder-decoder structure. The encoding path progressively reduces the resolution of the input signal while simultaneously expanding the channel depth [22, 12], thereby constraining the model to learn a highly dense latent representation. This stage of maximum dimensional abstraction is essential for extracting the global semantics and long-range data dependencies.

However, the spatial contraction inherent to the encoder causes the loss of high-frequency details [72]. To neutralise this degradation, the architecture integrates skip connections, which establish a direct and symmetrical routing between the encoding layers and their corresponding decoding stages [18, 12]. This mechanism allows high-resolution tensors to bypass the latent bottleneck entirely and be directly concatenated to the activation maps during the subsequent upsampling operations.

To process the data across the entire discrete time horizon $t \in [1, T]$, the network backbone employs a single set of shared parameters θ [22]. To condition the network on the temporal evolution of the process, the diffusion step t is specified to the architecture by adding the Transformer sinusoidal position embedding into each convolutional residual block [22].

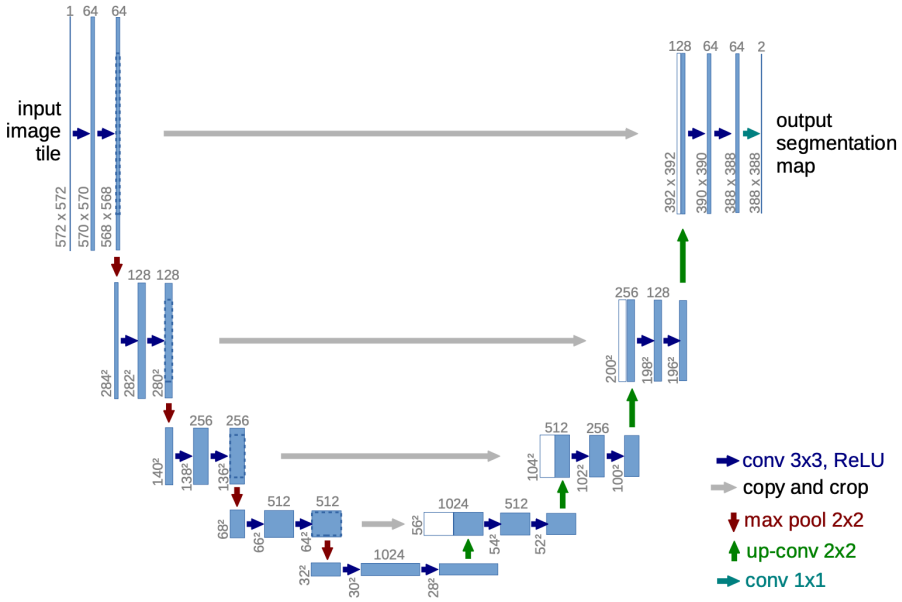


Figure 2.5: The original 2D U-Net architecture proposed for biomedical image segmentation. Reproduced from Ronneberger et al. [44].

2.3.5 Conditional Generation and CFG

While standard DDPMs demonstrate exceptional capabilities in modelling complex data distributions, their unconditioned nature poses a significant limitation for practical industrial applications. In the context of fault diagnostics, it is strictly necessary to control the generative process to synthesise specific mechanical signatures.

Initially, this requirement was addressed through *Classifier Guidance*, a technique that exploits a pre-trained external classifier $p_\phi(\mathbf{y}|\mathbf{x}_t, t)$ to guide the unconditional diffusion sampling process towards an arbitrary class label $\mathbf{y}$. As formally demonstrated [12], the sampling transitions can be explicitly conditioned using the gradients of the classifier $\nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t, t)$. However, this approach presents critical drawbacks: it requires training an additional classifier on heavily noised data, making the overall training process highly complicated [23]. Furthermore, updating the intermediate states via external gradients during inference can induce adversarial attack effects.

To overcome these issues, the Classifier-Free Guidance (CFG) paradigm was introduced [23]. The CFG effectively eliminates the need for an external classifier by jointly training the neural network to model both the conditional diffusion process $p_\theta(\mathbf{x}_t|\mathbf{c})$ and the unconditional diffusion process $p_\theta(\mathbf{x}_t)$, where $\mathbf{c}$ represents the target conditioning vector [23]. During the training phase, the conditioning

vector $\mathbf{c}$ is randomly replaced by a null token ($\emptyset$) with a certain dropout probability. Consequently, this stochastic masking forces the network to simultaneously learn both the baseline data distribution and the specific topological features dictated by the condition [23].

During the inference phase, the CFG mechanism simultaneously computes the conditional and unconditional noise estimates, respectively $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$ and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, extrapolating the final noise prediction via a linear combination weighted by the guidance scale parameter w:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c}) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \emptyset) + w \cdot [\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \emptyset)] \tag{2.12}$$

By decoupling the conditioning intensity from the network weights, this mathematical formulation provides the generative system with high operational flexibility. Specifically, the guidance scale w can be dynamically adjusted during the sampling phase, completely bypassing the need for model retraining and allowing an explicit trade-off between sample diversity and fidelity [12, 23].

2.4 Applications of Diffusion Models in Fault Diagnostics

Having established the theoretical foundations of DDPMs, this section presents how these generative architectures are practically deployed as advanced data augmentation frameworks to address the main challenges of condition monitoring. In particular, the discussion outlines the shift from fundamental approaches based on two-dimensional time-frequency representations toward advanced one-dimensional architectures designed to synthesise raw vibration signals and mitigate the phenomenon of domain shift. Finally, the discussion examines the capabilities of diffusion models in zero-shot generation for OOD scenarios.

2.4.1 Data Augmentation and 2D Domain Transition

The fundamental application of DDPMs in mechanical diagnostics, like other generative models, addresses the pervasive challenge of data scarcity and class imbalance [72]. Since diffusion models were initially developed and consolidated within the field of computer vision, a widely adopted methodology in the literature when working with time series involves translating the problem from the time domain to the time-frequency domain using the Continuous Wavelet Transform (CWT) [72].

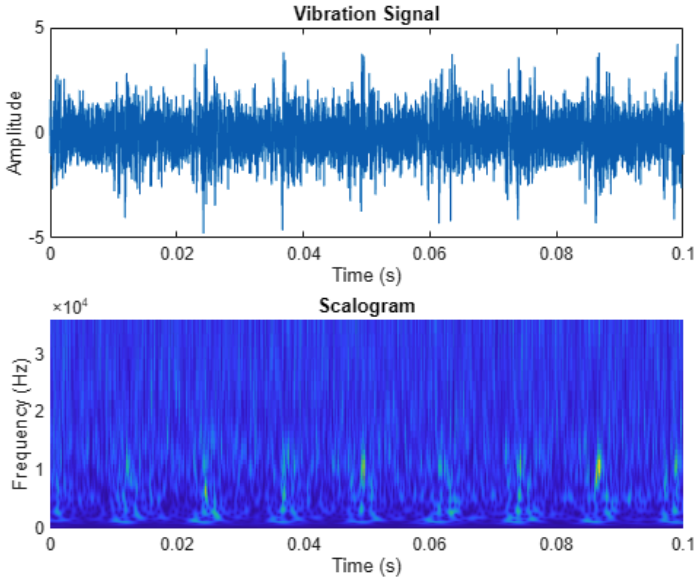


Figure 2.6: CWT scalogram of an outer raceway fault [54].

As illustrated in Figure 2.6, the CWT converts raw vibration signals into two-dimensional scalograms, capturing the temporal evolution of spectral energy across multiple frequency scales. This transformation captures time-frequency features, represents fault characteristics visually, and effectively separating noise [72]. By transferring the problem to the 2D domain, the diffusion framework can treat these scalograms as actual images, utilising stable and well-established computer vision architectures based on the two-dimensional U-Net topology.

Specifically, the integration of synthetic scalograms generated via DDPM into the CWRU dataset increased the diagnostic accuracy of a tailored 2D CNN from a baseline of 74%, achieved using exclusively real data, to 87.6% [72].

2.4.2 High-Diversity Conditional Generation and Transfer Learning

Although the baseline application of DDPMs on time-frequency scalograms provides a solid foundation for data augmentation, practical industrial diagnostics require strict control over the specific fault classes to be generated. However, generating conditioned data from limited empirical samples introduces a new generative challenge: the risk of synthesising highly homogeneous data distributions. This lack of statistical variance severely limits the generalisation capabilities of the downstream diagnostic classifier.

To overcome this limitation, a recent framework proposed the integration within the diffusion model of a Hybrid-Diversity (HD) loss function, an advanced metric that explicitly enforces sample diversity by incorporating penalties based on both intra-

class and inter-class cosine similarity during the reverse diffusion process [60]. As a result, the generated scalograms preserve the fundamental kinematic signatures of the specific bearing defect, while exhibiting unique stochastic variations that accurately reflect the complex background noise of real-world industrial environments.

In extreme *few-shot* scenarios, training a deep generative model from an extremely limited amount of empirical data is statistically impractical, making it necessary to bypass this bottleneck by employing a transfer learning strategy [60]. Following this methodology, the diffusion network is first pre-trained on a large-scale, publicly available bearing dataset to learn domain-invariant time-frequency representations. Subsequently, it undergoes a *fine-tuning* process on the sparse target dataset using a reduced learning rate. This paradigm effectively mitigates the limitations imposed by data scarcity, enabling the diffusion model to synthesise diverse, high-fidelity diagnostic samples [60].

This targeted data augmentation strategy, which effectively reduces the risk of overfitting by increasing the statistical variance of the training set, enables the downstream diagnostic classifier to achieve exceptional performance in fault detection [60, 72]. Nevertheless, a thorough analysis of these two-dimensional frameworks reveals significant computational and scalability limitations. The necessity to convert the entire vibration dataset into 2D scalograms prior to the training phase introduces a procedural overhead and adds considerable complexity to the overall system [72, 60]. This pre-processing burden, combined with the high computational cost and the inherent slowness resulting from the iterative sampling of DDPMs, constitutes an operational bottleneck. These critical issues confirm the methodological need to develop diffusion architectures capable of operating directly on raw one-dimensional time series, thereby bypassing the reliance on time-frequency transformations.

2.4.3 1D Diffusion and Multi-Source Domain Adaptation

In real-world industrial environments, machinery frequently operates under fluctuating kinematic conditions. Consequently, a diagnostic model trained on a specific operating regime is subject to the phenomenon of domain shift, suffering a drop in accuracy when tested at different speeds or loads.

To overcome these challenges, along with the computational limitations associated with 2D time-frequency transformations, recent research has proposed the application of diffusion models for the direct modelling of raw one-dimensional time series, employing the conditional denoising diffusion process to generate balanced shifted source domains [63]. By combining this generative step with Multi-Source Do-

main Adaptation (MSDA) techniques, the framework ensures that domain-invariant features can be extracted from multiple source domains [63].

A key innovation of this approach lies in the topological evolution of the denoising network. Standard DDPMs typically employ U-Net architectures based exclusively on CNNs. While CNNs are optimal for extracting local features in images, they exhibit limitations in capturing the long-range global correlations of vibration signals. To overcome this topological bottleneck, the Denoising Diffusion Multi-source Domain Adaptation (DDMDA) framework introduces a hybrid architecture called Utransnet [63]. By integrating a Transformer layer within the 1D U-Net topology, the network simultaneously leverages convolutions to extract local impulse transients, and self-attention mechanisms to capture the global periodicity of the raw vibration signal.

Once the data imbalance has been resolved through synthetic generation, the proposed framework addresses the domain shift through MSDA. A shared-weight feature extractor based on ResNet18 simultaneously processes raw data from the various domains. The extracted features are then fed into multi-level discriminators, whose task is to distinguish whether the data originates from the multiple source domains or the target domain [63]. To extract domain-invariant features, the overall network objective is to minimise the discriminator's accuracy. This adversarial process, aided by a gradient reversal layer, forces the extractor to confuse the discriminator, thereby aligning the statistical distributions of the different operating regimes. Finally, this mechanism ensures that the diagnostic features become domain-invariant, providing a robust and computationally efficient solution that operates directly on raw time-series data [63].

2.4.4 Spectral Integrity in 1D Diffusion

While the transition to 1D architectures significantly mitigates the computational burden associated with 2D time-frequency transformations, it introduces a novel physical challenge: the preservation of harmonic integrity. Unlike 2D representations, where frequency components are explicitly mapped as spatial features, standard 1D diffusion models rely primarily on the MSE calculated in the time domain. However, being a quadratic metric, this time-domain penalty disproportionately weights dominant low-frequency macro-dynamics due to their higher energy amplitude. Consequently, the model often struggles to capture low-amplitude micro-transients, inherently leading to the loss or oversmoothing of crucial high-frequency diagnostic signatures.

To overcome the limitations of standard diffusion models in capturing complex temporal dynamics, the Diffusion-TS framework [67] introduces a hybrid time-frequency optimisation paradigm. Furthermore, while standard models typically train the network to estimate the incremental noise $\boldsymbol{\epsilon}$ at each step, Diffusion-TS alters the predictive target to directly reconstruct the clean original sample $\mathbf{x}_0$ [67]. This formulation enables the direct integration of frequency-domain constraints into the training objective alongside the standard time-domain reconstruction loss.

By applying the FFT directly to the network's prediction $\hat{\mathbf{x}}_0(\mathbf{x}_t, t, \theta)$ and the ground truth $\mathbf{x}_0$, the network is optimised under a joint time-frequency constraint [67]. Alongside the standard time-domain reconstruction loss, the framework penalises the spectral discrepancy by calculating the L_2 loss directly in the frequency domain. This dual-domain constraint forces the model to simultaneously minimise temporal errors and preserve the harmonic spectrum, ensuring that the synthesised time series retain both the underlying periodic oscillations and the fine high-frequency transients [67].

2.4.5 Zero-Shot Generation and OOD Scenarios

In industrial practice, acquiring comprehensive vibration samples for every possible combination of mechanical load and rotational speed is an impractical task. Consequently, machinery frequently operates under previously unobserved kinematic conditions, giving rise to the complex problem of diagnostics in OOD scenarios.

To overcome this OOD diagnostic challenge, the Unknown Condition Diagnosis based on Diffusion Model (DiffUCD) framework has been introduced [18]. Instead of merely aligning known domains, DiffUCD leverages the generative capabilities of conditional diffusion models to synthesise vibration signal vectors corresponding to operating conditions never seen during training [18].

A crucial structural innovation of this framework lies in replacing the standard U-Net architecture with an advanced module called Condition-Guided Embedding UNet (CGE-UNet) [18]. Unlike models conditioned solely on discrete labels, CGE-UNet includes a condition encoder designed to project continuous physical parameters, such as rotational speed and applied mechanical load, along with the timestep embedding, into a condition vector. Subsequently, a cross-attention mechanism merges this condition vector, which acts as the Key and Value, directly with the encoded features extracted from the noisy state $\mathbf{x}_t$, which serve as the Query [18]. This gives the framework a zero-shot capability: the model becomes capable of generating the vibration signature $\mathbf{x}_0$ of a fault under operating conditions for which it has never received any empirical data, effectively generalising the physical dynamics learnt

from other known domains.

Finally, given the stochastic nature of OOD generation, the DiffUCD framework implements an unsupervised post-generation filtering module, called UCFilter [18]. This module first clusters the generated signals using the K-means algorithm and then evaluates the pairwise similarity distances. Specifically, it leverages dimensionality reduction principles by calculating the KL divergence between the joint probability distributions of the high-dimensional feature space and its low-dimensional mapping. By selecting the samples closest to the cluster centres, the filter discards poor quality signals and retains only those with the highest fidelity. Experimental tests confirm that diagnostic classifiers trained on these generated signals significantly outperform traditional generative architectures, such as GANs, proving to be an effective solution for the synthesis of diagnostic data in unknown operational scenarios [18].

Although frameworks such as DiffUCD have demonstrated the remarkable ability of DDPMs to synthesise vibration signals in OOD scenarios, they remain purely data-driven statistical estimators. A promising alternative is the development of the Physics-Informed Diffusion Models (PIDMs) [72, 5], which integrates physical constraints directly into the denoising network by penalising the residuals of the partial differential equations during training [5] or by guiding the sampling process [47].

Despite their potential, applying PIDMs to REB diagnostics faces severe practical limitations. Specifically, formulating precise differential equations for bearing dynamics is extremely difficult due to the chaotic rolling element kinematics, the unknown structural transfer functions across the machinery housing, and the noise induced by VFDs.

To overcome these specific limitations of industrial applications, the most scalable methodological evolution lies in adopting a condition-guided approach. Rather than relying on rigid differential equations, physical awareness can be dynamically injected into the denoising network through targeted conditioning. This effectively bridges the gap between pure statistical generation and physical coherence, offering an optimal compromise between diagnostic reliability, industrial scalability, and computational cost.

Chapter 3

Tools and Frameworks

"I believe that inside every tool is a hammer."

Adam Savage

The advancement in deep generative models applied to industrial diagnostics, particularly concerning the generation of vibration signals in the presence of bearing defects, heavily depends on the integration of efficient hardware and software solutions. The development of the proposed diffusion framework requires substantial computational resources due to the sequential nature of the sampling process, which requires 1,000 denoising steps, and the architectural complexity of the U-Net. Training these probabilistic models imposes strict requirements in terms of parallel computation and GPU memory capacity, essential respectively for limiting training times and for accommodating the massive intermediate activations and computational graphs required for backpropagation.

This chapter details the hardware resources that supported the computational load and provides an analytical overview of the software ecosystem employed for the training, hyperparameter optimisation, and evaluation of the conditional diffusion model.

3.1 Hardware Infrastructure

Training the proposed diffusion model from scratch required a high-performance remote computing infrastructure to overcome the hardware limitations of standard local machines, specifically in terms of Video Random Access Memory (VRAM) capacity and parallel processing capability.

The processing core of this infrastructure consists of a multi-GPU architecture

featuring two NVIDIA GeForce RTX 3080 units. Given the massive volume of matrix computations required to train the U-Net backbone, these graphics units enable the extensive parallelisation of tensor operations via the CUDA backend, significantly reducing the model's convergence time.

Each GPU is equipped with 10 GB of VRAM, providing the necessary capacity to process the high-dimensional tensors without exhausting the memory. Furthermore, 64 GB of system memory (RAM) enables efficient data caching and high-throughput Host-to-Device transfers, thereby preventing GPUs from sitting idle. Finally, the overall computational workflow is coordinated by an Intel Core i9-10900K processor (10 cores, 20 threads).

Since all interactions with the computational environment were conducted entirely remotely, access was established via the Secure Shell (SSH) protocol through the institute's Virtual Private Network (VPN) to comply with Information Technology security requirements. Moreover, to prevent long training sessions from being interrupted by sudden disconnections of the local client, model execution was encapsulated within persistent terminal sessions using the `tmux` multiplexer. Ultimately, the entire architectural and experimental development was tracked via `Git` version control, ensuring the robust management of the codebase.

3.2 Deep Learning Framework

The tailored diffusion model, including the 1D U-Net backbone and the reverse sampling process, was implemented entirely from scratch using PyTorch [40] (v2.9.0), the standard framework in the deep learning research community. The main advantage of this framework comes from its eager execution mode, which evaluates tensor operations directly at runtime and automatically builds the computational graph during each forward pass for subsequent backpropagation. This dynamic paradigm facilitates the implementation of the stochastic masking required by the CFG during training, allowing condition embeddings to be reset at runtime using native logical constructs, without the need to pre-allocate complex decision nodes in the graph.

Another significant advantage of PyTorch for this thesis purpose is its native `torch.fft` module, which supports the Autograd engine. This allows the spectral term of the developed hybrid loss function to be directly integrated without compromising the differentiability of the computational graph.

Supporting Libraries

To complete the deep learning framework, the project took advantage of Python's libraries for signal preprocessing, quantitative evaluation, and advanced data visualisation. Specifically, the `NumPy` library was utilised for data segmentation, global normalisation, and vectorial computations. The `scipy.io` module handled the extraction of the raw vibration signals from their native MATLAB (`.mat`) format, while `scipy.signal` was used to preprocess the raw data, specifically to downsample the signals through polyphase filtering and to apply zero-phase low-pass equalisation.

During the validation phase, `scipy.stats` was utilised to compute time-domain statistical features (such as kurtosis and skewness), while the `scikit-learn` library was employed for quantitative evaluation and classification assessment. Finally, visual inspection and interactive analysis in both the time and frequency domains were performed using `Plotly`, enabling detailed examination of generated signals and comparison with reference data.

3.3 Hyperparameter Optimisation: Optuna

The performance of complex generative architectures is highly sensitive to the choice of hyperparameters. Critical variables of the U-Net, such as the number of layers, the number of convolutional filters, and the size of the local self-attention windows, directly determine the generative fidelity of the model.

Given the high computational load of the proposed diffusion model, traditional exhaustive optimisation strategies, such as grid search, were deliberately discarded. Exploring the hyperparameter space through a rigid, brute-force grid would inevitably lead to infeasible training times and an inefficient use of hardware resources.

Similarly, random search was also ruled out as a viable alternative. Despite being highly parallelisable, this technique does not leverage information from prior trials, making its effectiveness heavily dependent on the inherent randomness of the hyperparameter selection process.

To overcome the limitations of both exhaustive and purely stochastic methods, the workflow was integrated with Optuna, an open-source framework dedicated to hyperparameter optimisation [2]. This choice was primarily driven by Optuna's implementation of Bayesian optimisation algorithms, which systematically explore the search space based on prior observations. Instead of selecting parameters at random, the system constructs a surrogate model that estimates the network's

performance over untested configurations. This approach naturally balances the exploration of unvisited hyperparameter combinations with the exploitation of regions already known to yield superior performance, guaranteeing a highly efficient and targeted search.

Specifically, the search space was sampled using the Tree-structured Parzen Estimator (TPE) algorithm. For each trial, the TPE models two probability density functions, $l(x)$ and $g(x)$, using KDE. The function $l(x)$ is fitted to the subset of hyperparameters associated with the best values of the objective function, while $g(x)$ models the complementary subset. The next configuration x to be tested is then selected by maximising the ratio $l(x)/g(x)$, thereby directing the search towards the most promising regions of the hyperparameter space.

Furthermore, Optuna was configured to apply an automatic early stopping mechanism for configurations that, even in the early stages, showed suboptimal performance compared to previous trials, thereby preventing the unnecessary waste of computational resources. Finally, to ensure the persistence and traceability of the experiment, the entire optimisation history was iteratively saved in a relational SQLite database, allowing the execution to be safely resumed in the event of unexpected interruptions.

Chapter 4

Dataset Selection

The diffusion model developed in this thesis was initially validated on the CWRU dataset [9], a recognised benchmark for evaluating deep learning architectures applied to the predictive maintenance of rolling bearings. This dataset provides an ideal baseline to test the generative capabilities of the model, as it features rigorously labelled, isolated defects of progressive severity.

In order to test the model's performance in operational contexts more representative of industrial reality, a separate training session was conducted using the EECSL dataset [32]. This choice enabled the validation of the proposed architecture even in the presence of inverter-driven motors. In such configurations, the vibration signals acquired using accelerometers positioned on the motor housing are heavily corrupted by both physical and electromagnetic noise. Consequently, this noise spectrum overlaps with the kinematic signature of bearing degradation. This reduction in the SNR complicates the generation of synthetic signals that accurately reflect real-world bearing dynamics.

4.1 CWRU Dataset

Experimental Setup

The CWRU experimental test stand, shown in Figure 4.1, consists of a 2 HP induction motor, a torque transducer, an encoder, a dynamometer, control electronics and accelerometers. The dynamometer is used to apply varying mechanical loads, while the torque transducer and encoder are employed to collect the operational data relating to the shaft torque and the actual rotational speed, respectively.

The motor shaft is supported by two test bearings: an SKF 6205-2RS JEM bearing at the Drive End (DE) and an SKF 6203-2RS JEM bearing at the Fan

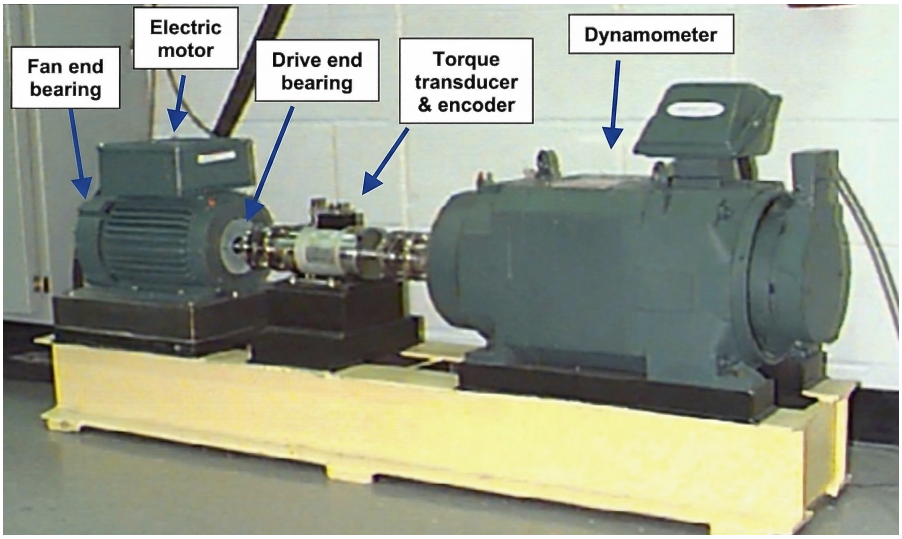


Figure 4.1: CWRU bearing test stand. Image adapted from [9].

Specification	*Drive End (DE) Bearing* (SKF 6205-2RS JEM)	*Fan End (FE) Bearing* (SKF 6203-2RS JEM)
Inside Diameter (in)	0.9843	0.6693
Outside Diameter (in)	2.0472	1.5748
Thickness (in)	0.5906	0.4724
Ball Diameter (in)	0.3126	0.2656
Pitch Diameter (in)	1.537	1.122
Number of Balls	9	8

Table 4.1: Geometric specifications of CWRU dataset bearings.

End (FE). The geometric specifications of both test bearings are summarised in Table 4.1, while the kinematic multipliers required for the analytical calculation of the characteristic fault frequencies are given in Table 4.2.

To replicate real-world mechanical degradation, single point faults were artificially introduced into the bearings using EDM. These localised defects were created separately on the inner raceway, the rolling elements, and the outer raceway, with increasing diameters of 0.007, 0.014, 0.021, 0.028 and 0.040 inches. Specifically, the original SKF bearings were utilised for defect sizes up to 0.021 inches, whereas equivalent NTN bearings were installed to accommodate the larger 0.028 and 0.040-inch faults.

Defect Frequency Multipliers	*Drive End (DE)* (SKF 6205-2RS JEM)	*Fan End (FE)* (SKF 6203-2RS JEM)
Inner Ring (BPFI)	5.4152	4.9469
Outer Ring (BPFO)	3.5848	3.0530
Rolling Element (BSF)	4.7135	3.9874
Cage Train (FTF)	0.39828	0.3817

Table 4.2: Characteristic fault frequency multipliers for the CWRU bearings. The actual fault frequencies are derived by scaling the motor rotational speed by the respective kinematic multiplier (k), such that $f_{\text{defect}} = k \cdot f_r$.

Data Acquisition and Organisation

Once the artificially faulted bearings were prepared, they were reinstalled into the test stand for the data acquisition phase. Vibration signals was acquired using accelerometers positioned vertically at 12 o'clock on both the DE and FE housings, with a supplementary sensor mounted on the motor supporting base plate (BA) only in specific experiments.

Data were collected using a 16-channel DAT recorder at sampling rates of both 12 kHz and 48 kHz for DE bearing fault experiments, while data on FE bearing faults were acquired exclusively at 12 kHz. In addition to the fault scenarios, a normal baseline dataset was collected under normal operating conditions exclusively at 48 kHz.

As detailed in Table 4.3, the experiments were performed at four different motor loads ranging from 0 to 3 Horsepower (HP), resulting in a progressive and slightly non-linear reduction in shaft speed.

The entire dataset is organised into MATLAB files, where each unique ID (e.g., `105`) corresponds to a specific experimental condition. For each condition, the vibration signals recorded by the sensors follow a standardised naming convention (`X105_DE_time`, `X105_FE_time`, and `X105_BA_time`).

Fault Kinematics and Vibration Signatures

Since single-point faults were introduced on three critical components: the inner raceway, rolling elements, and outer raceway, understanding the kinematic behaviour of these faults is crucial for interpreting the resulting vibration signatures. From a kinematic perspective, the outer raceway is a stationary component press-fitted into

Motor Load	*Rotational Speed*
0 HP	∼1797 RPM
1 HP	∼1772 RPM
2 HP	∼1750 RPM
3 HP	∼1730 RPM

Table 4.3: Approximate motor rotational speeds corresponding to the applied loads during the CWRU dataset experiments.

the stator housing. Consequently, the force of gravity acting on the rotor weight establishes a static and constant load zone located in the lower portion of the bearing. Since any defect on the outer race remains spatially fixed, the severity of its impacts with the rolling elements, and the resulting vibration response, depends strictly on its position relative to this load zone. To account for this physical phenomenon, experiments on outer raceway defects were conducted at specific angular positions: at 6 o'clock (centred directly in the maximum load zone), at 3 o'clock (orthogonal to the load zone), and at 12 o'clock (opposite to the load zone). When the rolling elements pass over a defect located at the 6 o'clock position, the severe impact generates high-amplitude impulsive responses in the vibration signal. Conversely, the 12 o'clock position produces much weaker impulses, as the rolling elements cross the defect with negligible contact force.

Unlike the outer raceway, the inner raceway is rigidly mounted to the motor shaft and consequently rotates at the same speed. A fault located on this component continuously enters and exits the static load zone during each shaft revolution, thereby causing a marked periodic amplitude modulation in the peaks of the resulting vibration signal.

The most complex kinematic behaviour is observed in faults affecting the rolling elements, as the balls simultaneously undergo an orbital motion around the motor shaft and a rotation around their own axes. Since a ball is free to rotate in any direction, a surface defect generates highly non-stationary impacts against both the inner and outer raceways, occurring at the Ball Spin Frequency (BSF). Furthermore, the amplitude modulation of these impacts is dictated by the Fundamental Train Frequency (FTF), or cage frequency, which describes the rate at which the specific faulty ball periodically enters and exits the load zone.

Despite these challenges, which make the vibration signatures of ball defects inherently difficult to isolate, a well-documented anomaly within the CWRU dataset

further complicates the diagnostic analysis. Specifically, the disassembly and reassembly procedures required to create the defects frequently caused unintended damage, known as brinelling, to the inner and outer raceways [48]. Consequently, the vibration patterns induced by these accidental raceway faults make it more challenging for the generative model to learn the true signature of the ball defect.

Data Selection Strategy

To assess the versatility of the proposed generative model, the experiments utilise vibration signals sampled at both 12 kHz and 48 kHz, enabling a comparative analysis of how the DDPM architecture responds to varying temporal resolutions, information densities, and intrinsic noise profiles.

In traditional vibration diagnostics, high sampling rates represent the standard industrial practice, as bearing faults generate broadband impulsive responses that excite structural resonances of the bearing and housing assembly, typically located at high frequencies. However, a comprehensive critical evaluation of the dataset [48] highlights distinct anomalies within the CWRU data sampled at 48 kHz. Specifically, spectral analysis reveals strong discrete components in the high-frequency band, primarily attributable to Electromagnetic Interference (EMI) induced by the dynamometer controller, as well as high-frequency harmonics caused by mechanical looseness in the test stand. Consequently, analysing data at 48 kHz becomes inherently more complex than analysing recordings at 12 kHz, where such high-frequency noise is naturally cut off by the Nyquist limit.

Moreover, high sampling rates pose significant challenges for the training of diffusion models, as the extreme temporal density at 48 kHz severely oversamples the background noise between consecutive impacts, diluting the useful kinematic information across a large number of samples. Conversely, the 12 kHz data provide greater temporal compactness, enhancing the network's capacity to capture the long-range periodicities associated with bearing faults. Validating the model on both sampling frequencies demonstrates that the proposed DDPM framework is capable not only of mapping fault signatures across different temporal resolutions, but also of faithfully replicating the specific noise characteristics of the test stand.

Given the decision to consider both sampling frequencies, the data selection focuses exclusively on DE bearing faults (acquired at both 12 kHz and 48 kHz), deliberately omitting FE bearing faults (acquired only at 12 kHz). Furthermore, focusing on the DE configuration reflects real-world industrial applications where the DE bearing exhibits a significantly higher probability of failure compared to the

Fault Diameter	*Fault Location*	*Motor Load (HP) / Speed (RPM)*			
		0 (1797)	*1 (1772)*	*2 (1750)*	*3 (1730)*
0.007"	Inner Race	105	106	107	108
	Ball	118	119	120	121
	Outer Race	130	131	132	133
0.014"	Inner Race	169	170	171	172
	Ball	185	186	187	188
	Outer Race	197	198	199	200
0.021"	Inner Race	209	210	211	212
	Ball	222	223	224	225
	Outer Race	234	235	236	237

Table 4.4: Summary of the selected CWRU vibration signals sampled at 12 kHz. Numbers indicate the unique ID of the `.mat` file.

FE counterpart.

For these selected faults, the vibration signals acquired by the DE sensor were considered. Selecting the DE sensor over the FE counterpart maximises the SNR, as the defect-induced shock waves propagate to its housing with minimal structural attenuation, providing the generative model with the clearest and highest-energy kinematic signatures available on the test stand.

Regarding the specific fault configurations, their selection was driven by the need to ensure class balance during the training phase. Specifically, for faults affecting the outer raceway, only recordings from the 6 o'clock position were extracted. This choice was dictated by a structural constraint within the original CWRU dataset: the 6 o'clock location is the only configuration offering a complete set of measurements across all three considered severity levels (0.007, 0.014, and 0.021 inches). Including the orthogonal (3 o'clock) or opposite (12 o'clock) positions would inevitably introduce missing data for the intermediate 0.014-inch diameter. Such a discrepancy would severely unbalance the dataset, thereby compromising the network's ability to learn the progressive evolution of mechanical degradation.

Finally, the detailed list of the vibration signals extracted from the CWRU dataset to create the training and validation sets is summarised in Table 4.4 and Table 4.5 for the 12 kHz and 48 kHz sampling frequencies, respectively.

Fault Diameter	*Fault Location*	*Motor Load (HP) / Speed (RPM)*			
		0 (1797)	*1 (1772)*	*2 (1750)*	*3 (1730)*
0.000" (Normal)	None	097	098	099	100
0.007"	Inner Race	–	110	111	112
	Ball	–	123	124	125
	Outer Race	–	136	137	138
0.014"	Inner Race	–	175	176	177
	Ball	–	190	191	192
	Outer Race	–	202	203	204
0.021"	Inner Race	–	214	215	217
	Ball	–	227	228	229
	Outer Race	–	239	240	241

Table 4.5: Summary of the selected CWRU vibration signals sampled at 48 kHz. Numbers indicate the unique ID of the `.mat` file.

4.2 EECSL Dataset

Experimental Setup

The EECSL experimental test stand consists of a four-pole induction motor (5.5 kW), mechanically coupled to a 30 kW DC generator to regulate the applied load. During the testing phase, the mechanical load is precisely regulated at 0%, 50%, 75%, and 100% of the rated capacity.

Different fault conditions were emulated during the start-up and steady-state operations. Specifically, steady-state data were acquired in the presence of various conditions: bearing defects, dynamic eccentricity, broken rotor bars, imbalance, misalignment, and structural looseness, alongside a nominal healthy baseline. The database also provides start-up transient recordings, but these dynamic profiles are strictly limited to the healthy state, imbalance, looseness, and misalignment conditions.

The operating conditions include a 60-Hz DOL power supply and a VFD [6], with the latter configured for constant volts-per-hertz open-loop control at 20, 40, and 60 Hz.

Data Acquisition and Organisation

During the experiments, both prototype and conventional industrial sensors were utilised. Specifically, the wireless prototype module contains exclusively a tri-axial MEMS capacitive accelerometer and a Hall-effect sensor to measure vibrations and radial leakage flux, respectively. Signals from these prototype sensors were acquired for 6 seconds at sampling rates of 800 Hz and 600 Hz, respectively [6].

Conversely, the conventional instrumentation provides a comprehensive measurement of the system's electromechanical state. It comprises three-axis piezoelectric accelerometers alongside 300-turn and 5-turn search coils to measure, respectively, radial leakage and air-gap magnetic fluxes. Furthermore, the electrical parameters, namely the three-phase line currents and line-to-line voltages, were continuously monitored using dedicated current transformers and voltage probes [6]. All signals from the conventional instrumentation were sampled at 10.24 kHz using a 16-bit data acquisition system, providing 50-second recordings for steady-state operations and 10-second recordings for start-up transients.

To facilitate systematic data retrieval, the database recordings are saved following a rigorous naming convention structured as follow:

`Field1_Field2_Field3_Field4`

where

- `Field1` = `Fault_type`
- `Field2` = `Power_supply_&_frequency`
- `Field3` = `Loading_%`
- `Field4` = `Measured_signal_type`

Data Selection Strategy

The primary data selection criterion involves isolating steady-state recordings from the start-up transients. This decision is fundamentally driven by a dataset limitation, as the start-up recordings do not encompass the bearing fault condition. Furthermore, this exclusion is supported by rigorous diagnostic reasoning: the severe mechanical and electromagnetic interference generated during the inverter's acceleration ramps would completely mask the weak transients caused by incipient bearing defects. Consequently, omitting these dynamic profiles prevents the neural network from

merely learning the macroscopic variations in speed, forcing it instead to focus on the characteristics of the defects.

The generative capabilities are tested on three distinct machine states: a nominal healthy baseline, a 0.6 × 4 mm (depth × length) outer race bearing defect (induced via EDM on a 6208 bearing), and a 30% dynamic rotor eccentricity [6].

The decision to include eccentricity, as well as the bearing fault, in the training set is motivated by a desire to test the versatility of the generative model on two anomalies that produce fundamentally different signatures in the frequency and time domains. The bearing defect is a purely mechanical anomaly, characterised by impulsive transients and sharp impacts capable of exciting structural resonances of the bearing components and the motor housing. In contrast, dynamic rotor eccentricity is an electromechanical fault arising from geometric misalignment of the rotor relative to the stator, creating a non-uniform air gap. Unlike bearing defects, eccentricity does not produce impulsive transients, but rather a continuous oscillation modulated at the rotor speed. This demonstrates the model's capacity to learn and generate realistic vibration signatures across fundamentally different fault mechanisms.

Finally, a single vibration channel was selected: the vertical radial axis, along which the impact energy is most effectively transmitted and captured. This choice ensures strict methodological consistency across the datasets. In the CWRU test stand, reference vibration data were exclusively measured using vertically oriented accelerometers. Extracting the same vertical signal from the EECSL database guarantees that any variation in the network's performance depends solely on the fault dynamics, completely avoiding distortions related to different sensor placements.

As summarised in Table 4.6, the selected vibration data include recordings under both inverter-fed power supplies (VFD) at 20, 40, and 60 Hz, and direct-on-line control (DOL) at 60 Hz, in order to validate the proposed diffusion model also in the presence of noise caused by the PWM modulation.

Fault Type	*Supply Type*	*Frequency (Hz) / Load (%)*		
		20 Hz	*40 Hz*	*60 Hz*
Healthy	VFD	0, 50, 75, 100	0, 50, 75, 100	0, 50, 75, 100
	DOL	-	-	0, 50, 75, 100
Eccentricity	VFD	0, 50, 75, 100	0, 50, 75, 100	0, 50, 75, 100
	DOL	-	-	0, 50, 75, 100
Bearing (0.6 mm)	VFD	0, 50, 75, 100	0, 50, 75, 100	0, 50, 75, 100
	DOL	-	-	0, 50, 75, 100

Table 4.6: Summary of the operating conditions selected from the EECSL dataset.

Chapter 5

Implementation

"Sometimes, the elegant implementation is just a function. Not a method. Not a class. Not a framework. Just a function."

John Carmack

This chapter details the complete development pipeline required to translate the theoretical foundations of diffusion models into a functional framework addressing the problem of data scarcity in diagnostic applications, as schematically illustrated in Figure 5.1.

It begins with the preprocessing of the raw vibration signals, encompassing polyphase decimation, spectral equalisation, tensor segmentation, and normalisation. Subsequently, it presents the development of the custom-designed conditional diffusion framework, including the architecture of the 1D U-Net, the hybrid loss function, the noise scheduling strategy, and the CFG-guided sampling process.

After outlining the training strategy and the hyperparameter optimisation phase, the chapter finally concludes by describing the evaluation framework designed to assess the morphological fidelity and diagnostic utility of the synthesised vibration signals, alongside the zero-shot generation capability of the proposed diffusion model.

5.1 Preprocessing of Vibration Signals

In this section, the issue of missing healthy baseline signals in the CWRU 12 kHz dataset is first addressed. Specifically, a polyphase decomposition algorithm combined with a Kaiser window for anti-aliasing is applied to the healthy 48 kHz signals in order to accurately downsample them to 12 kHz. Subsequently, a spectral discrepancy near

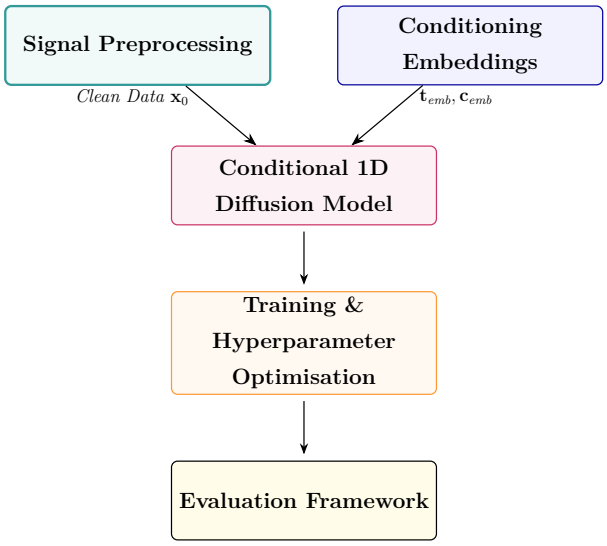


Figure 5.1: High-level overview of the proposed generative framework.

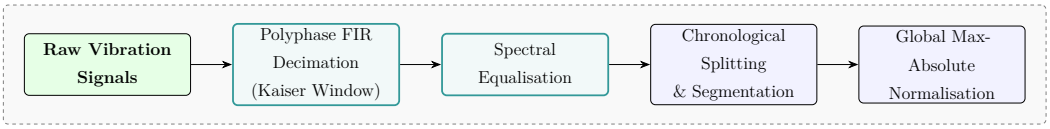


Figure 5.2: Preprocessing pipeline for the raw vibration signals. Blue blocks represent steps exclusively applied to the CWRU dataset.

the Nyquist limit, observed between the native 12 kHz signals and the downsampled ones, is corrected by applying a 5th-order Butterworth filter to ensure spectral consistency. Once a uniform dataset is established, the signals for each operating condition are chronologically split and partitioned into fixed-length segments. Finally, to optimise the performance and stability of the developed diffusion model, the segmented data underwent amplitude normalisation using the optimal technique identified for this study. All steps are summarised in Figure 5.2.

5.1.1 Polyphase FIR Decimation

The lack of native 12 kHz vibration signals reflecting machinery health conditions in the CWRU dataset necessitated their construction through the digital downsampling of the corresponding signals originally acquired at 48 kHz. This process of converting a digital signal from a higher sampling rate to a lower one is known as *decimation* [10]. According to the Nyquist-Shannon sampling theorem, reducing the sampling rate by a factor of $M = 4$ without aliasing requires pre-filtering the original signal with a low-pass filter whose cutoff frequency does not exceed the new Nyquist limit, defined

as:

$$f_c = \frac{f_s}{2M} = \frac{48 \text{ kHz}}{8} = 6 \text{ kHz} \tag{5.1}$$

If this anti-aliasing filter is omitted and one simply retains one sample out of every four, the high-frequency components incompatible with the new Nyquist limit would overlap into the baseband spectrum, causing the destructive phenomenon of aliasing and thus irreversibly compromising the integrity of the decimated signal [56].

An ideal low-pass filter has an impulse response of infinite duration in the time domain; since this is not realisable in practice, its response must be truncated. However, an abrupt truncation (rectangular windowing) generates the Gibbs phenomenon, which manifests as undesirable ripples at the band edges. To mitigate this phenomenon, an Finite Impulse Response (FIR) filter designed with the Kaiser window method is utilised. This technique avoids abrupt truncations by gradually smoothing the filter coefficients towards zero. The main advantage of the Kaiser window lies in the flexibility of its shape parameter, β, which, together with the window length, provides direct control over the trade-off between stopband attenuation and transition bandwidth [56]. In this framework, the standard value of $\beta = 5.0$ is adopted, corresponding to a stopband attenuation of approximately 54 dB.

The choice of an FIR architecture is primarily justified by its linear phase response, a property not easily achievable with Infinite Impulse Response (IIR) realisations, which is crucial to prevent phase distortion [56, 10]. By introducing a constant group delay across all frequency components, the FIR filter shifts the entire waveform in time without distorting its shape.

Once the FIR filter has been designed, applying it directly to the entire 48 kHz signal prior to downsampling would be highly inefficient, as 75% of the computed time-domain samples would be discarded. To optimise this process, the `scipy.signal.resample_poly` function is employed. Specifically, this polyphase decomposition algorithm reorganises the low-pass filter designed with the Kaiser window by dividing it into $M = 4$ polyphase components, which act as smaller sub-filters operating in parallel [56, 10]. The fundamental advantage of this structure is that it shifts the filtering operations to the lower, decimated sampling frequency (12 kHz) [10]. By calculating only the single valid output sample out of every four, the overall computational burden is reduced by a factor of $M = 4$, making the entire downsampling pipeline highly optimised.

Finally, a methodological challenge regarding the filtering of finite-length signals is addressed. During the convolution process, the filter requires access to data points beyond the signal boundaries. By default, the `resample_poly` function resolve

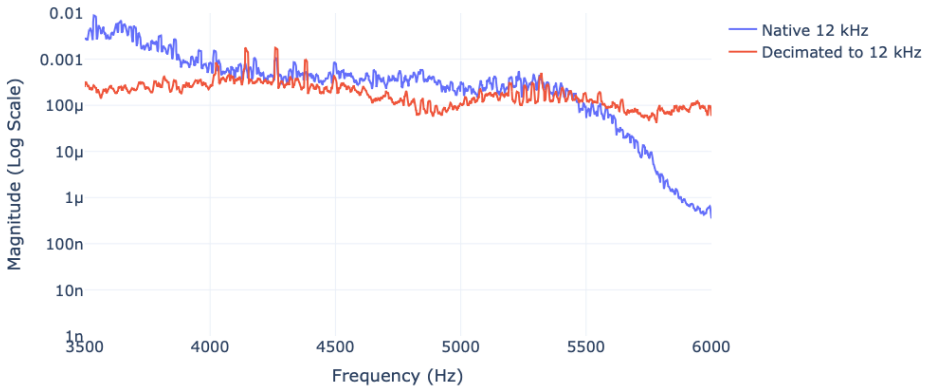


Figure 5.3: Magnitude spectra comparison before equalisation.

this by applying zero-padding. However, this sudden transition creates a sharp mathematical discontinuity that the filter interprets as a high-frequency step input, thereby inducing artificial transient oscillations known as edge-ringing. To mitigate this phenomenon, a linear boundary extension is applied prior to the filtering process. By mathematically extrapolating the initial and final trends of the vibration signal, the filter encounters a gradual, continuous ramp rather than a sharp step. This approach eliminates the artificial high-frequency components that traditional zero-padding would have introduced, ensuring that the signals obtained post-decimation remain free of numerical artefacts at the boundaries.

5.1.2 Spectral Equalisation

The combination, within the same training set, of 48 kHz vibration signals downsampled to 12 kHz and natively acquired 12 kHz vibration data introduces the risk of shortcut learning in the diffusion generative framework. This issue arises from the systematic correlation between the signal acquisition chain and the condition: all healthy baseline signals are processed through the digital decimation pipeline, while all fault signals are natively downsampled through the internal hardware decimation filter of the ADC. Consequently, the diffusion model may learn to associate each operating condition not only with its underlying physical vibration characteristics, but also with the spectral signature of the respective acquisition filter.

An empirical comparative analysis, conducted under identical operating conditions on the CWRU dataset (1 HP load, 1772 RPM), highlights a marked spectral mismatch near the Nyquist frequency (Figure 5.3). Specifically, the figure compares

the amplitude spectrum of a signal acquired natively at 12 kHz (ID 106, Inner Race) with that of a signal downsampled from 48 kHz to 12 kHz using polyphase decomposition and a Kaiser window FIR filter (ID 098, Healthy). To ensure a rigorous comparison, the analysis was performed on signal segments of equal length of 50,000 samples. The normalised magnitude spectra are computed after applying a Hanning window to each segment, thereby mitigating spectral leakage caused by edge discontinuities. Finally, to facilitate visual interpretation, a 50-point moving average is applied to the spectra. Given the frequency resolution of 0.24 Hz/bin, this operation results in a narrow 12 Hz smoothing window that effectively reduces local spectral variance without altering the macroscopic differences in the high-frequency profiles.

An examination of these magnitude spectra reveals that the native 12 kHz signal exhibits a steep spectral roll-off starting at approximately 5.3 kHz. In contrast, the digitally downsampled signal maintains a flat passband response across the full frequency range up to the transition band of the Kaiser FIR filter, centered near the 6 kHz Nyquist limit, retaining significantly more high-frequency energy than the natively acquired signal.

To rule out the possibility that the observed spectral discrepancy is attributable to fault-induced vibration signatures rather than hardware filtering, an experiment is conducted. A healthy signal and a faulty signal, both natively acquired at 48 kHz and subsequently downsampled to 12 kHz using the identical polyphase Kaiser FIR decimation pipeline, are compared in the frequency band between 5 and 6 kHz. Since both signals share the same digital acquisition chain and decimation filter, any spectral difference between them would reflect exclusively the physical effect of the fault condition. The two magnitude spectra exhibit no significant discrepancy in this critical frequency band, confirming that the roll-off observed in the natively acquired 12 kHz signals is due to the internal hardware anti-aliasing filter of the CWRU data acquisition system.

To eliminate this artificial spectral discrepancy, spectral equalisation was applied to the entire CWRU dataset at 12 kHz. To this end, a fifth-order Butterworth digital low-pass filter was implemented, with a cut-off frequency empirically set at 5.3 kHz. The main reason behind selecting the Butterworth filter lies in the mathematical nature of its frequency response, which exhibits a maximally flat and strictly monotonic trend in both the passband and the stopband [42]. This characteristic clearly distinguishes it from other classical solutions, such as Chebyshev IIR filters or optimal linear-phase FIR filters, which inherently introduce periodic

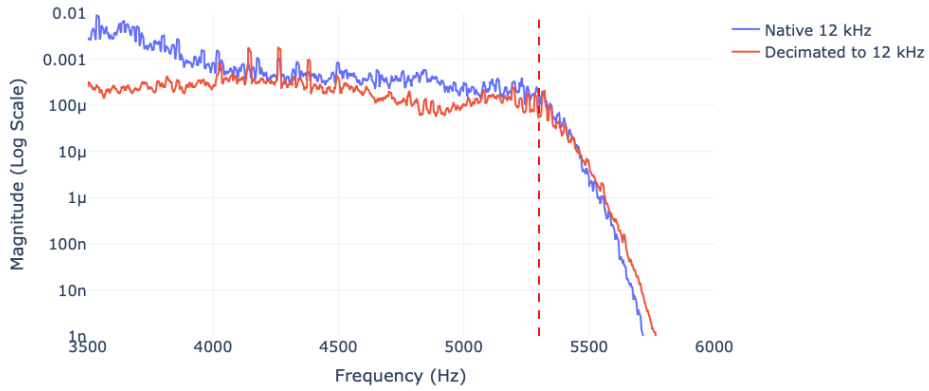


Figure 5.4: Post-equalisation amplitude spectra of the signals presented in Figure 5.3.

fluctuations, known as ripples, within the passband and the stopband [42].

Moreover, as an IIR architecture, the Butterworth filter offers superior computational efficiency compared to FIR designs [42]. Achieving an equally sharp roll-off at 5.3 kHz using an FIR filter would require designing a system with hundreds or thousands of coefficients. In contrast, the 5th-order IIR implementation achieves the required steep attenuation with negligible memory allocation and computational cost, making it the optimal choice for processing large datasets.

However, IIR filters possess a known disadvantage: they inherently introduce non-linear phase distortion. To neutralise this critical issue and preserve the exact temporal location of the impulse peaks, the filtering was performed using a forward-backward algorithm (`filtfilt`). This procedure achieves zero-phase filtering by first applying the IIR filter to the signal (forward pass), reversing the filtered signal in time, reapplying the same filter to the time-reversed signal (backward pass), and finally reversing the result to restore the original temporal order.

Ultimately, this approach successfully combines the maximal flatness of the Butterworth filter with the precise temporal alignment required to accurately capture bearing faults. The results of this equalisation are presented in Figure 5.4, which demonstrates a consistent spectral roll-off near the Nyquist limit for both signals, confirming the successful elimination of the acquisition-dependent discrepancy. Consequently, the representational capacity of the generative model is directed exclusively toward the physically relevant spectral features, free from the spurious discrepancies introduced by the different anti-aliasing filters.

Parameter (per file)	*12 kHz Model*	*48 kHz Model*
Signal Acquisition		
Total signal length (L)	121,991 samples	486,224 samples
Physical duration	$\sim$ 10 seconds	$\sim$ 10 seconds
Data Splitting		
Training Set (60%)	73,194 samples	291,734 samples
Validation Set (20%)	24,398 samples	97,244 samples
Test Set (20%)	24,399 samples	97,246 samples
Training Segmentation		
Window size (N)	1024 samples	4096 samples
Stride (S)	96 samples	384 samples
Total training tensors	**752**	**750**
Validation/Test Segmentation		
Window size (N)	1024 samples	4096 samples
Stride ($S = N$)	1024 samples	4096 samples
Total validation tensors	**23**	**23**
Total test tensors	**23**	**23**

Table 5.1: Summary of the chronological splitting and segmentation results for an individual CWRU operating condition.

5.1.3 Chronological Splitting and Tensor Segmentation

To generate the training, validation, and test sets for the diffusion model, the signals are partitioned chronologically. In particular, the initial 60% of the samples are allocated to the training set, the subsequent 20% to the validation set, and the remaining 20% to the test set. Following this chronological division, which ensures evaluation on strictly unseen, temporally subsequent data, the distinct sets are segmented into fixed-length tensors to meet the input requirements of the U-Net architecture. The detailed parameter specifications and the resulting tensor counts for an individual operating condition are summarised in Table 5.1. The remaining operating conditions follow the same partitioning scheme.

To evaluate the generative framework on the two different sampling frequencies available in the CWRU dataset (12 kHz and 48 kHz), the window size (N) and the stride (S) are scaled proportionally to guarantee that in both cases the model

analyses the exact same time window. Specifically, for the dataset operating at 12 kHz, a window size of $N = 1024$ samples is set, while for the 48 kHz dataset, this size is quadrupled to $N = 4096$ samples. This configuration yields a constant time window of approximately 85 ms for each extracted tensor. This duration is optimally sized to capture multiple complete mechanical cycles of the rotating shaft. Even under the dataset's most restrictive condition, namely the minimum rotational speed of 1730 RPM under maximum load where a single cycle lasts 34.6 ms, the selected time window successfully covers nearly two and a half mechanical shaft cycles. Furthermore, considering that the outer race fault represents the lowest-frequency kinematic anomaly evaluated, an 85 ms window guarantees approximately 9 complete fault cycles. This ensures that each extracted signal segment contains a statistically significant and highly periodic kinematic signature, providing the network with adequate contextual repetitions to learn the underlying fault patterns effectively.

For the training set, a highly overlapping sliding window approach is employed by setting a stride of 96 samples for the low-resolution signals and 384 samples for the high-resolution signals. This significant overlap serves as an effective data augmentation strategy. Conversely, the validation and test sets are extracted by imposing a strict non-overlap condition, setting the stride exactly equal to the window size.

The same chronological partitioning and sliding window segmentation logic was applied to prepare the EECSL dataset and the results are summarised in Table 5.2.

In addition to bearing faults, the analysis of the EECSL dataset considers rotor eccentricity. The diagnostic signature of this defect manifests at significantly lower frequencies, as it is intrinsically linked to the mechanical rotational speed and the electrical supply frequency. To effectively detect the spatial asymmetry of the air gap caused by eccentricity, the time window must encompass at least one full mechanical revolution of the rotor. To satisfy this requirement under the most restrictive condition, namely the minimum operational speed of a 20 Hz supply frequency, a window size of 2048 samples (200 ms) was selected. For a four-pole induction motor (two pole pairs), a supply frequency of 20 Hz yields a synchronous speed of $n_s = f/p = 20/2 = 10$ Hz.

However, due to the inherent slip of asynchronous motors operating under load, the mechanical rotational frequency of the rotor is lower than the synchronous speed. Consequently, a single mechanical cycle lasts slightly longer than 100 ms. Therefore, the chosen 200 ms time window is rigorously sized to successfully capture almost

Parameter (per file)	
Signal Acquisition	
Total signal length (L)	512,000 samples
Physical duration	50 seconds
Data Splitting (60% / 10% / 30%)	
Training Set	307,200 samples
Validation Set	51,200 samples
Test Set	153,600 samples
Training Segmentation	
Window size (N)	2048 samples
Stride (S)	512 samples
Total training tensors	**597**
Validation/Test Segmentation	
Window size (N)	2048 samples
Stride ($S = N$)	2048 samples
Total validation tensors	**25**
Total test tensors	**75**

Table 5.2: Summary of the chronological splitting and segmentation results for an individual EECSL operating condition.

two complete mechanical cycles even under this worst-case scenario. Conversely, at the maximum supply frequency of 60 Hz, the same window encompasses nearly six complete shaft revolutions.

The significantly longer acquisitions in the EECSL dataset compared to the CWRU dataset (50 s versus approximately 10 s) motivated two adjustments in data preparation. First, to ensure a more robust evaluation, the test set proportion is increased to 30%, with the validation set consequently adjusted to 10%. Second, a lower overlap ratio is sufficient to generate an adequate volume of training tensors. Consequently, a stride of 512 samples is adopted for the training set, successfully balancing computational efficiency with data augmentation. In contrast, for the validation and test sets, a strict non-overlap condition is maintained by setting the stride equal to the window size.

5.1.4 Amplitude Normalisation

The final stage of the preprocessing pipeline involves amplitude normalisation. In the field of vibration diagnostics, applying local normalisation, calculated and applied independently for each time window, would eliminate the information related to the signal's absolute energy. This would prevent the network from learning and reproducing the true physical severity of the faults, forcing the generative model to capture solely the geometric morphology of the waveforms. To preserve the relative physical intensity of the defect signatures across different operating conditions, a global scaling approach is strictly required.

After comparing various normalisation techniques, including global Z-score standardisation and global min-max scaling to the interval $[-1, 1]$, global max-absolute normalisation was selected as the most suitable method for the proposed diffusion model. This technique divides each sample x by the absolute maximum value $|X_{max}|$ computed exclusively on the training set and subsequently applied also to the validation and test sets to prevent data leakage, according to the following formulation:

$$x_{norm} = \frac{x}{|X_{max}|} \tag{5.2}$$

This approach provides crucial advantages. First, it maps the data within the $[-1, 1]$ interval, ensuring that in the reverse denoising process the network operates on consistently scaled inputs starting from the standard normal prior [22]. Furthermore, since vibration signals acquired with AC-coupled accelerometers exhibit zero mean by nature, the global max-absolute normalisation, being a purely multiplicative

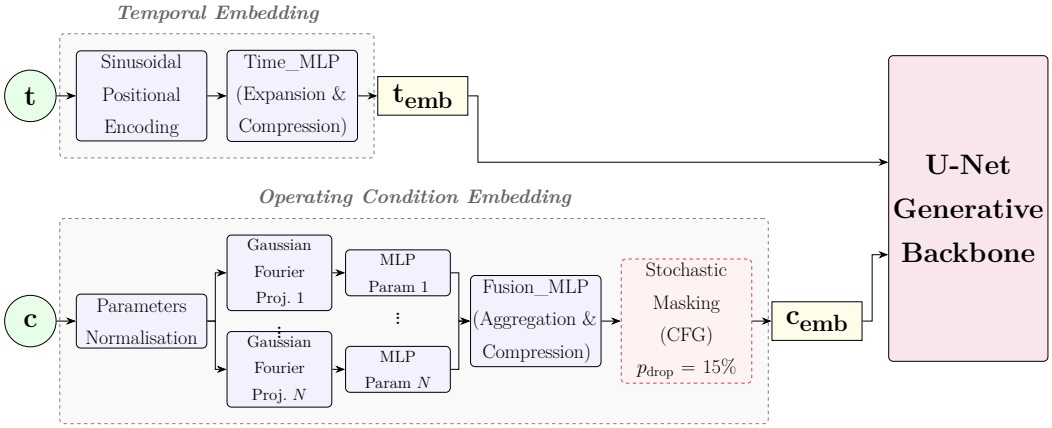


Figure 5.5: Conditioning pipeline for extracting the temporal ($\mathbf{t}_{\text{emb}}$) and operating condition ($\mathbf{c}_{\text{emb}}$) embeddings prior to their injection into the U-Net residual blocks.

transformation, preserves both the zero mean and the relative energy differences between fault types and severities.

5.2 The Conditioning Pipeline

To enable conditional signal generation, the diffusion framework incorporates a dedicated module that transforms the diffusion timestep and the operational variables into dense latent embeddings, which explicitly guide the U-Net during both the forward diffusion training and the reverse denoising inference.

Specifically, a dual-branch module is responsible for extracting the temporal ($\mathbf{t}_{\text{emb}}$) and conditioning ($\mathbf{c}_{\text{emb}}$) embeddings. These representations are subsequently injected into the U-Net residual blocks via the AdaGN modulation mechanism.

5.3 Operating Condition Embedding

In conditional diffusion models, the network's ability to generate signals consistent with specific operational states strictly depends on how the physical parameters are encoded and injected into the U-Net denoising architecture.

The developed framework implements a conditioning pipeline structured into three sequential stages: normalisation of the operating condition parameters, dimensional expansion via Gaussian Fourier projections, and feature fusion into the final conditioning embedding.

5.3.1 Parameter Normalisation

The physical variables describing the state and the operating conditions of the machinery exhibit heterogeneous units of measurement, numerical ranges, and orders of magnitude. To ensure numerical stability during training and prevent parameters with larger numerical ranges from disproportionately influencing the AdaGN modulation of the feature maps, each scalar parameter x is normalised within the continuous interval $[-1, 1]$ using a symmetric min-max scaling transformation:

$$x_{\text{norm}} = 2 \cdot \left(\frac{x - x_{\min}}{x_{\max} - x_{\min}} \right) - 1 \tag{5.3}$$

The key advantage of this linear transformation is that it rescales each conditioning parameter independently without distorting its internal distribution, preserving the relative distances between operating states within each parameter's domain. Since the available physical parameters heavily depend on the experimental setup, the composition and dimensionality of the conditioning vector depend on the specific dataset under analysis.

CWRU Dataset. To construct the conditioning vector, three distinct features are processed through the normalisation scheme defined in Eq. (5.3):

- **Rotational speed:** Normalised using the actual minimum and maximum operational limits recorded in the dataset, specifically $x_{\min} = 1730$ RPM and $x_{\max} = 1797$ RPM.
- **Defect diameter:** Scaled by setting $x_{\min} = 0.0$ inches and $x_{\max} = 0.021$ inches, respectively corresponding to the baseline defect-free state and the maximum severity level observed.
- **Kinematic fault frequency multiplier:** Rather than employing a discrete categorical treatment (e.g., one-hot encoding), the defect location (*inner race*, *outer race*, *ball*) is converted into its respective theoretical fault frequency multiplier. The assigned coefficients are 5.415 for the BPFI, 3.585 for the BPFO, 4.714 for the BSF, while the healthy baseline condition is assigned a multiplier of 0.0, representing the absence of any characteristic fault frequency. Consequently, the resulting continuous defect feature is normalised by setting $x_{\min} = 0.0$ and $x_{\max} = 5.415$. By replacing categorical labels with physically meaningful multipliers, this strategy provides the generative network with a direct quantitative reference for the characteristic fault frequencies.

EECSL Dataset. The conditioning vector integrates four operational features. While continuous variables are normalised according to Eq. (5.3), categorical variables are mapped into non-ordinal representations to prevent the network from inferring spurious ordinal relationships between independent fault topologies.

- **Supply frequency:** Scaled using the operational bounds $x_{\min} = 20$ Hz and $x_{\max} = 60$ Hz.
- **Motor load:** Normalised based on the motor's rated capacity, setting the bounds to $x_{\min} = 0.0$ % and $x_{\max} = 100.0$ %.
- **Fault class:** The discrete categories relating to the machine's status, namely *healthy*, *bearing fault*, and *eccentricity*, are encoded using a one-hot representation to prevent the network from inferring non-physical ordinal relationships between independent fault topologies. Consequently, this feature is expanded into a three-dimensional binary vector: $[1, 0, 0]$ for the healthy state, $[0, 1, 0]$ for bearing faults, and $[0, 0, 1]$ for eccentricity.
- **Power type:** The power supply method introduces distinct temporal and spectral signatures in the vibrational domain. This feature is encoded using a polar mapping: -1 for motors connected directly to the mains (DOL) and 1 for motors driven by inverters. Positioning these two states at opposite ends of the scalar range provides the network with a maximally contrasting representation of their distinct effects on the vibration signal.

5.3.2 Gaussian Fourier Projection

Following the normalisation process, initial experiments demonstrated that directly feeding the normalised conditioning vector into an MLP to obtain the final conditioning embedding proved ineffective for capturing the dependencies between the operating parameters and the resulting vibration signals.

Standard MLPs suffer from the phenomenon of spectral bias: the inherent tendency to primarily and rapidly learn low-frequency target functions, while severely struggling to resolve the high-frequency variations from low-dimensional inputs [52]. In this context, this overly smooth mapping translates into the network's inability to express the sharp transitions in the conditioning space that arise when even minimal variations in operating variables (e.g., a slight RPM shift) correspond to entirely distinct vibration signatures. Consequently, signals generated for different

operational conditions lose their defining features and become excessively similar to one another.

To overcome the spectral bias, the continuous input parameters x_{norm} are pre-processed through a *Gaussian Fourier projection* module before being fed into the MLP. This mapping operation expands each continuous scalar variable into a high-dimensional feature space through a set of sinusoidal basis functions [52], according to the following equation:

$$\mathbf{f}(x_{\text{norm}}) = [\sin(2\pi \mathbf{b} x_{\text{norm}}), \cos(2\pi \mathbf{b} x_{\text{norm}})]^T \tag{5.4}$$

In this formulation, $\mathbf{b} \in \mathbb{R}^m$ is a static, non-trainable weight vector, whose elements are sampled stochastically from an isotropic normal distribution $\mathcal{N}(0, \sigma^2)$. Once sampled, these values are fixed and serve as the fundamental frequencies of the projection.

A key advantage of Gaussian Fourier feature mapping is the precise control it allows through the hyperparameter σ [52]. Since the projection frequencies are sampled from $\mathcal{N}(0, \sigma^2)$, the standard deviation σ dictates the sensitivity of the MLP to variations in the input conditions.

Adjusting σ controls the trade-off between oversmoothing and over-sensitivity. A low σ confines the projection frequencies to a narrow band, causing the conditioning embedding to vary too slowly across the input space and blurring the distinctions between closely adjacent operational states. Conversely, an excessively high σ introduces projection frequencies of excessive magnitude, causing the embedding to vary too rapidly and leading the network to generate artefacts or inconsistent representations for operational conditions that should produce similar outputs.

In the developed framework, this property is exploited to apply domain-specific scaling. The primary conditioning variable, namely rotational speed for the CWRU dataset and supply frequency for the EECSL dataset, is deliberately initialised with a doubled standard deviation ($\sigma = 16.0$) compared to the remaining continuous physical parameters ($\sigma = 8.0$). This differential scaling strategy is driven by a specific physical consideration: rotational speed dictates the fundamental driving frequency of the system and acts as a multiplier for all characteristic fault frequencies. Consequently, even minimal fluctuations in the operational speed induce significant spectral shifts across the entire vibration spectrum. By selectively doubling the value of σ for this variable, the network is mathematically driven to develop higher sensitivity towards fluctuations in this critical kinematic parameter.

5.3.3 Late Feature Fusion

The processing of the conditioning vector follows a late fusion approach, where continuous and discrete physical variables are initially processed through independent paths.

Regarding the continuous variables, each parameter is first mapped via a Gaussian Fourier projection module into a high-dimensional feature vector, and subsequently processed by a dedicated MLP module, consisting of fully connected layers interleaved with the Sigmoid Linear Unit (SiLU) nonlinear activation function. Specifically, the model dynamically instantiates N Gaussian Fourier projection modules, the first with $\sigma = 16.0$ and the rest with $\sigma = 8.0$, alongside an equal number of MLP blocks, where N denotes the total number of continuous variables within the conditioning vector. By projecting and processing each parameter through an independent branch, the framework prevents the random Fourier weights from unpredictably mixing distinct physical variables before the MLP can extract a domain-specific representation for each operating condition.

Concurrently, discrete variables follow a separate branch where each categorical parameter is mapped into a dense, high-dimensional vector via a single linear layer paired with the SiLU activation function.

Once all the independent semantic representations have been extracted, the resulting tensors are concatenated and fed into a final aggregation module (`Fusion_MLP`). This module serves a twofold purpose. First, it models the non-linear physical interdependencies among the variables, capturing complex coupled dynamics such as the interaction between high motor load and severe fault conditions. Second, it applies progressive dimensional compression, reducing the concatenated tensor to the embedding dimensionality required by the U-Net.

The `Fusion_MLP` outputs a single integrated conditioning embedding $\mathbf{c}$, which encodes the global operating state of the machinery into a unified conditioning representation. Prior to its injection into the U-Net architecture, this vector undergoes a stochastic masking procedure that randomly replaces it with a null vector during training.

Conditioning Dropout for Classifier-Free Guidance

In conditional diffusion models, balancing sample diversity with fidelity to the imposed conditioning represents a fundamental challenge. To achieve an optimal trade-off, the proposed architecture adopts the CFG technique [23].

However, to implement the CFG, the denoiser must learn to estimate both the conditional denoising direction $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$ and the unconditional denoising direction $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$. To achieve this, the conditioning embedding $\mathbf{c}$ is subjected to stochastic dropout during training.

Specifically, in the proposed framework, the conditioning vector $\mathbf{c}$ is suppressed with a fixed probability of 15% ($p_{\text{drop}} = 0.15$), using a stochastic binary mask m. When suppression occurs, $\mathbf{c}$ is replaced by a *null token* $\boldsymbol{\emptyset}$, a trainable parameter vector optimised jointly with the remainder of the network parameters.

The masking operation yielding the effective operating conditions embedding $\mathbf{c}_{\text{emb}}$ injected into the U-Net is defined as:

$$\mathbf{c}_{\text{emb}} = m\boldsymbol{\emptyset} + (1 - m)\mathbf{c} \tag{5.5}$$

where $m \sim \text{Bernoulli}(p_{\text{drop}})$ is the binary scalar mask.

This strategy enables the network to adapt its learning objective according to the state of the mask. When the conditioning vector is active ($\mathbf{c}_{\text{emb}} = \mathbf{c}$), the U-Net learns to predict the noise conditioned on the specific operational state, modelling $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$. Conversely, when the null token is injected ($\mathbf{c}_{\text{emb}} = \boldsymbol{\emptyset}$), the network learns to predict the noise independently of any operating condition, modelling $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ and capturing the underlying structure of the vibration signals.

5.4 Temporal Embedding: Sinusoidal Positional Encoding

In the application of diffusion models to time series, the variable $t \in [1, T]$ does not represent the sampling instant of the physical signal, but rather the index of the iterative diffusion process. It is a discrete index that quantifies the level of artificial noise injected into the data at that specific diffusion step. Since the neural network's objective is to estimate the noise profile that varies at each timestep, it is essential to provide the U-Net with a unique and clear vectorial representation of the current level of degradation.

To this end, the scalar value t is mapped into a continuous, high-dimensional vector representation using a *sinusoidal positional encoding* [58, 22]. To construct this vector, sine and cosine functions are computed at different frequencies, mathematically defined as follows:

$$\begin{cases} PE(t, 2i) & = \sin\left(\frac{t}{10000^{\frac{2i}{D}}}\right) \\ PE(t, 2i+1) & = \cos\left(\frac{t}{10000^{\frac{2i}{D}}}\right) \end{cases} \tag{5.6}$$

where $i \in \{0, 1, \ldots, \frac{D}{2} - 1\}$ represents the vector component index. This formulation ensures that each value of i corresponds to two components, one sinusoidal and one cosinusoidal, effectively covering the entire dimensionality D of the embedding space.

This sinusoidal encoding yields a highly expressive, multi-scale representation of the timestep. The high-frequency components, associated with small values of i, vary rapidly and allow the network to distinguish between adjacent diffusion steps. The low-frequency components, associated with large values of i, vary slowly and provide macroscopic information regarding the global phase of the denoising process.

Furthermore, the smooth and structured nature of the sinusoidal formulation ensures that embeddings of adjacent timesteps remain geometrically close in the embedding space, providing the network with a consistent and well-ordered representation of the noise schedule that facilitates learning the progressive noise decay dynamics across the diffusion trajectory.

Following the deterministic sinusoidal positional encoding, the resulting tensor is processed by a *Time_MLP* module characterised by an expansion-compression architecture. The primary function of this module is to transform the deterministic sinusoidal encoding into a learned representation specifically adapted to the network's architecture and the denoising task.

The *Time_MLP* module comprises an initial linear layer that doubles the dimensionality of the sinusoidal timestep embedding, a SiLU activation function, and a final linear compression layer that restores the tensor to its original dimension. The expansion phase increases the representational capacity of the encoding, allowing the subsequent compression to extract the most relevant temporal features for guiding the denoising process.

5.5 The Generative Backbone: 1D Conditional U-Net

This section introduces the generative backbone of the diffusion model: a fully customised 1D Conditional U-Net designed to process vibration signals. To provide a clear overview, the discussion adopts a bottom-up approach. The discussion first examines the individual building blocks of the denoiser, laying out the theoretical and diagnostic rationale behind each component choice. Once these foundational elements are established, the focus shifts to the overall U-Net architecture.

5.5.1 The Residual Convolutional Block

The residual convolutional block serves as a fundamental building block of the proposed architecture, designed to extract local temporal patterns from the vibration signal while integrating the temporal and conditional embeddings through the AdaGN modulation mechanism. The overall architecture of this module is illustrated in the block diagram in Figure 5.6, while its specific design choices are detailed below.

Multi-Scale Feature Extraction via Dilated Convolutions

Within the residual convolutional block, 1D dilated convolutions are employed to expand the Receptive Field (RF) without increasing the number of trainable parameters or losing the temporal resolution of the signal [66].

A dilated convolution introduces a dilation factor d that expands the kernel's receptive field, creating a gap of $d-1$ spaces between consecutive kernel weights. For an input signal $\mathbf{x}$ and a filter $\mathbf{w}$ of length k, the operation evaluated at index i is defined as:

$$\mathbf{y}[i] = \sum_{j=0}^{k-1} \mathbf{x}[i + d \cdot j] \cdot \mathbf{w}[j] \tag{5.7}$$

Consequently, the RF of a single convolutional layer is no longer constrained solely by the kernel size, but also depends on the chosen dilation factor.

Specifically, the U-Net architecture adopts a mirrored base-2 exponential dilation scheme ($d = 2^i$), where the dilation factor increases progressively with the depth of the encoder layers, reaching its maximum at the bottleneck, and subsequently decreases along the upsampling path. This dilation strategy operates in perfect synergy with the U-Net downsampling operations, implemented via strided 1D convolutions (stride = 2). As each downsampling stage halves the temporal dimension of the feature maps, the dilated convolutions expand the RF to cover increasingly wider temporal windows. Consequently, this mechanism allows the shallower layers to capture localised micro-transients, while the deeper layers aggregate these into broader temporal patterns.

Finally, to maintain the temporal dimension of the output tensor equal to that of the input, the padding is computed at initialisation based on the dilation factor:

$$\text{padding} = \left\lfloor \frac{(k-1) \cdot d}{2} \right\rfloor \tag{5.8}$$

This temporal alignment ensures that the signal lengths remain consistent across the block, thereby enabling the proper execution of the residual shortcut connection.

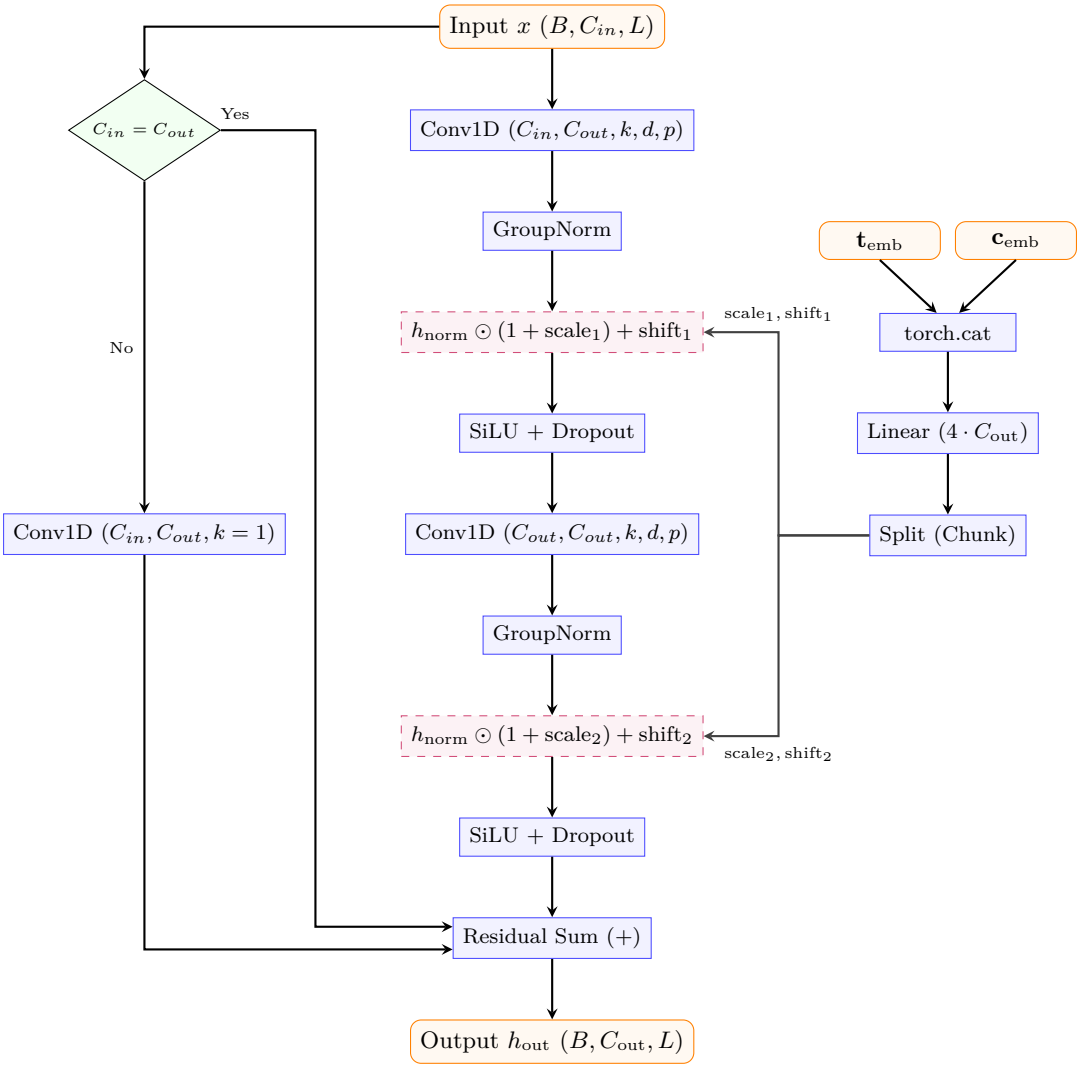


Figure 5.6: Architecture of the Residual Convolutional Block with dual-stage AdaGN conditioning. The temporal ($\mathbf{t}_{\text{emb}}$) and operational ($\mathbf{c}_{\text{emb}}$) embeddings are concatenated and linearly projected to generate the scale and shift parameters, which are used to modulate the normalised intermediate feature maps.

Feature Normalisation and AdaGN

Once dilated convolutions extract the feature maps, they must be normalised to facilitate the training process. In diffusion models, the substantial memory requirements of processing high-dimensional inputs through deep architectures often necessitate small batch sizes, making the application of standard Batch Normalisation (BN) highly suboptimal. Calculating the mean and variance over a limited number of samples yields unstable statistical estimates, ultimately compromising the network convergence [61]. More critically, during training each sample within a batch corresponds to a different noise level; BN would incorrectly compute statistics by mixing samples at different degradation levels, introducing a systematic bias in the normalisation.

To overcome both issues, the proposed framework adopts Group Normalisation (GN). Instead of normalising data across the batch dimension, GN operates on individual samples by dividing the channel axis into independent groups (set to $G = 8$). Within each group, the mean $\boldsymbol{\mu}_{\text{group}}$ and standard deviation $\boldsymbol{\sigma}_{\text{group}}$ are calculated locally to obtain the normalised activation:

$$\mathbf{h}_{\text{norm}} = \frac{\mathbf{h} - \boldsymbol{\mu}_{\text{group}}}{\sqrt{\boldsymbol{\sigma}^2_{\text{group}} + \epsilon}} \tag{5.9}$$

This approach ensures robust normalisation and makes the block's performance independent of the batch size. Following this local normalisation, the network must integrate the conditioning information. Based on preliminary architectural experiments, simple concatenation of the conditioning vectors to the block input was found to yield inferior generative performance compared to the AdaGN modulation approach. Consequently, the temporal ($\mathbf{t}_{\text{emb}}$) and the conditional ($\mathbf{c}_{\text{emb}}$) embeddings are leveraged to modulate the normalised feature maps through the Adaptive Group Normalisation (AdaGN) mechanism [37].

To this end, $\mathbf{t}_{\text{emb}}$ and $\mathbf{c}_{\text{emb}}$ are first concatenated and processed by a linear layer which projects them into a higher-dimensional space. In contrast to standard single-stage approaches [12], the proposed implementation adopts a two-stage modulation strategy, partitioning the projected tensor into four distinct modulation parameters: two scale factors ($\mathbf{scale}_1, \mathbf{scale}_2$) and two shift factors ($\mathbf{shift}_1, \mathbf{shift}_2$). These parameters integrate the conditioning information directly into the intermediate feature maps by modulating the normalised activations, $\mathbf{h}_{\text{norm}}$, according to the following formulation:

$$\text{AdaGN}(\mathbf{h}_{\text{norm}}, \mathbf{scale}, \mathbf{shift}) = \mathbf{h}_{\text{norm}} \odot (\mathbf{1} + \mathbf{scale}) + \mathbf{shift} \tag{5.10}$$

This dual-stage application within each residual block ensures a fine-grained integration of the conditioning context. Guided by the specific operational conditions and the current diffusion timestep, the first modulation ($\mathbf{scale}_1, \mathbf{shift}_1$) performs an initial alignment of the features extracted by the primary convolutional layer, while the second modulation ($\mathbf{scale}_2, \mathbf{shift}_2$) refines the patterns captured by the subsequent convolution.

Activation and Residual Connection

Once normalised and conditioned, the tensor is processed by the SiLU activation function, defined as $f(x) = x \cdot (1 + e^{-x})^{-1}$. SiLU was chosen for its continuous differentiability, which facilitates a smoother gradient flow compared to the conventional Rectified Linear Unit (ReLU).

The last architectural design choice within the convolutional block is the residual connection, which incorporates a pointwise convolution to align the channel dimensions when the input and output sizes differ [20].

The fundamental advantage of employing the shortcut connection is that the convolutional block, rather than attempting to directly approximate the complex target mapping $\mathcal{H}(\mathbf{x})$, focuses exclusively on learning the residual function $\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$. This approach significantly facilitates network optimisation and effectively overcomes the degradation problem, a phenomenon observed in traditional plain architectures where, as network depth increases, performance saturates and then rapidly degrades [20]. Without residual connections, optimising the proposed deep U-Net architecture would have been significantly more challenging.

5.5.2 1D Self-Attention Mechanisms: Local and Global

While the use of dilated convolutions effectively expands the receptive field, purely convolutional operators remain intrinsically limited in modelling long-range data dependencies. Furthermore, significantly increasing the dilation factor without a corresponding increase in kernel parameters can trigger the so-called *gridding effect*, a phenomenon that restricts sampling to isolated points, thereby ignoring the intermediate context.

To overcome this limitation, the architecture pairs dilated convolutions with a Local Self-Attention mechanism, which densely analyses the context within a restricted temporal window. This hybrid approach successfully leverages the strengths of both paradigms. On one hand, dilated convolutions provide a strong inductive bias and translation invariance, which are ideal for extracting local physical patterns, such

as the high-frequency impact peaks typical of bearing faults. On the other hand, Local Self-Attention computes input-adaptive attention weights by evaluating pairwise interactions between all samples within the local window. By densely processing every data point in that region, the model effectively counteracts the sparse sampling of dilated convolutions, precisely capturing the morphological signature of an anomaly and recognising the coupling between closely spaced signal events, even under slight temporal or phase shifts.

The inherent limitation of these localised operations lies in their restricted temporal horizon, which prevents them from capturing long-range data dependencies. However, in the proposed architecture, this constraint is treated as a deliberate design choice. In the shallower layers of the network, localised processing is intentionally exploited to focus the high temporal resolution on the specific morphology of individual impulsive transients. Conversely, the modelling of macro-periodicities and the global dynamics of the signal is delegated to a 1D global attention module, strategically positioned at the bottleneck of the U-Net where the time sequence undergoes the greatest downsampling and the features reach their highest level of abstraction.

Local Self-Attention

Following each residual convolutional block, the architecture integrates a 1D Local Self-Attention module, which adapts the Window-based Multi-Head Self-Attention (W-MSA) [35] paradigm to the one-dimensional domain.

The input tensor is first zero-padded to ensure its temporal length is perfectly divisible by the window size W. It is then partitioned into fixed, non-overlapping temporal windows of size W (in our configuration, $W = 128$). Attention is calculated independently within each individual partition, thereby restricting each element to attend exclusively to other elements within its own local window. By constraining the field of view to this fixed time interval, the computational complexity of the attention mechanism is drastically reduced from $\mathcal{O}(L^2)$ to $\mathcal{O}(L \cdot W)$, which scales linearly with respect to L for the fixed window size W.

Within each local window, the extraction of temporal dependencies proceeds in sequential steps. First, the windowed signal segment acts as the input for a single linear projection that expands the channel dimension by a factor of three. This expanded tensor is subsequently chunked into the *Query* (Q), *Key* (K), and *Value* (V) matrices.

Rather than computing a single global attention map, the module employs a multi-

head formulation to simultaneously focus on different spectral and morphological features within the same window. To this end, the Q, K, and V matrices are divided across $h = 8$ parallel attention heads. Consequently, each head operates in a reduced feature subspace of dimension $d_k = C/h$ (where C is the total channel depth).

Finally, within each independent head, these dimensionally reduced matrices are fed into the standard Scaled Dot-Product Attention mechanism [58], which computes the output as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.11}$$

The dot product QK^T computes pairwise similarity scores between temporal positions within the local window. However, its variance scales proportionally with d_k, producing values of large magnitude that push the softmax into saturated regions with negligible gradients. To prevent this, the scaling factor $\sqrt{d_k}$ is applied to normalise the dot product variance, thereby preserving numerical stability and ensuring effective backpropagation.

Finally, the output from all parallel heads is concatenated, the partitioned windows are merged to restore the original sequence length L, and the tensor is passed through a final linear projection.

Global Self-Attention

To capture the long-range dependencies and the macro-periodicity of the vibration signal, the diffusion framework employs a 1D Global Self-Attention module. This module leverages the same Multi-Head Attention (MHA) paradigm used in the local module, but it removes the window constraint to compute attention across the entire time sequence.

This unconstrained approach requires the calculation of the dot product between every pair of temporal samples, resulting in a quadratic space and time complexity of $\mathcal{O}(L^2)$. Considering input sequences up to 4096 samples, applying global attention in the shallower layers would yield attention matrices of up to 4096×4096 elements, making the operation computationally prohibitive. The global self-attention mechanism is therefore positioned exclusively at the bottleneck, where successive downsampling operations have reduced the sequence length to its minimum.

Operating on the compressed features at the bottleneck, the global attention module correlates data points across the full temporal horizon. Because this is a self-attention mechanism, the *Query* (Q), *Key* (K), and *Value* (V) matrices are all generated from the same input tensor. These matrices are then divided along

the channel dimension to feed $h = 4$ parallel attention heads, whose outputs are subsequently concatenated and passed through a final linear projection W^O:

$$\text{MHA}(Q, K, V) = \text{Concat}(\mathbf{head}_1, \ldots, \mathbf{head}_h)W^O \tag{5.12}$$

Finally, to stabilise training and ensure numerical regularity, the module includes a residual connection and a Layer Normalisation stage:

$$\mathbf{h}_{\text{out}} = \text{LayerNorm}(\mathbf{h}_{\text{in}} + \text{MHA}(Q, K, V)) \tag{5.13}$$

From a diagnostic perspective, placing the global attention mechanism at the bottleneck perfectly compensates for the restricted temporal scope of the earlier layers, allowing the model to capture the underlying macro-periodicity of the vibration signal.

5.5.3 Overall Architecture: The Conditional 1D U-Net

Having established the foundational building blocks of the U-Net denoiser, the focus now shifts to its overall macro-architecture.

Designed to map a noisy time series $\mathbf{x_t}$ to an estimate of the noise $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$, the network adopts a symmetric topology divided into three macro-stages: the contraction path (*encoder*), the deep feature processing (*bottleneck*), and the expansion path (*decoder*). A distinctive structural feature of this network is the distributed conditioning strategy. As previously detailed, the temporal and conditional embeddings do not act as mere global inputs; rather, they are injected with the AdaGN operator into every residual convolutional block across the encoder, bottleneck, and decoder. This distributed injection ensures that feature modulation occurs consistently across all temporal resolution scales.

In the initial contraction path, the signal passes through a modular structure whose depth is dynamically defined by the `num_layers` parameter. As the signal flows through these layers, its temporal resolution is progressively reduced in favour of higher feature abstraction along the channel axis. Feature extraction is performed through the combined action of dilated convolutions and local self-attention modules, both equipped with residual connections. Prior to each spatial reduction, the high-resolution feature maps are preserved in memory to be forwarded via skip-connections during the subsequent decoding phase. Furthermore, rather than relying on standard max-pooling operators, which discards non-maximal activations and lacks learnable parameters, the architecture performs downsampling via learnable strided convolutions.

At the end of the contraction path, the tensor reaches the bottleneck, the stage characterised by the lowest temporal resolution and the highest semantic density. Processing begins with a non-dilated residual convolutional block. The output is then sequentially processed by a local self-attention module and a global self-attention module. From a diagnostic perspective, this dual-attention topology allows the model to simultaneously capture local impact morphologies and global fault-driven periodicities. Recognising these underlying physical structures within the noisy input, the network can accurately predict the stochastic noise component $\hat{\boldsymbol{\epsilon}}$ superimposed on the signal.

The expansion path projects the deep latent representations back to the signal's original temporal resolution. The upsampling process relies on transposed convolution operators (*ConvTranspose1d*), which double the temporal resolution while halving the number of channels. To compensate for the loss of fine-grained, high-frequency components caused by the previous downsampling operations, the architecture integrates *skip connections*. Specifically, the tensor propagating from the bottleneck is concatenated along the channel axis with the corresponding high-resolution feature map previously preserved by the encoder. This fusion integrates two complementary information streams: the highly abstract, long-range representations encoded by the bottleneck and the fine-grained morphological details preserved by the encoder's skip connections.

Finally, the output of the decoder is processed by a terminal convolutional layer with a kernel size of 1. This operation produces a tensor of identical dimensions to the noisy input $\mathbf{x_t}$, representing the final noise estimate $\hat{\boldsymbol{\epsilon}}$.

5.6 Denoising Diffusion Probabilistic Model

To adapt the standard diffusion framework to the characteristics of the vibration domain, the proposed implementation deviates from standard formulations by introducing three key architectural enhancements: a non-linear noise variance schedule, a hybrid time-frequency loss function, and a conditional sampling process driven by CFG.

5.6.1 Cosine Noise Scheduling

The forward corruption process employs a cosine noise schedule [37]. Unlike the original linear schedule, which tends to cause a premature collapse of the SNR, this squared-cosine scheme ensures a much more gradual information decay. Specifically,

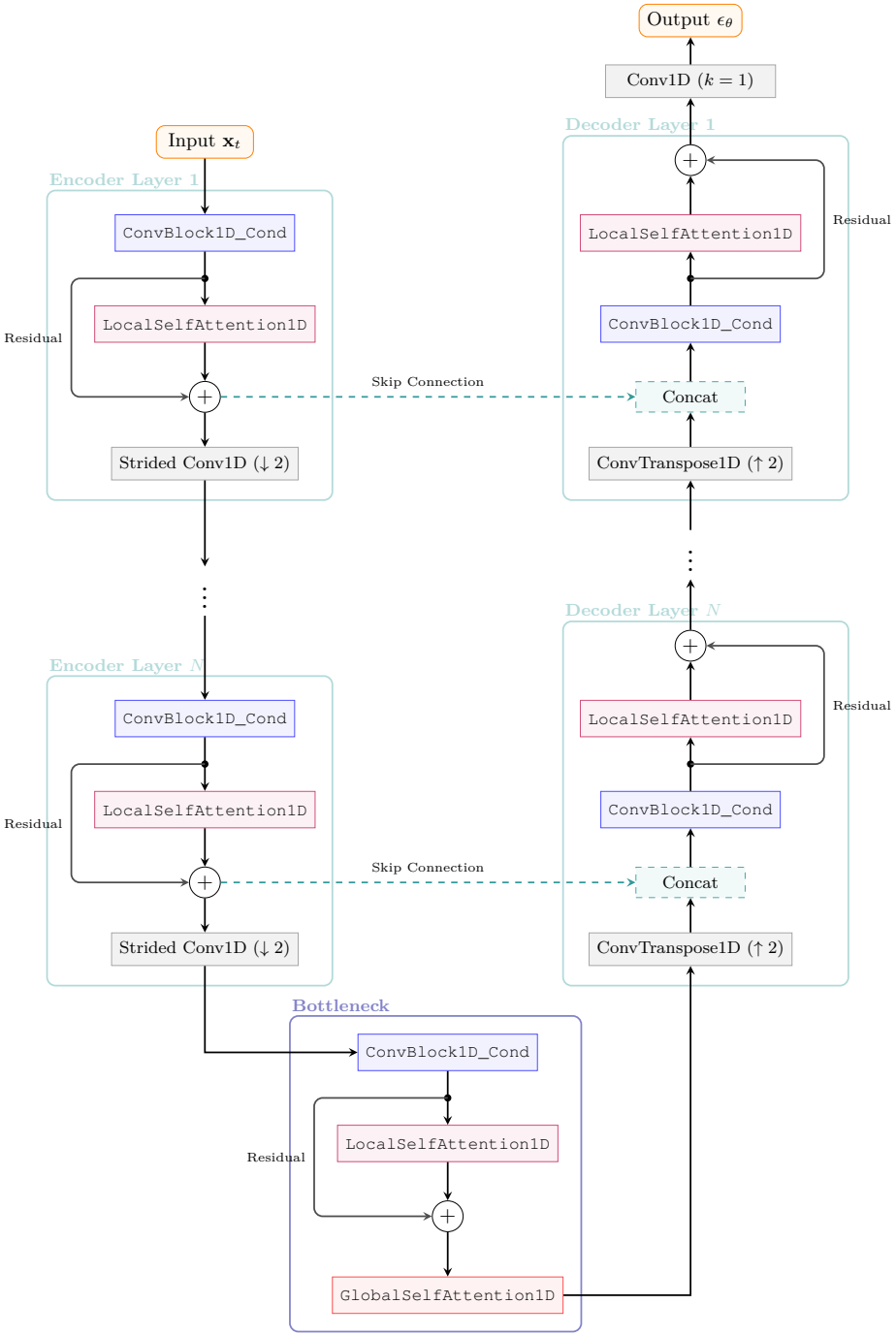


Figure 5.7: Architecture of the Conditional 1D U-Net denoiser mapping the noisy input $\mathbf{x}_t$ to the estimated noise $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$. The encoder extracts multi-scale features through the residual convolutional blocks (Fig. 5.6) and local self-attention modules. The bottleneck applies global self-attention to capture long-range temporal dependencies. The decoder restores temporal resolution via skip connections that merge high-resolution encoder features with the deep bottleneck representations.

it yields a nearly linear decay of information in the middle of the corruption process, while limiting rapid changes in the noise levels at the extremes.

Preliminary tests conducted in this work demonstrated that this schedule yields superior generative performance compared to both linear and sigmoid schedules. In a diagnostic context, preventing an abrupt early drop in SNR is crucial to guarantee the structural coherence and temporal localisation of impulsive transients. Preserving these components for a longer duration of the diffusion process allows the neural network to effectively learn their underlying morphological features.

Mathematically, the corruption process is governed by a time-dependent function $f(t)$, analytically defined as:

$$f(t) = \cos^2\left(\frac{t/T + s}{1+s} \cdot \frac{\pi}{2}\right) \tag{5.14}$$

where T represents the total number of diffusion steps, and $s = 0.008$ is a small constant offset introduced to prevent the initial noise injection from being practically zero. From this function, the cumulative parameter $\bar{\alpha}_t$ is computed to determine the fraction of the original signal preserved at step t:

$$\bar{\alpha}_t = \frac{f(t)}{f(0)} \tag{5.15}$$

Consequently, the variance of the noise added at each step, denoted as β_t, is derived as follows:

$$\beta_t = 1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}} \tag{5.16}$$

To ensure numerical stability during training, the calculated values of β_t are typically clipped to a maximum value of 0.999. This crucial implementation detail prevents singularities near the end of the diffusion process, where $\bar{\alpha}_t$ approaches zero.

Finally, using the cumulative parameter $\bar{\alpha}_t$, the direct sampling of the noisy signal $\mathbf{x}_t$ at any arbitrary timestep t from the clean input $\mathbf{x}_0$ is expressed in closed form as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{5.17}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the standard Gaussian noise that the neural network is trained to estimate.

5.6.2 Hybrid Loss Function

The optimisation of the U-Net parameters is driven by a loss function formulated to simultaneously minimise the error in both the time and frequency domains. The temporal component is defined by the MSE calculated between the injected Gaussian

noise $\boldsymbol{\epsilon}$ and the denoiser's estimate $\hat{\boldsymbol{\epsilon}}$. Minimising the noise MSE is mathematically equivalent to minimising a timestep-weighted reconstruction error of the clean signal $\mathbf{x}_0$ [22].

According to Parseval's theorem, the total signal energy in the time domain is equivalent to that in the frequency domain [39], meaning that the temporal MSE implicitly penalises the aggregate spectral error. However, this global energy equivalence does not guarantee that individual frequency components are accurately reconstructed, particularly when the error is concentrated in sparse, high-amplitude spectral regions.

A further limitation of the temporal MSE arises from its strict sensitivity to the phase: if the network generates a morphologically accurate transient but misaligns it even slightly in time, the squared penalty is severe. To minimise this penalty, the model tends to produce more conservative, over-smoothed signals, suppressing the impulsive transients that are critical for fault diagnosis, as observed in preliminary experiments conducted during the architectural design phase.

To overcome this limitation, the developed loss function integrates an explicit, phase-invariant penalty in the frequency domain. During each training step, the model uses the noise prediction and the corrupted tensor $\mathbf{x}_t$ to derive a closed-form estimate of the original signal $\hat{\mathbf{x}}_0$ (as shown in Code 5.1). The amplitude spectra of both the original signal and its estimate are then calculated using the *real Fast Fourier Transform* (rFFT) and, finally, the reconstruction error is quantified by applying the Mean Absolute Error (MAE) between the logarithmic magnitude spectra of $\hat{\mathbf{x}}_0$ and that of the target signal $\mathbf{x}_0$.

The combined use of the MAE metric and the logarithmic scale addresses two specific diagnostic needs. While the resonance peaks excited by fault impacts carry diagnostically relevant information on the severity, the features that discriminate between fault types and locations are encoded in smaller-amplitude components: low-order fault-related harmonics and modulation sidebands surrounding the structural resonances. By compressing the dynamic range of the spectrum, the logarithmic transformation prevents large resonance peaks from dominating the loss gradient. Furthermore, using the MAE rather than a quadratic metric helps to stabilise the gradients during training. Since the error grows linearly rather than quadratically, this metric is much more tolerant of large spectral outliers, such as spurious peaks generated by the network during the early training epochs.

Finally, a critical issue tied to the diffusion process must be addressed. For values of t close to T, the signal is almost entirely masked by noise and the estimate of $\hat{\mathbf{x}}_0$

is highly imprecise. To account for this uncertainty and avoid destructive gradients, the spectral loss is dynamically weighted by the cumulative parameter $\bar{\alpha}_t$, which acts as an indicator of the SNR. When the noise is very high, this coefficient drastically attenuates the impact of the spectral loss, leaving the optimisation entirely driven by the temporal MSE.

The overall formulation of the hybrid loss function, implemented in Code 5.1, is therefore:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}} \left[\text{MSE}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c)) + \lambda_{\text{freq}} \cdot \bar{\alpha}_t \cdot \text{MAE}(\log(|\mathcal{F}(\hat{\mathbf{x}}_0)| + \delta), \log(|\mathcal{F}(\mathbf{x}_0)| + \delta)) \right] \tag{5.18}$$

where $\delta = 10^{-7}$ denotes a numerical stability factor introduced to prevent mathematical singularities.

```python
# MSE loss
mse_loss = F.mse_loss(noise_pred, noise)

# Estimate the original signal from the predicted noise
sqrt_alpha_bar_safe = torch.clamp(sqrt_alpha_bar, min=1e-5)
x_0_pred = (x_t - sqrt_one_minus_alpha_bar * noise_pred) /
    sqrt_alpha_bar_safe

# Magnitude Spectrum through rFFT
fft_pred = torch.abs(torch.fft.rfft(x_0_pred, dim=-1))
with torch.no_grad():
    fft_real = torch.abs(torch.fft.rfft(x_0, dim=-1))
# Logarithmic scaling (adding 1e-7 to prevent log(0) instability)
log_fft_pred = torch.log(fft_pred + 1e-7)
log_fft_real = torch.log(fft_real + 1e-7)

# Absolute Error
sample_loss_freq = F.l1_loss(log_fft_pred, log_fft_real, reduction='
    none')
# Mean along the frequency and channel dimensions, final shape: [B]
sample_loss_freq = sample_loss_freq.mean(dim=(1, 2))

# Dynamic SNR weighting (alpha_bar has dimension [B])
weighted_sample_loss = sample_loss_freq * alpha_bar
loss_freq = torch.mean(weighted_sample_loss)

# Hybrid loss: MSE loss + frequency loss
loss_total = mse_loss + (lambda_freq * loss_freq)
```

Code 5.1: Hybrid training loss function: a weighted combination of standard DDPM temporal MSE and SNR-weighted frequency loss.

5.6.3 Signal Synthesis via Classifier-Free Guidance

During the inference phase, the synthesis of the new vibration signal $\mathbf{x}_0$ is achieved by applying the *denoising* process, starting from a pure noise tensor sampled from the standard Gaussian distribution $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To ensure that the generated signal strictly adheres to the requested operational parameters, the algorithm integrates *Classifier-Free Guidance* (CFG) [23]. When CFG is applied ($\omega > 1$), the network calculates two distinct noise predictions within each sampling step: a conditioned estimate guided by the conditioning embedding ($\boldsymbol{\epsilon}_{\text{cond}}$) and an unconditional estimate ($\boldsymbol{\epsilon}_{\text{uncond}}$). Specifically, the unconditional estimate is obtained by replacing the actual conditioning embedding with a specific embedding learned during the training phase to represent the absence of conditioning.

The final noise prediction ($\hat{\boldsymbol{\epsilon}}_\theta$) is then calculated through the following linear combination, weighted by the guidance scale parameter ω:

$$\hat{\boldsymbol{\epsilon}}_\theta = \boldsymbol{\epsilon}_{\text{uncond}} + \omega(\boldsymbol{\epsilon}_{\text{cond}} - \boldsymbol{\epsilon}_{\text{uncond}}) \tag{5.19}$$

In this formulation, the unconditional estimate provides a baseline that captures the general structure of the vibration signals independently of any specific operational condition. Meanwhile, the difference term estimates the directional gradient guiding the generation towards the target operating state. For $\omega > 1$, the equation performs an extrapolation beyond the conditioned estimate, amplifying the features specific to the target operating condition.

A key advantage of the CFG strategy is the ability to adjust the parameter ω at inference time, which enables a dynamic balance between stochastic exploration (diversity) and strict adherence to the target vibration signature (fidelity).

This property proves particularly valuable during the zero-shot generation evaluation, where the guidance scale is tuned to maximise conditioning fidelity for operating conditions not encountered during training, counteracting the model's tendency to generate samples that drift towards the nearest observed condition in the conditioning space.

The complete reverse process, including the application of the CFG, is detailed in the Code 5.2.

```
1 @torch.no_grad()
2 def sample(self, cond_vector, signal_length, omega = 1.0):
3 
4     self.eval()
5     device = next(self.denoiser.parameters()).device
6     cond_vector = cond_vector.to(device)
```

```
7
8       if cond_vector.dim() == 1:
9           cond_vector = cond_vector.unsqueeze(0)
10
11      B = cond_vector.shape[0]
12      x = torch.randn(B, 1, signal_length, device = device)
13
14      for t in reversed(range(self.num_timesteps)):
15          t_batch = torch.full((B,), t, device=device, dtype=torch.long)
16
17          if omega > 1.0:
18              # Conditioned noise estimation
19              noise_cond = self.denoiser(
20                  x, t_batch, cond_vector, force_uncond=False
21              )
22              # Unconditional noise estimation
23              noise_uncond = self.denoiser(
24                  x, t_batch, cond_vector, force_uncond=True
25              )
26              # Classifier-Free Guidance
27              noise_pred = noise_uncond + omega * (
28        noise_cond - noise_uncond)
29          else:
30              # Conditioned noise estimation
31              noise_pred = self.denoiser(
32                  x, t_batch, cond_vector, force_uncond=False
33              )
34
35          # Retrieve noise scheduling parameters for timestep t
36          alpha = self.alphas[t]
37          alpha_cumprod = self.alphas_cumprod[t]
38          beta = self.betas[t]
39
40          # Stochastic noise term for the reverse step
41          noise = torch.randn_like(x) if t > 0 else 0
42
43          # Reverse transition step: computes x_{t-1} from x_t
44          coef1 = 1 / torch.sqrt(alpha)
45          coef2 = (1 - alpha) / torch.sqrt(1 - alpha_cumprod)
46          x = coef1 * (x - coef2 * noise_pred) + torch.sqrt(beta) *
    noise
47
48      return x.cpu().squeeze().numpy()
```

Code 5.2: Reverse sampling process with CFG.

5.7 Training Strategy and Optimisation

To minimise the hybrid loss function, the network parameters of the U-Net were updated using the AdamW optimiser [36]. Compared to standard adaptive gradient algorithms such as Adam, this variant decouples the weight decay term from the gradient-based update, providing more robust regularisation [36]. This approach penalises excessive weight growth, thereby reducing the risk of overfitting and preventing the model from memorising specific noise realisations.

In addition to the choice of optimiser, two further algorithmic control strategies are employed to manage the complex training dynamics typical of diffusion models and to prevent divergence during gradient descent. First, the learning rate is not kept constant but is modulated using a *cosine annealing scheduler*. The learning rate starts from a predefined initial value and decreases progressively following a cosine profile. This strategy forces the model through an initial phase of exploration of the solution space, followed by an increasingly conservative refinement of the weights in the final epochs. Additionally, to counteract the problem of exploding gradients, which is inherent in deep architectures, the *gradient clipping* technique is applied. During each backward pass, the L_2 global norm of all parameter gradients is clipped to a predefined maximum threshold, facilitating stable convergence of the generative model.

Having defined the training strategy, the next step involved optimising the model's hyperparameters to refine the architecture and ensure maximum fidelity in the generation of vibration signals. To keep the computational load within acceptable limits, the search focused on the network's most influential architectural hyperparameters, excluding training hyperparameters, such as the initial learning rate and the weighting coefficient λ_{freq}.

Specifically, the decision not to let Optuna search for the optimal learning rate reflects an established best practice in the development of DDPM. Training diffusion models produces inherently noisy and variable gradients, as the error is calculated over timesteps featuring ever-changing amounts of stochastic noise. To manage this mathematical instability, it is essential to set a conservative, modest base learning rate a priori. In this work, the convergence dynamics are already well-managed by the combined action of the AdamW optimiser, which dynamically adapts the step size for each parameter, and the cosine annealing scheduler, which enforces a global reduction of the learning rate. Exploring random orders of magnitude for the learning rate through Optuna would have increased the risk of divergence without providing meaningful improvements to training stability.

Hyperparameter	Search Space	Optimal Value
`base_ch`	{16, 32, 64}	64
`num_layers`	[3, 5]	4
`attn_window_size`	{32, 64, 128}	128

Table 5.3: Hyperparameters evaluated with Optuna and their optimal values.

Similarly, the weighting coefficient λ_{freq} was also excluded from the search space and calibrated empirically. It was adjusted to ensure that the frequency term accounted for approximately 20% of the total loss, which proved to be the most effective balance. This balance is crucial because the two components of the loss function, the mean squared error on the noise and the mean absolute error on the logarithmic spectrum, differ significantly in magnitude. An excessive value of λ_{freq} would have caused the spectral component to completely overshadow the temporal one. By maintaining a controlled weight of 20%, however, the frequency loss acts as a useful auxiliary regulariser, favouring the faithful reconstruction of harmonic peaks, without ever compromising the overall stability of convergence.

The search space explored by Optuna focused on the three most impactful architectural hyperparameters. The first is the size of the base channels (`base_ch`), tested at values of 16, 32 and 64 to determine the initial number of convolutional filters. The second is the overall network depth (`num_layers`), varied within a range of 3 to 5. Finally, significant focus was placed on the size of the attention window (`attn_window_size`), evaluated with sizes of 32, 64 and 128 elements.

The `objective()` function minimised by Optuna is the hybrid loss evaluated on the validation set. The best hyperparameter configuration identified on the CWRU dataset (Table 5.3) was considered the final architectural setup and was kept unchanged for the subsequent training phase on the EECSL dataset. The decision not to perform a new optimisation of the architectural hyperparameters for the second dataset reflects the desire to evaluate the robustness and intrinsic generalisation capabilities of the model. Furthermore, it represents a pragmatic choice to limit the severe computational burden associated with hyperparameter optimisation in diffusion models.

Unlike the CWRU domain, signals in the EECSL dataset come from motors driven by inverters. PWM introduces significant high-frequency harmonic content that alters the signal morphology and the background noise spectrum. Maintaining the architectural configuration unchanged, same number of layers, base channels,

Hyperparameter	Value	Description
n_channels	1	Input channels (single-channel signal)
base_ch	64	Base filters of the initial convolutional layer
num_layers	4	Resolution levels (downsampling / upsampling stages)
time_emb_dim	128	Temporal embedding dimension
cond_emb_dim	128	Conditional embedding dimension
attn_window_size	128	Temporal window size (local self-attention)
heads (Local)	8	Number of attention heads (local self-attention)
heads (Global)	4	Number of attention heads (global self-attention)
kernel_size	5	Spatial dimension of filters within residual blocks
dropout	0.1	Dropout probability for network regularisation
norm_groups	8	Number of groups for GN

Table 5.4: Architectural hyperparameter configuration of the U-Net employed as the denoiser.

and attention window size, enables an assessment of the framework's versatility and robustness across different signal morphologies, independently of the machine's power supply type.

Similarly, the hyperparameters related to the training dynamics, such as the learning rate and the optimiser configurations, were also kept constant. The only exception concerned the hybrid loss function, and in particular the weighting coefficient λ_{freq}. Due to the different spectral content and harmonic amplitudes introduced by the inverter, this parameter was empirically recalibrated for the EECSL dataset. This adjustment was necessary to ensure that the contribution of the frequency component to the total loss remained constant (around 20%). This recalibration ensured consistent convergence behaviour across both datasets, maintaining the same balance between temporal and spectral objectives.

To ensure full reproducibility of the experiments, the final configurations adopted are presented in two main categories: the architectural hyperparameters of the denoiser (Table 5.4), and those governing the diffusion process and the training strategy (Table 5.5).

5.8 Evaluation Framework

Following the hyperparameter optimisation, a comprehensive evaluation framework was established to measure the quality of the generated signals using the test set. It

Hyperparameter	Value	Description
`num_timesteps` (T)	1000	Total steps for the forward corruption and reverse generation processes
Beta Schedule	Cosine	Variance schedule for the forward noise process
`batch_size`	32	Number of samples processed per iteration
`num_epochs`	300	Total training epochs
`learning_rate` (η)	1×10^{-4}	Initial base learning rate
Optimiser	AdamW	Adaptive gradient algorithm (weight decay $=1 \times 10^{-4}$)
LR Scheduler	Cosine	Learning rate decay profile ($T_{max} = 300$, $\eta_{min} = 1 \times 10^{-6}$)
`cond_drop_prob`	0.15	Conditional dropout probability during training (CFG)
λ_{freq} (CWRU)	0.005	Spectral component weight for CWRU dataset
λ_{freq} (EECSL)	0.01	Spectral component weight for EECSL dataset

Table 5.5: Hyperparameter configuration for the diffusion process and the training strategy.

is essential to demonstrate that the generated signals preserve the correct temporal morphology and the spectral content typical of mechanical faults. This assessment builds on established evaluation methodologies for vibration signal synthesis, measuring the statistical fidelity of the generated samples with respect to unseen real-world distributions, as well as their practical utility in downstream fault diagnostic tasks. Unlike image generation, where visual inspection and standardised metrics like the Fréchet Inception Distance (FID) provide straightforward evaluation criteria, synthesising physical signals presents severe testing challenges.

Given this complexity, the performance analysis does not rely on a single global metric, but adopts a progressive and rigorous assessment strategy. First, the signal fidelity is examined through a qualitative visual inspection paired with quantitative statistical indicators in both the time and frequency domains. Subsequently, t-SNE visualises the overlap of real and synthetic data in the space of diagnostic statistical indicators and in that of deep representations extracted from the CNN. The diagnostic utility of the generated samples is then empirically evaluated using a downstream classifier. To rule out the risk of dataset memorisation, the nearest-neighbour distances between the generated samples and the original training data were computed. Finally, the evaluation framework concludes with a zero-shot generation test, designed to demonstrate that the model has learnt the underlying

physical dynamics and can synthesise coherent signals for operating conditions never seen during training.

5.8.1 Signal Fidelity Evaluation

The preliminary stage of the evaluation strategy focuses on the fundamental morphological and harmonic fidelity of the generated signals. Initially, this involves a qualitative visual inspection in both the time and frequency domains. By directly comparing the temporal waveforms and the magnitude spectra of real and synthetic signals, it is possible to visually verify that the generative model successfully replicates macro-level signal characteristics. These include the signature of the mechanical faults, the stochastic distribution of broadband noise, and the excitation of high-frequency structural resonance bands typical of the system's dynamic response.

Nevertheless, visual inspection alone is inherently insufficient for a rigorous evaluation framework. The highly stochastic nature of mechanical vibrations, driven by non-linear friction, transient load fluctuations, and inverter-induced noise, masks subtle amplitude variations that visual inspection cannot reliably capture. Furthermore, qualitative visual analysis is incapable of objectively quantifying cumulative energy discrepancies across the frequency spectrum. Consequently, these preliminary observations serve only as a baseline and must be strictly confirmed by precise numerical evaluations.

To quantitatively assess the generative model's ability to preserve key physical and statistical properties, a systematic feature extraction was performed. In the time domain, this analysis employed the classic indicators of mechanical diagnostics [8, 15]: **RMS**, **kurtosis**, **crest factor**, **skewness** and **peak**. These features quantify the overall energy content (RMS), the presence of impulsive mechanical impacts (kurtosis, crest factor), the statistical asymmetry of the amplitude distribution (skewness), and the maximum absolute deviation of the signal (peak). Furthermore, to evaluate the signals in the frequency domain, three spectral shape features were extracted from the power spectrum to capture global profile, dispersion, and complexity of the energy distribution: Root Mean Square Frequency (RMSF), Root Variance Frequency (RVF), and normalised spectral entropy [15].

Specifically, assuming P_k is the spectral energy at the k-th frequency bin f_k (for $k = 1, \ldots, K$), these features are calculated as follows:

RMSF: Defined as the energy-weighted quadratic mean frequency of the spectrum. By weighting higher-frequency components more heavily than the Frequency Centroid

(FC), it effectively indicates position shifts of the main frequency components.

$$\text{RMSF} = \sqrt{\frac{\sum_{k=1}^{K} f_k^2 P_k}{\sum_{k=1}^{K} P_k}}$$

RVF: Measures the spectral dispersion around the spectral centroid (or FC), computed as:

$$\text{RVF} = \sqrt{\frac{\sum_{k=1}^{K} (f_k - \text{FC})^2 P_k}{\sum_{k=1}^{K} P_k}}$$

where the FC is defined as:

$$\text{FC} = \frac{\sum_{k=1}^{K} f_k P_k}{\sum_{k=1}^{K} P_k}$$

Spectral entropy: Quantifies the randomness and complexity of the energy distribution in the spectrum. It is computed as Shannon entropy divided by the maximum theoretical entropy:

$$H_{\text{norm}} = -\frac{\sum_{k=1}^{K} p_k \log_2(p_k)}{\log_2(K)}$$

where p_k represents the power spectrum normalised as a probability mass function:

$$p_k = \frac{P_k}{\sum_{j=1}^{K} P_j}$$

Finally, to comprehensively visualise the distributions of both time and frequency features, strip plots overlaid on box plots were generated for each operating condition, comparing the synthetic samples and the real data from the test set.

In addition, the analysis employed two complementary metrics to quantify the exact bin-by-bin alignment of the frequency components: the **Log-Spectral Distance (LSD)** and the **Spectral Cosine Similarity (SCS)**. To extract a robust characteristic signature for each fault class, the metrics were computed using the mean magnitude spectra (for the SCS) and the mean power spectra (for the LSD), averaged over a statistically significant set of real and synthetic samples.

The LSD assesses whether the generative model has accurately reproduced the correct amount of energy across the power spectrum. Let $P_r(k)$ and $P_s(k)$ be the mean power spectra of the real and synthetic signals, the LSD is defined as the root mean squared error calculated on a logarithmic scale across the K valid frequency bins [17]:

$$LSD = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left[10 \log_{10}(P_r(k)) - 10 \log_{10}(P_s(k))\right]^2} \quad \text{[dB]}$$

If two signals possess an identical power spectral density distribution, this metric yields a value of 0 dB. The use of a logarithmic scale is highly advantageous in

vibration analysis; by compressing the wide dynamic range of the signal, it prevents the distance metric from being disproportionately dominated by the high-amplitude components. This ensures that lower-energy features critical for early fault detection, such as sideband modulations around structural resonances and localised broadband energy shifts, are properly weighted in the evaluation.

While the LSD focuses on relative energy variations across the spectrum, the SCS evaluates the shape of the spectrum. It is a scale-invariant metric that assesses the similarity of harmonic peak positions independently of their absolute amplitude. If the generative model synthesises the fault peaks at the correct frequencies but with a slightly divergent energy level, the LSD will penalise the generative model, whereas the SCS will yield a high score, confirming that the fundamental harmonic structure of the fault has been successfully learned. Specifically, the SCS metric considers the two mean magnitude spectra as K-dimensional vectors and calculates the cosine of the angle between them. Its value ranges from 0, in the case of completely orthogonal vectors that do not share any common frequency, to 1 when the vectors are perfectly proportional, meaning the synthetic spectrum perfectly replicates the shape and relative proportions of the real harmonic peaks. Let $S_r(k)$ and $S_s(k)$ be the mean magnitude spectra over the K valid frequency bins, the SCS is defined as their normalised inner product:

$$SCS = \frac{\sum_{k=1}^{K} S_r(k) \cdot S_s(k)}{\sqrt{\sum_{k=1}^{K} S_r(k)^2} \cdot \sqrt{\sum_{k=1}^{K} S_s(k)^2}}$$

5.8.2 Diagnostic Utility Assessment via Downstream Classification

In the diagnostic field, the ultimate proof of the quality of the synthetic vibrational signals lies in their practical utility. If a classifier, trained exclusively on synthetic samples, successfully recognises physical faults on real machinery, it provides empirical evidence that the generative model has learned and replicated the correct physical dynamics of the system. To translate this principle into an objective metric, it was necessary to design the downstream classifier that would serve as a proper performance measurement tool.

Downstream Classifier Architecture

To carry out the downstream classification task, a one-dimensional CNN based on the WDCNN architecture [70] was developed. Targeted architectural enhancements were introduced to improve the model's accuracy, using only real samples as reference. The final architecture is shown in Figure 5.8.

The network features an initial convolutional layer with a notably large receptive field, employing a kernel size of 64 samples. Applying such a wide filter directly to the raw vibration time windows inherently acts as a low-pass filter, effectively averaging and attenuating the high-frequency stochastic noise. In contrast to the original baseline model, BN was deliberately omitted in this first layer, and the convolutional stride was significantly reduced ($s = 4$ instead of 16). Together, these two modifications yielded an improvement in the classifier's overall accuracy.

This initial stage is followed by four standard convolutional layers, featuring small kernels (size of 3), designed for the hierarchical extraction of complex, high-level features. Unlike the first layer, these subsequent stages are regularised through BN to stabilise the training process and accelerate convergence.

A further structural design choice is the introduction of the adaptive max pooling operator to downsample the resulting tensor into a fixed temporal dimension of 4 timesteps. This operation makes the network flexible and invariant to the length L of the input time window. By forcing this fixed output size, the subsequent flattening phase consistently yields a feature vector of dimension 512 (128 channels $\times$ 4 timesteps), thereby guaranteeing structural compatibility with the dense classification layers, regardless of the input tensor length.

Finally, classification is handled by a fully connected MLP, regularised with a dropout rate of approximately 50% to mitigate the risk of overfitting. The model is trained by minimising the Cross-Entropy loss function using the Adam optimiser.

Diagnostic Utility Assessment

To quantify both the diagnostic utility and the distributional fidelity of the synthetic signals, the developed CNN was utilised within four distinct evaluation scenarios. To ensure maximum methodological rigor, this evaluation strategy is built upon two fundamental premises. First, the synthetic datasets (training, validation, and test) were generated to match the exact size of their real counterparts. An equal number of samples was generated for each class to perfectly mirror the balanced structure of the real datasets, thereby preventing class imbalance bias and ensuring that overall accuracy is a meaningful performance metric. Second, it is important to clarify that the real test set used for the final evaluation phase was strictly isolated from any prior processing. By excluding these real samples from the diffusion U-Net's training phase and hyperparameter optimisation, the risk of data leakage was mitigated.

With these methodological foundations established, the four evaluation scenarios are defined as follows.

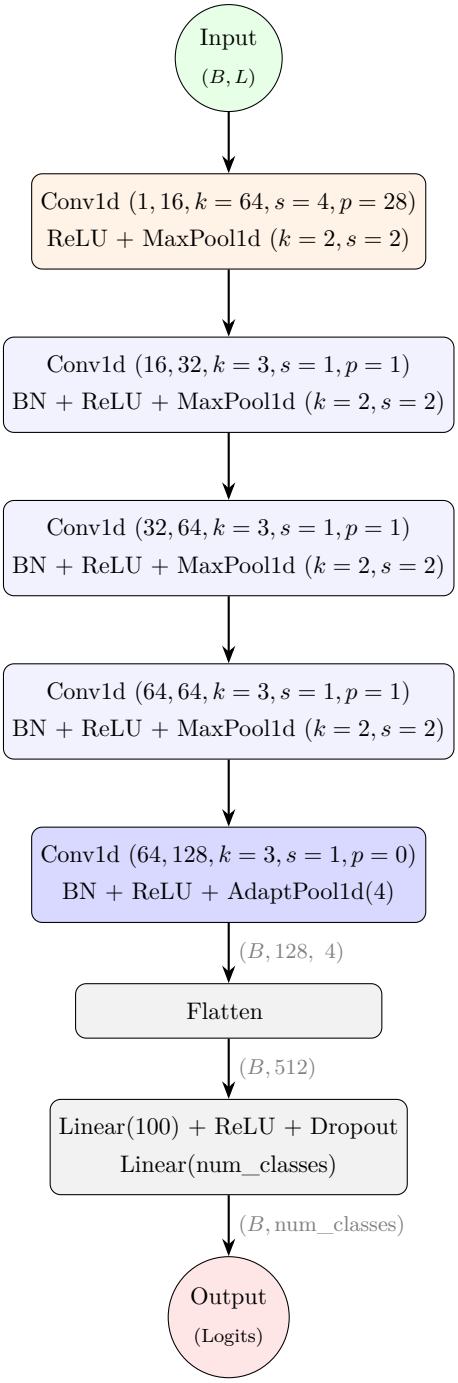


Figure 5.8: Architecture of the CNN downstream classifier. The numerical values on the arrows represent the dimensions of the output tensors from each layer.

The baseline configuration, named *Train on Real – Test on Real* (TRTR), involves training and testing the classifier exclusively on the original dataset, thereby establishing an upper-bound performance reference for the architecture's diagnostic capability.

The most critical scenario for evaluating the generative model is *Train on Synthetic – Test on Real* (TSTR) [14]. By training the network only on synthetic samples and evaluating it on the real test set, this setup directly measures the practical utility of the generated signals. High accuracy confirms that the synthetic samples contain sufficient discriminative information to train a classifier that generalises to real data, suggesting that the diffusion model has captured the fault-relevant structure of the various operating conditions. Furthermore, if TSTR performance approaches the TRTR baseline, it indicates that the synthetic distribution is rich and diverse enough to serve as a valid substitute for real data in classifier training.

To complement this analysis, the *Train on Real – Test on Synthetic* (TRTS) scenario evaluates the discriminative fidelity of the generated signals. By testing a model trained exclusively on real samples against synthetic samples, this scenario assesses whether the generation process has reproduced the class-discriminative structure of the original data without introducing unphysical artefacts. A high accuracy score confirms that the synthetic samples lie within the decision boundaries learnt by the classifier on real data.

Finally, the *Train on Synthetic – Test on Synthetic* (TSTS) configuration assesses the internal consistency and class separability of the synthetic dataset. Strong performance in this scenario indicates that the generative model produces well-structured and discriminative representations across all fault classes and operating conditions.

5.8.3 t-SNE Visualisation of Distribution Alignment

To visually evaluate the similarity and alignment between the real and synthesised data, the t-Distributed Stochastic Neighbour Embedding (t-SNE) dimensionality reduction technique was employed [57]. The primary objective of this approach is to map the complex neighbourhood relationships present in high-dimensional spaces into a two-dimensional plane. The t-SNE algorithm excels at positioning highly similar points close together on the map while clearly separating dissimilar points. In this new space, the axes do not carry physical meaning and inter-point distances in the projected space should not be interpreted as proportional to the original high-dimensional dissimilarities.

To examine the quality of the generation, t-SNE was employed across two distinct information domains: the space of diagnostic features and the space of deep representations extracted from the developed CNN.

The first phase of the analysis focused on the diagnostic features. For each vibration segment, whether real or synthetic, an eight-dimensional vector was utilised, comprising the classic indicators of mechanical diagnostics: RMS, kurtosis, crest factor, skewness, peak, RMSF, RVF, and spectral entropy. To balance the contribution of metrics characterised by heterogeneous orders of magnitude, the vectors were standardised using the Z-score method. The standardisation parameters (mean and standard deviation) were derived exclusively from the real data and subsequently applied to both real and synthetic samples. A successful generation is reflected in a substantial overlap between synthetic points and the clusters formed by real signals of the same class. This suggests that the model has not merely replicated the marginal distributions of individual features, but successfully captured the joint distribution of all features simultaneously.

Subsequently, the t-SNE algorithm was utilised to visualise the space of the non-linear abstractions learnt by the developed CNN. In this phase, the CNN, trained on empirical data, was employed as a feature extractor for both synthetic and real test samples. Instead of completing the forward pass, execution was truncated at the penultimate dense layer to extract a 100-dimensional feature vector. This operation isolates the deep features, which represent the high-level semantic abstractions that the network has developed to uniquely describe the classes before they are converted into the class logits. Consequently, if the resulting t-SNE plot shows that the distribution of the generated samples perfectly overlaps and clusters within the same areas as their corresponding real classes, it indicates that the diffusion model generates signals whose deep representations are indistinguishable from those of real samples belonging to the same fault class. This structural overlap suggests that the generative model has learnt to reproduce the same abstract, discriminative signatures that define the physical fault conditions from the perspective of a diagnostic network.

5.8.4 Memorisation Risk Assessment

In the evaluation of deep neural architectures, achieving high-fidelity generation raises the critical issue of memorisation. A model that fails to learn the true underlying distribution of the physical phenomenon, merely memorising the training samples and reproducing them exactly or with negligible variations, lacks generative value. To verify that each generated sample is a novel instance rather than a copy of a

training sample, a 1-Nearest Neighbour (1-NN) distance analysis is carried out.

To quantify the similarity, the analysis relies on the Euclidean distance. A standard point-to-point evaluation in the time domain would be highly sensitive to temporal misalignments, artificially increasing the distance between vibration time windows that are physically identical but temporally shifted. Consequently, the evaluation is transferred to the frequency domain by exploiting the time-shifting property of the Fourier transform [39]. Since a time delay translates exclusively into a phase shift in the frequency domain, computing the Euclidean distance on the normalised magnitude spectra renders the metric invariant to temporal translations.

The memorisation assessment is performed by computing the cross-distances between the synthetic spectra and the real training spectra. For each generated sample, the shortest distance to any real training sample is recorded as a direct indicator of potential data replication. To contextualise these results, the 1-NN distance distribution of the synthetic data is visualised using KDE plots, superimposed with two reference distributions: the intra-training distribution (Train-to-Train) and the test reference (Test-to-Train).

The intra-training distribution is calculated by finding the nearest neighbour for each train window within the remainder of the training set. In the specific case of the CWRU dataset, where a reduced sliding-window stride is applied, adjacent temporal segments share a high degree of overlap. This overlap causes the baseline to exhibit a sharp peak very close to zero. A successful generative model is expected to safely avoid this near-zero zone.

The second baseline is the test reference (Test-to-Train), which shows the distance distribution between unseen real windows and the training data. Since the test samples are extracted using non-overlapping windows, their distances are not artificially shifted towards zero.

The visual interpretation of the final KDE plot is straightforward: the generation is considered free from memorisation if the synthetic distance distribution does not exhibit a peak near zero. Moreover, if the synthetic distance distribution aligns with that of the test set, the generated samples possess the same natural variability and novelty as unseen real samples.

5.8.5 Zero-Shot Generation

To evaluate the diffusion model's ability to learn the underlying physical dynamics of the faults, a zero-shot generation experiment is conducted. The objective of this test is to demonstrate that the model is capable of reproducing complex mechanical

behaviours even for conditions it has never encountered during training.

To this end, the model is trained by excluding specific operating conditions from the training set. For the CWRU dataset, all samples corresponding to a 0.014-inch severity fault and a motor load of 2 HP are omitted. This exclusion is applied consistently across all three bearing fault topologies: inner raceway, ball, and outer raceway. For the EECSL dataset, the vibration signals corresponding to the induction motor driven by the VFD at 40 Hz under a 75% load are withheld across all baseline and fault conditions (healthy, eccentricity, and bearing fault).

By injecting the target conditioning embeddings corresponding to the withheld operating conditions into the trained U-Net, synthetic signals are generated. If the downstream classifier trained on these synthetic data achieves high accuracy when evaluated on real test samples of the excluded classes, it provides empirical evidence that the diffusion model has synthesised vibration signals with sufficient discriminative structure even for operating conditions never seen during training.

To successfully generate these unseen operating conditions, the proposed diffusion model must infer the underlying physical dynamics from other operating conditions. To guide the generation towards the unseen target states, the CFG mechanism is specifically employed: by tuning the guidance scale hyperparameter w, the trade-off between sample diversity and conditioning fidelity is explicitly controlled. A higher guidance scale increases adherence to the target conditioning embedding at the cost of reduced sample diversity. Excessively high values may push the generation beyond the learned distribution, potentially compromising physical realism. To empirically determine the optimal guidance scale, a parametric sweep is conducted across a discrete set of guidance scale values ($w \in \{1.0, 1.3, 2.0, 3.0, 4.0\}$). The quality of the resulting synthetic data is assessed through the downstream classifier's accuracy, indicating that $w = 2.0$ achieves the optimal compromise between conditioning fidelity and generative quality within this experimental setting.

This zero-shot generation experiment concludes the implementation chapter of this thesis, providing evidence that the proposed architecture successfully emulates the complex physical reality of rotating machinery. Ultimately, this generative capability enables diagnostic classifiers to be trained on synthetic data representing operating conditions for which no real measurements were collected, thereby reducing the need for exhaustive experimental campaigns covering every possible combination of fault type, severity, and operating condition.

Chapter 6

Results

"Somewhere, something incredible is waiting to be known."

Carl Sagan

This chapter presents the results of the proposed conditional diffusion model, structured in strict accordance with the evaluation framework established in Chapter 5. Through this rigorous assessment, the analysis provides empirical evidence that the framework successfully generates physically consistent and diagnostically reliable surrogate data. Ultimately, the model demonstrates zero-shot generation capabilities by successfully synthesising reliable signals for operating conditions that were explicitly excluded from the training set.

6.1 Signal Fidelity Analysis

Before proceeding with the analysis of quantitative metrics, it is essential to visually verify that the diffusion model generates signals characterised by a coherent temporal structure and a faithful spectral energy distribution. This section will discuss highly representative case studies, chosen to encompass the distinct mechanical and electrical challenges posed by the datasets.

For the CWRU dataset, a qualitative comparison between the real signals from the test set (green) and those synthesised by the diffusion model (red) is conducted under an incipient fault condition, specifically a fault severity of 0.007 inches at a motor load of 2 HP. This specific configuration was selected to demonstrate the model's ability to capture the physical dynamics of early-stage degradation across all components. The analysis includes all dataset conditions: baseline, inner raceway

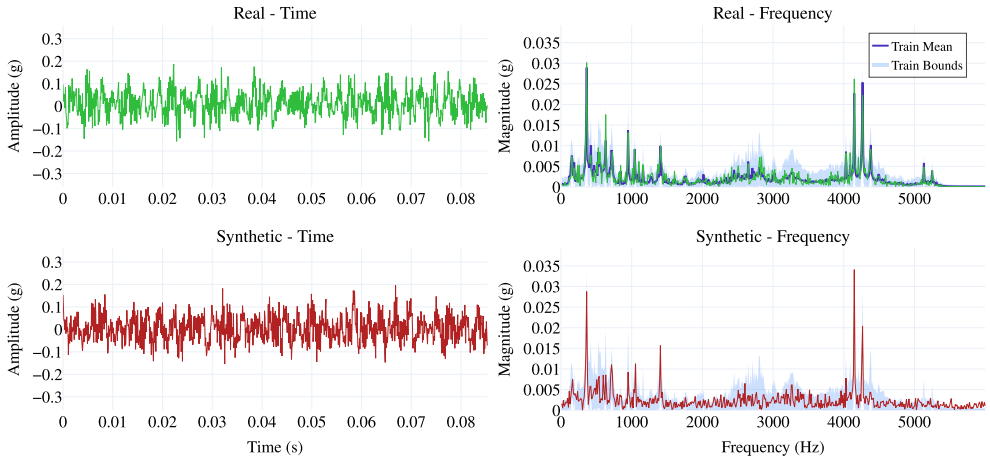
fault, outer raceway fault, and ball fault (Figure 6.1).

Before detailing the individual fault scenarios, it is necessary to clarify the expected characteristics of the magnitude spectrum of the vibration signal acquired by an accelerometer mounted on the motor housing. In a vibration amplitude spectrum, the fundamental kinematic fault frequencies are generally not distinguishable as discrete spectral peaks. Each defect-induced contact produces a short-duration mechanical impulse which, by the properties of the Fourier transform, distributes its energy across a wide frequency band rather than concentrating it at a single frequency. As a result, the spectral amplitude at any individual kinematic frequency remains low and is typically masked by the broadband noise floor of normal machine operation. The mechanical structure through which the impulse must propagate to reach the sensor behaves as a frequency-selective transmission path, attenuating most frequencies while amplifying those coinciding with its structural resonances. Consequently, the dominant features observed in the magnitude spectrum are not the kinematic fault frequencies themselves, but rather the structural resonance frequencies, periodically excited by the repetitive fault-induced impulses.

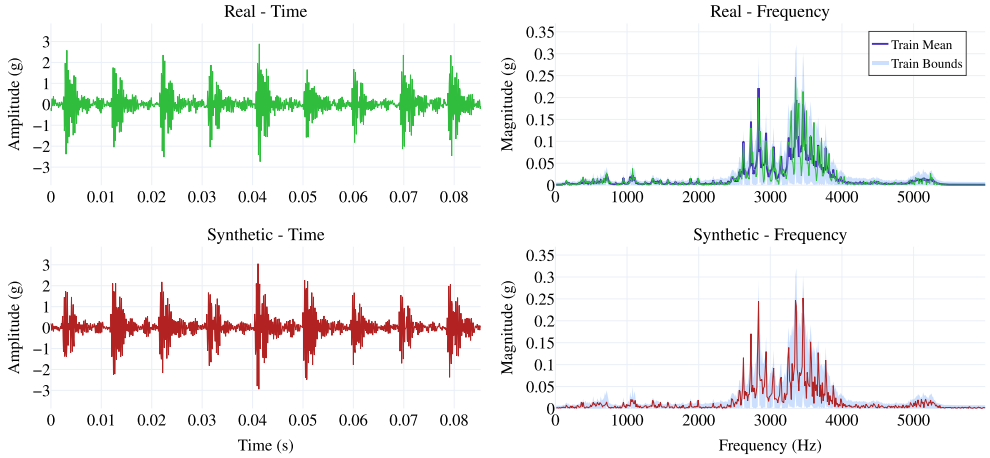
By analysing the operational conditions across all four scenarios illustrated in Figure 6.1, it can be observed that the spectral amplitude distribution of the synthetic signals remains strictly bounded within the statistical tolerance bands derived from the training set (grey area). This agreement confirms that the generative model has successfully captured the characteristic magnitude spectra of each specific fault condition.

The healthy baseline (Figure 6.1a) represents the first stage of the evaluation, verifying that the model does not introduce spurious transients in the time domain or spectral artefacts in the frequency domain in the absence of a defect. In the time domain, the synthetic signal is free from any impulsive content, while in the frequency domain, the magnitude spectrum exhibits extremely low amplitude levels, with the maximum peaks confined below approximately 0.03 g.

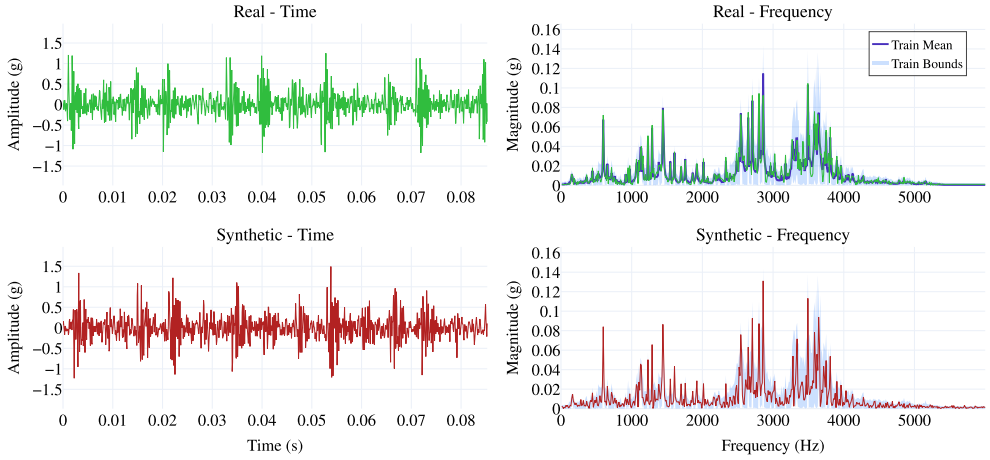
The outer race fault (Figure 6.1b) is stationary and maintains a fixed position relative to the maximum load zone. Consequently, it generates equally spaced impulses with a nearly constant amplitude. The synthetic signal accurately reproduces this characteristic repetition interval, which corresponds to the inverse of the BPFO, measuring 0.0095 s in the analysed case (load 2 HP). In the frequency domain, due to the fixed and shorter transmission path compared to the other defects, the impact energy directly excites the structural resonance frequencies, yielding the highest spectral amplitudes that reach up to 0.3 g.



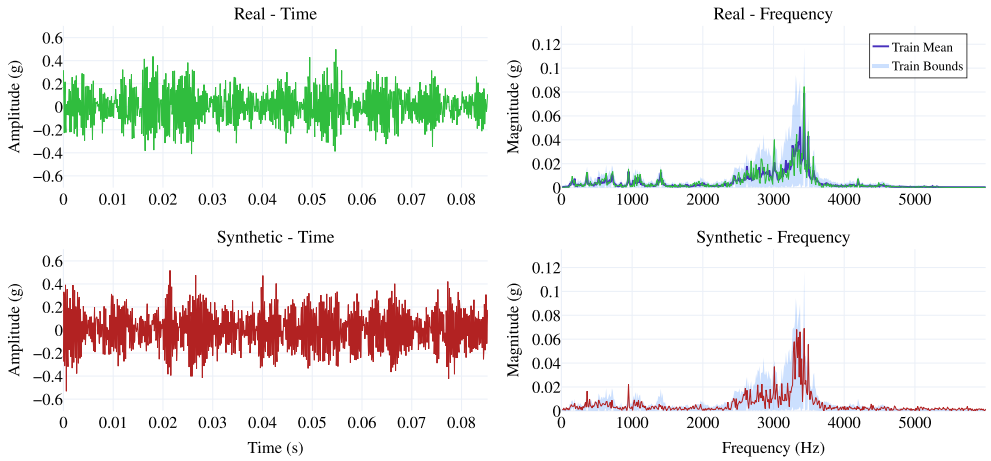
(a) Healthy Baseline (Load 2 HP)



(b) Outer Race Fault (0.007", Load 2 HP)



(c) Inner Race Fault (0.007", Load 2 HP)



(d) Ball Fault (0.007", Load 2 HP)

Figure 6.1: Comparison of real and synthetic vibration signals for healthy and faulty conditions (CWRU 12 kHz, Load: 2 HP, Diameter: 0.007"). The y-axis represents the acceleration in g ($1\text{g} = 9.8\ \text{m/s}^2$).

The inner race fault (Figure 6.1c) involves a more complex transmission path, as the defect-induced impacts must propagate through the rolling elements and the lubricant film before reaching the housing. Furthermore, because the defect rotates, its position relative to the load zone varies continuously, causing the impact amplitudes to be modulated at the shaft rotation frequency. In the time domain, the synthetic signal correctly shows periodic impulses spaced approximately $1/\text{BPFI} = 0.0063$ s, exhibiting varying amplitudes where the maximum values correspond to the load zone and the minimum values to the unloaded region. As expected, the spectral amplitude is more contained, peaking at 0.14 g, because part of the impact energy is attenuated along the longer transmission path. Notably, unlike the other fault types, the inner race defect also excites a distinct structural resonance around 1500 Hz, likely associated with the shaft assembly, which is correctly reproduced in the synthetic signals.

Finally, the complex dynamics of the ball fault (Figure 6.1d) are also correctly reproduced by the generative model. Since the defect on the rolling element alternately strikes the inner and outer raceways, the resulting impact pattern is significantly less regular and periodic compared to raceway faults. In the time domain, the vibration amplitude is bounded within ± 0.6g, showing lower values than the raceway defects because the impact energy undergoes greater dissipation before reaching the sensor. Accordingly, the magnitude spectrum of the synthetic signal reflects this behaviour,

with the amplitude distribution remaining more contained and individual peaks restricted below 0.12 g.

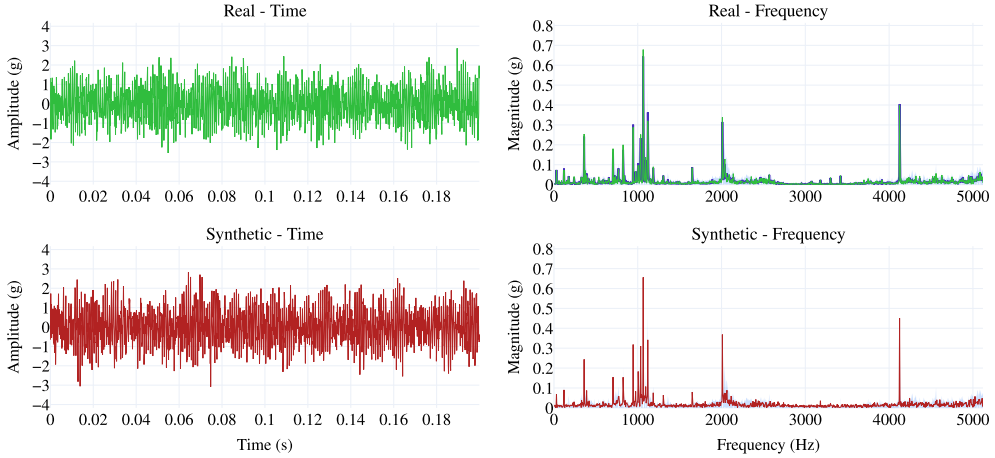
Moving to the EECSL dataset, the qualitative analysis encompasses not only localised bearing defects but also distributed anomalies such as motor eccentricity. The time-domain waveforms and magnitude spectra are first presented under DOL operation (Figure 6.2) and subsequently compared with their VFD counterparts (Figure 6.3), both under full load conditions.

Starting with the healthy baseline condition, the model correctly captures the differences between DOL and VFD operating regimes. Specifically, under VFD it accurately reproduces the higher overall vibration amplitude levels, the structural resonances excited by the high-frequency PWM switching harmonics, and the electromagnetic noise floor characteristic of inverter-driven operation

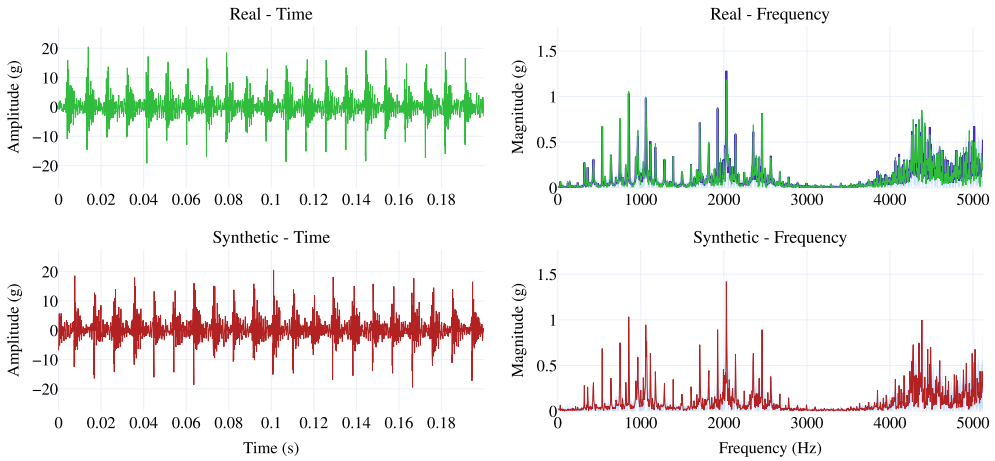
Regarding the outer race fault, the synthetic signals under both DOL and VFD conditions correctly reproduce the characteristic fault impulses at the BPFO of 109 Hz, corresponding to an impact period of approximately 0.009 s. Although the mechanical impact energy of the defect remains consistent across both operating modes, the overall vibration amplitude in the VFD regime is higher. The generative model successfully captures this phenomenon, which is driven by the superposition of the mechanical impacts with the additional vibration components introduced by the VFD, specifically radial electromagnetic forces and torque ripple.

Finally, the generative model demonstrates its ability to synthesise vibration signatures for eccentricity, an electromechanical fault whose physical nature is completely different from impulsive bearing defects. In the presence of dynamic eccentricity, the synchronous rotation of the minimum air-gap position with the shaft induces a periodic amplitude modulation of the vibration signal at the rotational frequency, a phenomenon successfully replicated by the model in the time domain. Furthermore, the model accurately reproduces the substantial increase in spectral amplitudes in the 4000–5000 Hz band, under VFD operation compared to DOL driving, which is caused by the interaction between the non-uniform air gap of the eccentric rotor and the high-frequency PWM switching harmonics.

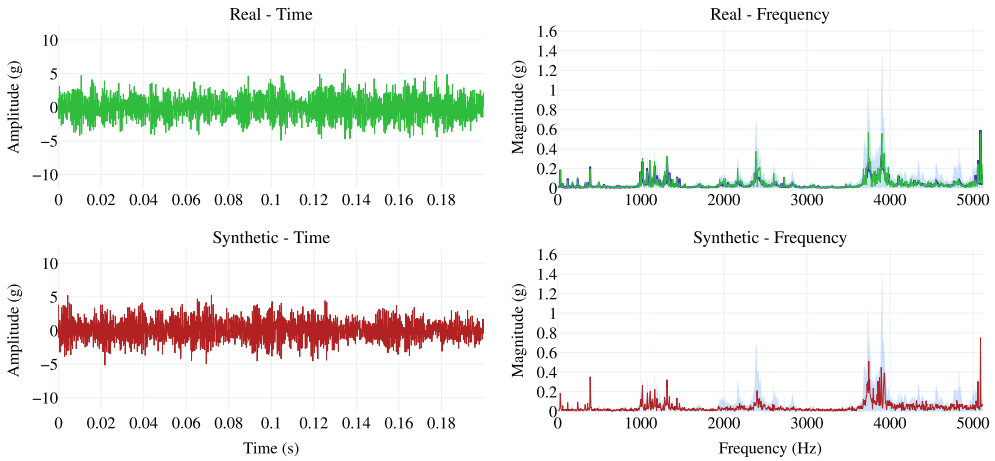
This comparative analysis confirms that the generative model accurately synthesises vibration signals under both DOL and VFD driving conditions, thereby demonstrating robustness across distinct fault types as well as adaptability to fundamentally different motor driving configurations.



(a) Healthy Baseline

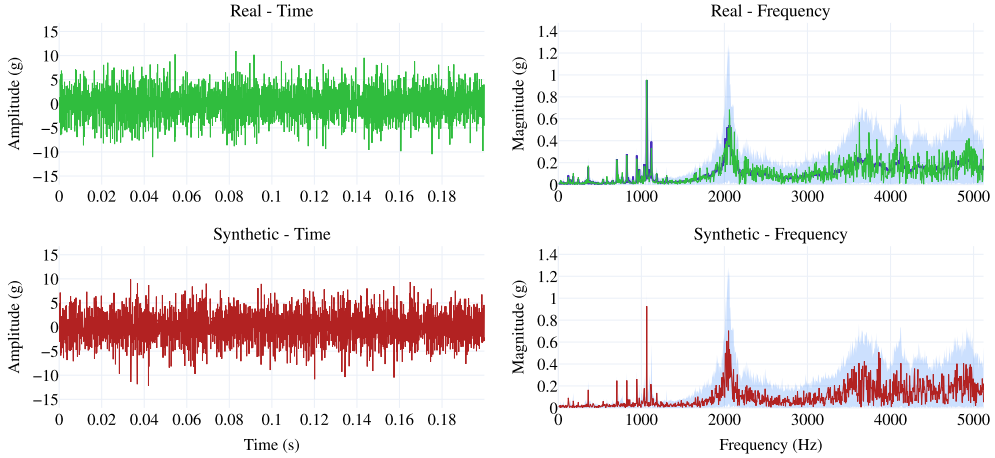


(b) Outer Race Fault

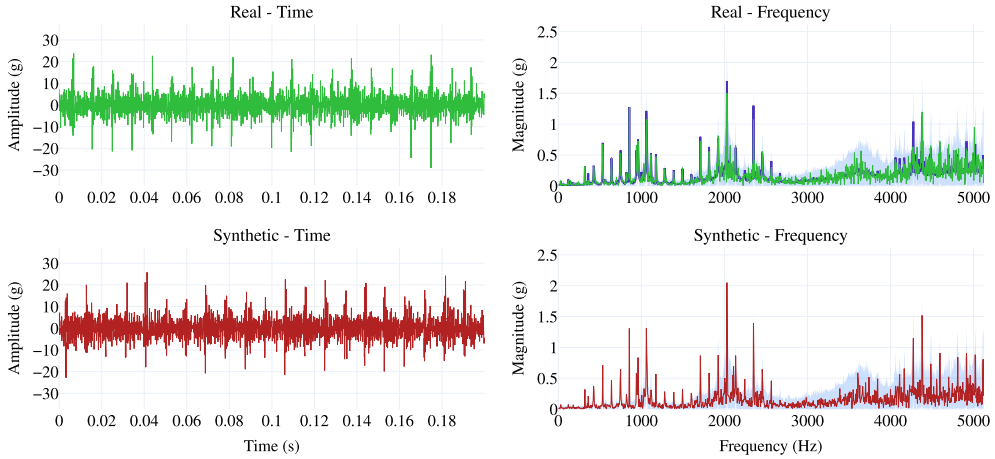


(c) Eccentricity Fault

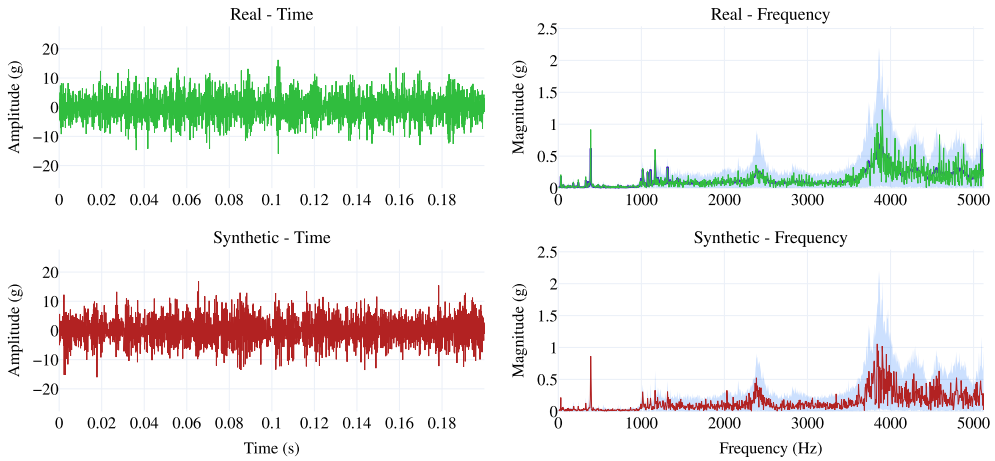
Figure 6.2: Comparison of real and synthetic vibration signals under DOL operation (EECSL, supply frequency: 60 Hz, Load: 100%).



(a) Healthy Baseline



(b) Outer Race Fault



(c) Eccentricity Fault

Figure 6.3: Comparison of real and synthetic vibration signals under VFD operation (EECSL, supply frequency: 60 Hz, Load: 100%).

6.2 Statistical and Spectral Feature Analysis

For synthetic data to be viable for data augmentation, its diagnostic features, in both the time and frequency domains, must closely match those of the real signals under analogous operating conditions.

To ensure robust statistical evaluation, the distributions of the features extracted from the synthetic data are compared against those from the real test set (23 samples for CWRU, 75 for EECSL), as well as the entire available real dataset (117 samples for CWRU, 250 for EECSL), for each operating condition.

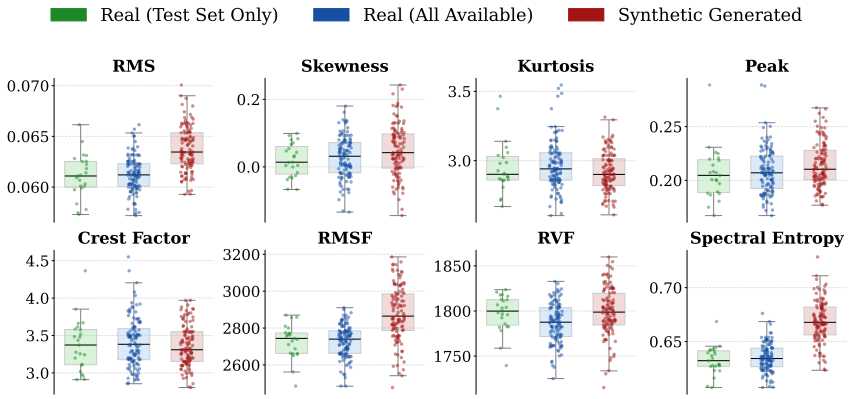
For the complete real dataset, no window overlap is applied. Furthermore, to guarantee a fair and robust evaluation, the number of generated synthetic samples is set equal to the total number of real samples.

The feature distributions are visualised using strip plots overlaid on box plots. This combined visualisation enables an assessment that goes beyond simple median alignment, allowing for the verification of statistical dispersion, interquartile ranges, and distribution tails. A close agreement between the real and synthetic distributions confirms that the generative model has not only learned the average signal characteristics but also the variability and statistical spread inherent to each specific operating and fault conditions.

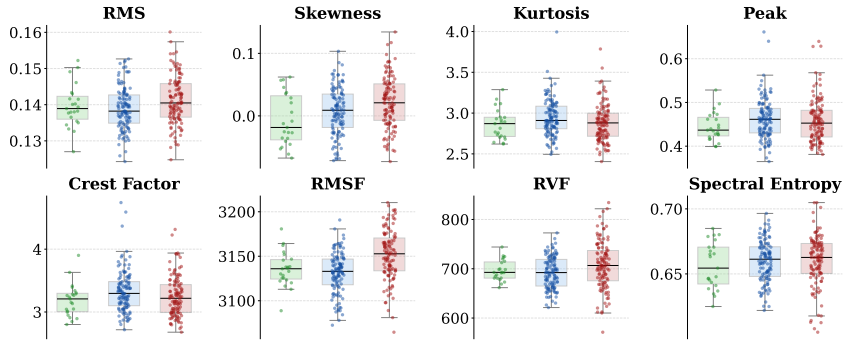
Regarding the CWRU dataset (Figure 6.4), the distributions of the extracted features are presented across four operational scenarios covering the machinery's dynamic range: from optimal health conditions to severe damage (outer race 0.021" fault). The synthetic signals exhibit feature values that are consistent with the real data across all evaluated states.

Although the overall distributional overlap is satisfactory, slight discrepancies can be observed in the healthy baseline scenario, where the synthetic model tends to slightly overestimate the RMSF and the spectral entropy. However, it is crucial to highlight that these minor deviations are predominantly confined to the healthy state. In contrast, discrepancies are almost negligible in the anomalous vibration scenarios, which represent the primary generative objective of this study.

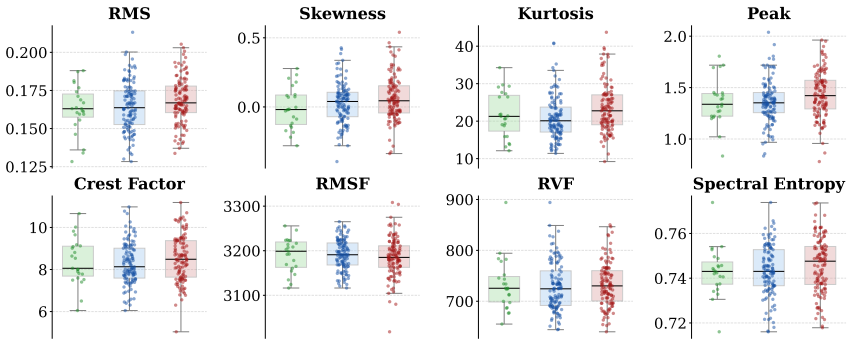
From a physical perspective, the vibration energy of a healthy baseline is inherently low. The generative model successfully reproduces these contained Root Mean Square (RMS) values, deviating from the real median by only 3%. Conversely, in scenarios characterised by evident fault impulses, such as inner and outer race faults, the overall signal energy substantially increases. This physical phenomenon is correctly reflected by a proportional rise in the RMS of the synthetic data. Impulsivity metrics, such as kurtosis and peak value, appropriately register low values for healthy



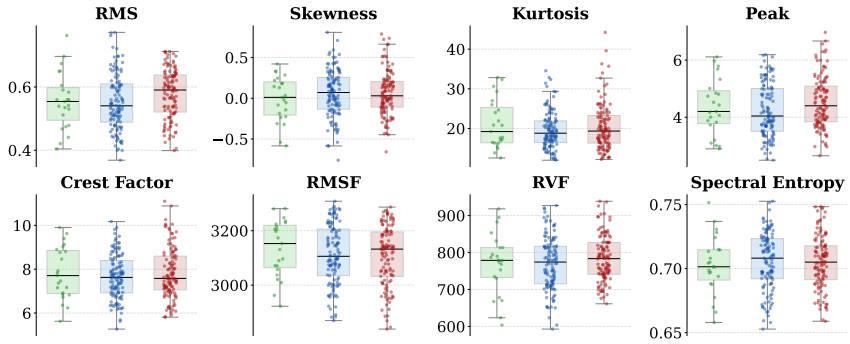
(a) Healthy Baseline (Load 1 HP)



(b) Ball Fault 0.007" (Load 1 HP)



(c) Inner Race Fault 0.014" (Load 1 HP)



(d) Outer Race Fault 0.021" (Load 1 HP)

Figure 6.4: Comparison of real and synthetic diagnostic feature distributions (CWRU dataset, 1 HP load) across different fault severity levels.

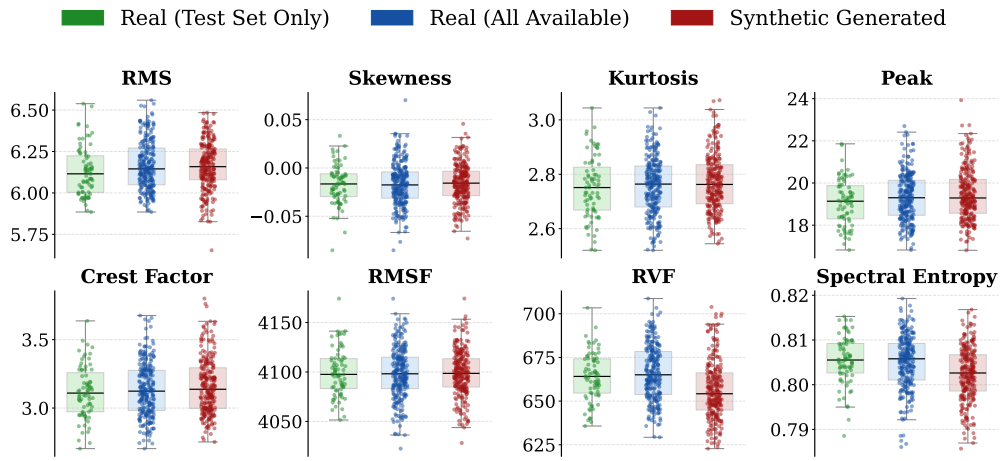
baseline signals and scale up accurately with fault severity, reaching their maxima in the severe outer race fault scenario (0.021"). In the frequency domain (RMSF, RVF, and spectral entropy), the generator accurately captures the feature distributions of the real data. This indicates that the model does not merely replicate the global energy in the time domain, but successfully allocates this energy across the correct frequency bands, while preserving the complexity and noise profile of the original power spectrum.

Moving to the EECSL dataset (Figure 6.5), while the analysis was conducted across all experimental configurations, the results obtained under a VFD regime at 40 Hz and a 75% load are presented here as a representative case. Under these conditions, the analysis evaluates three machinery states: healthy, eccentricity fault, and bearing fault. Consistent with the previously analysed DOL scenarios from the CWRU dataset, a strong agreement is observed also between the real and synthetic feature distributions under the VFD operation. Despite the inherent complexity of replicating spectral characteristics (RMSF, RVF, and spectral entropy), which are highly sensitive to the overall shape of the power spectrum, the synthetic distributions remain consistent with the real data across all considered operational states.

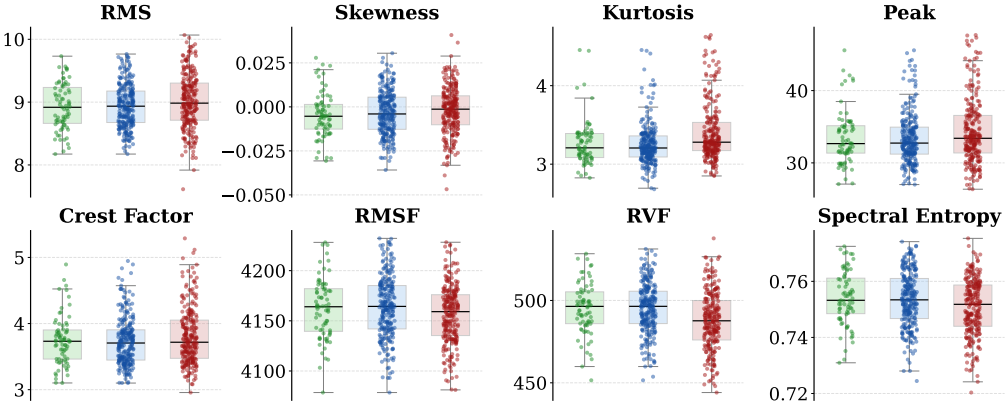
To further quantify the spectral fidelity of the generated signals, the heatmaps of the SCS and LSD metrics are presented for both the CWRU 12 kHz and the EECSL datasets in Figure 6.6.

Focusing first on the CWRU dataset, the mean synthetic spectra exhibit a strong shape alignment with those of the real samples, with SCS values consistently close to unity. The lowest similarity score is observed for the 3 HP baseline condition at 0.968, while particularly remarkable results are achieved for the most severe defects, such as the 0.021-inch inner race and outer race faults, which feature SCS values of approximately 0.998. Regarding the LSD, the heatmap generally shows small distances between the real and synthetic mean power spectra. The lowest distance values are observed for the 0.021-inch outer race fault, consistently falling below 1.0 dB.

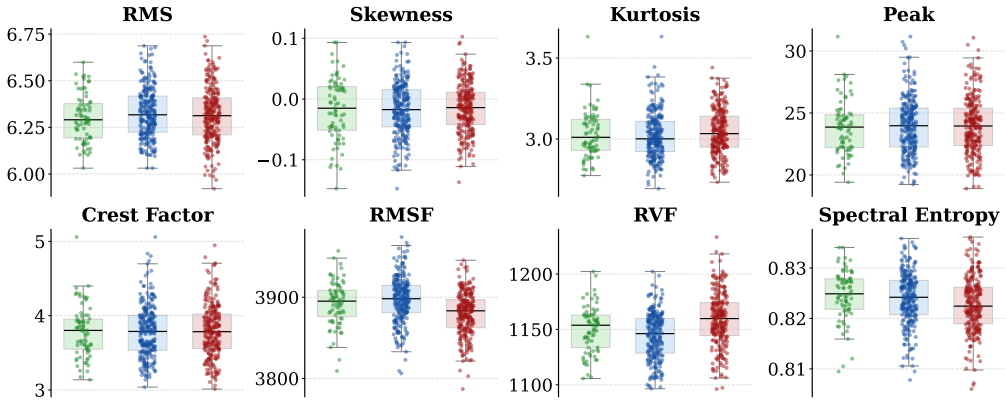
Extending the analysis to the EECSL dataset allows for the evaluation of the generative architecture under more complex, industrial-like operating conditions. Even with the introduction of power electronics, the SCS remains stable at optimal values across all VFD-driven regimes, consistently exceeding 0.990 and notably outperforming the DOL configurations. This dynamic is further confirmed by the LSD results, where the best overall performance is recorded for bearing faults under VFD supply. These conditions yield exceptional LSD values below 1.0 dB, reaching



(a) Healthy Baseline (VFD 40Hz, 75% Load)

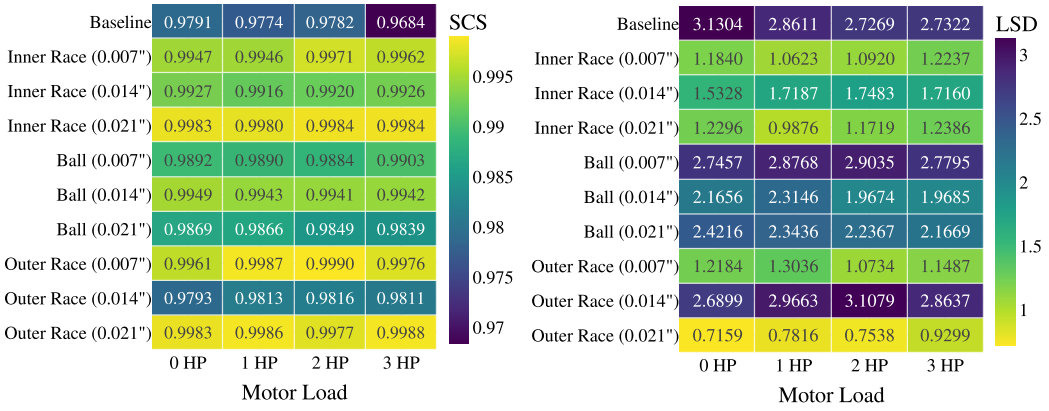


(b) Eccentricity Fault (VFD 40Hz, 75% Load)

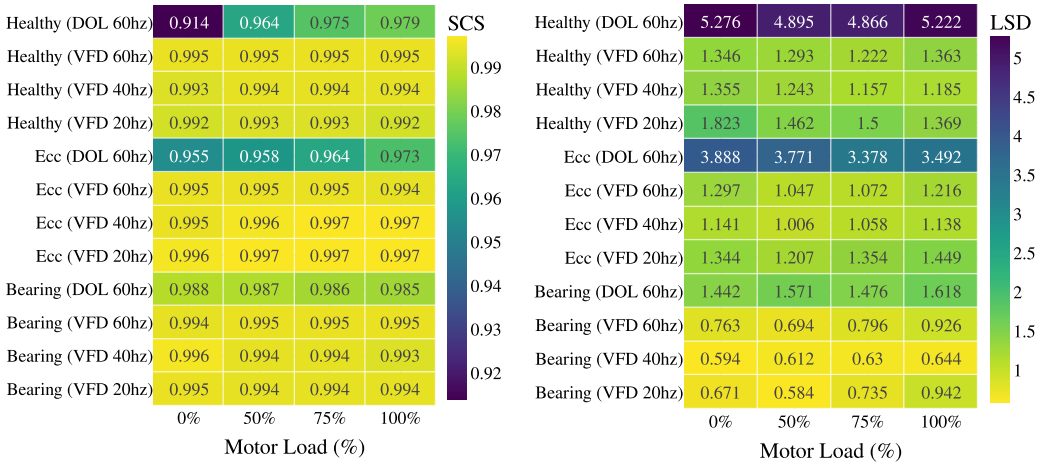


(c) Outer Race Fault (VFD 40Hz, 75% Load)

Figure 6.5: Comparison of real and synthetic diagnostic feature distributions (EECSL dataset, VFD 40Hz, Load 75%).



(a) SCS **(b)** LSD [dB]



(c) SCS **(d)** LSD [dB]

Figure 6.6: Heatmaps of SCS and LSD metrics for both the CWRU (top) and EECSL (bottom) datasets.

	Intra-Domain		Cross-Domain	
Dataset	**TRTR**	**TSTS**	**TSTR**	**TRTS**
CWRU (12 kHz)	99.02%	98.37%	98.26%	96.85%
CWRU (48 kHz)	98.39%	97.39%	98.54%	90.03%
EECSL	99.00%	98.13%	99.14%	97.11%

Table 6.1: Diagnostic performance across intra- and cross-domain scenarios for all the evaluated vibration datasets.

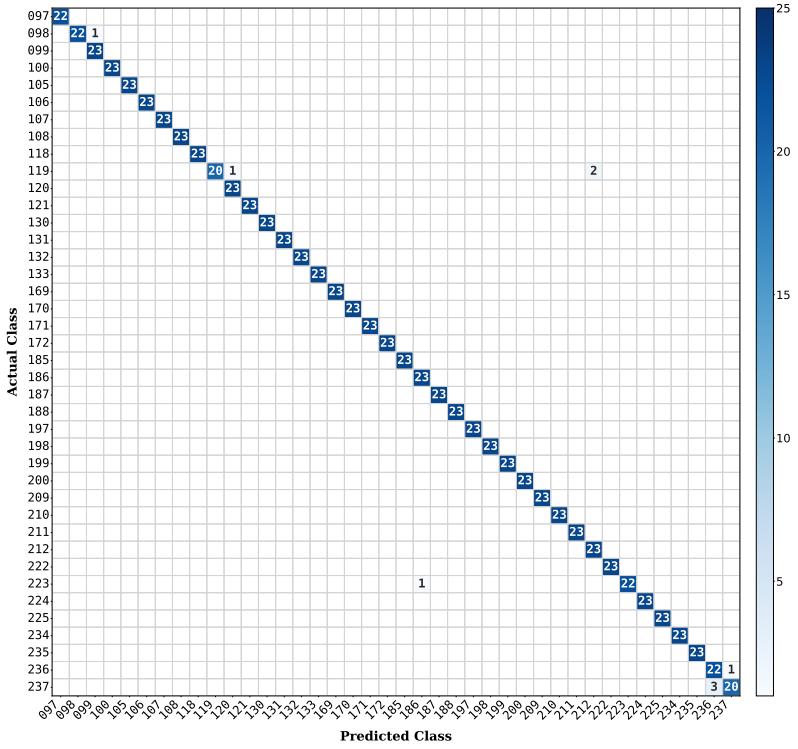
a minimum of just 0.584 dB at 20 Hz with a 50% load. Conversely, the healthy DOL 60 Hz class exhibits the highest mean spectral distances, spanning between 4.8 and 5.3 dB. Similarly, the eccentricity fault under the DOL regime features higher discrepancies (between 3.3 and 3.9 dB) compared to its VFD-driven counterpart (between 1.0 and 1.45 dB).

6.3 Downstream Classification Performance

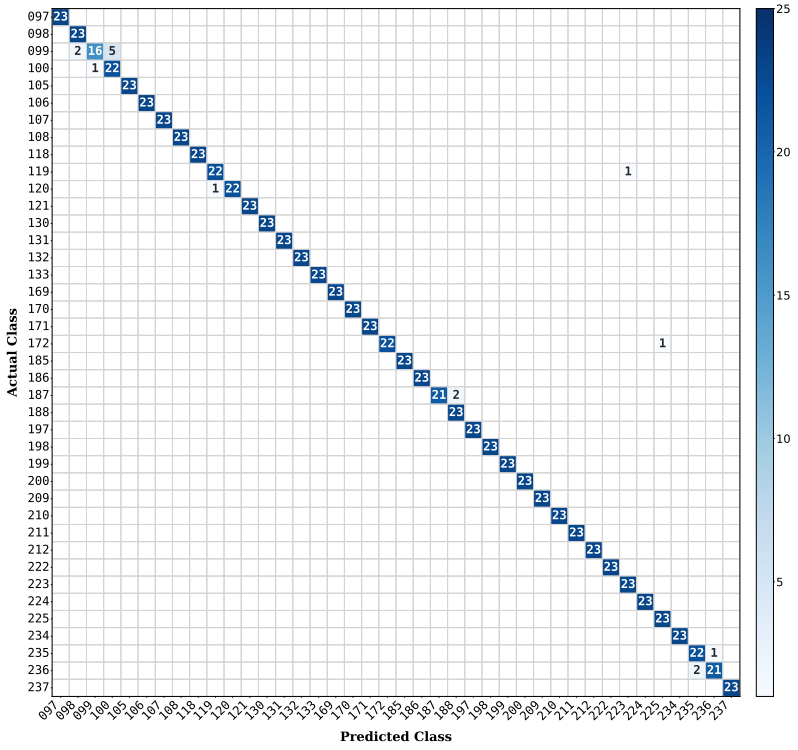
The definitive test of the diffusion model's performance lies in the practical diagnostic utility of its synthesised signals. The tailored CNN was therefore applied to the CWRU (12 kHz and 48 kHz) and the EECSL datasets across the four scenarios (TRTR, TSTS, TSTR, TRTS) defined in Section 5.8.2. The overall accuracy scores are summarised in Table 6.1, while the detailed classification behaviour for each dataset are visualised through dedicated confusion matrices in Figures 6.7, 6.8, and 6.9.

In the TRTR scenario, accuracy scores are high, and the few classification errors recorded relate to closely similar operating conditions, with minimal differences in terms of load. For instance, in the CWRU 12 kHz dataset, errors occur primarily between classes associated with files 236 and 237, both representing a 0.021-inch outer race fault differentiated only by motor load (2 HP vs. 3 HP). Similarly, in the EECSL dataset, isolated misclassifications are concentrated exclusively between adjacent load conditions (75% and 100%) at the same operating frequencies (20 Hz and 40 Hz) for the eccentricity defect.

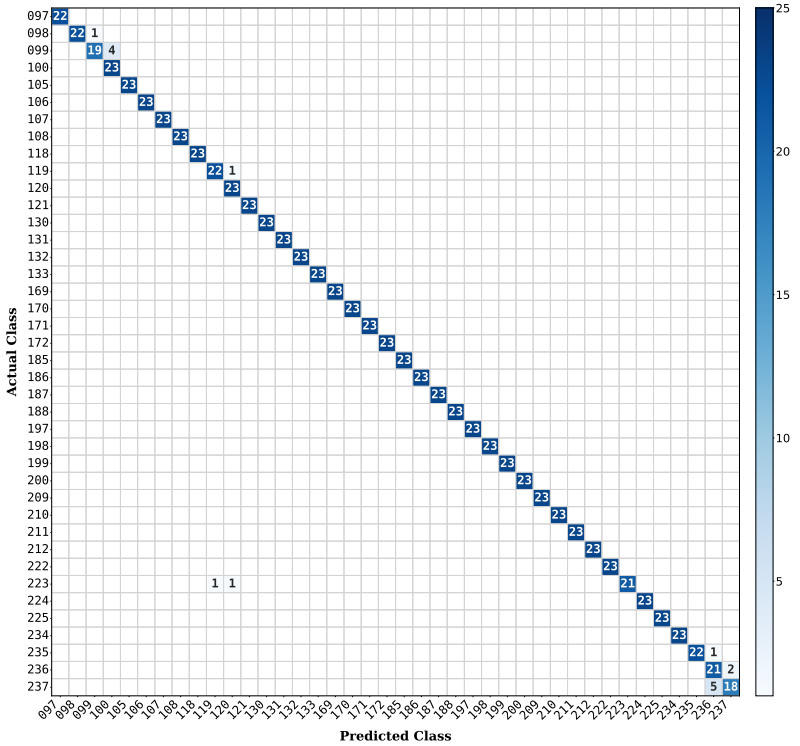
The crucial scenario for demonstrating the practical utility of the generated data is the TSTR configuration. Here, the results are consistently high across all datasets, with the classifier reaching an accuracy of 98.26% on the CWRU 12 kHz, 98.54% on the CWRU 48 kHz, and 99.14% on the EECSL dataset. This strong


Actual Class
Predicted Class
097 098 099 100 105 106 107 108 118 119 120 121 130 131 132 133 169 170 171 172 185 186 187 188 197 198 199 200 209 210 211 212 222 223 224 225 234 235 236 237
25
20
15
10
5

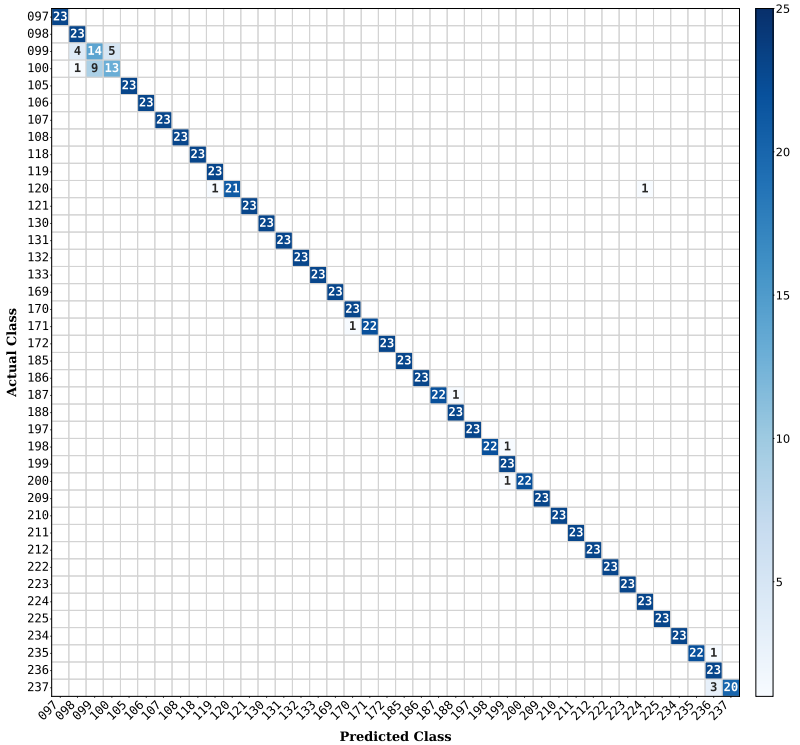
(a) TRTR


Actual Class
Predicted Class
097 098 099 100 105 106 107 108 118 119 120 121 130 131 132 133 169 170 171 172 185 186 187 188 197 198 199 200 209 210 211 212 222 223 224 225 234 235 236 237
25
20
15
10
5

(b) TSTS

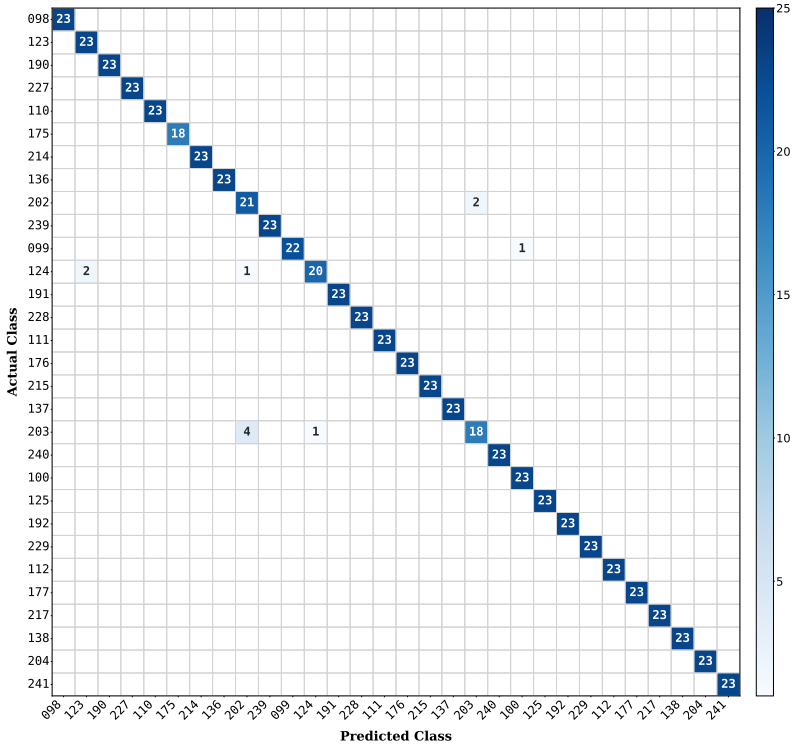


(c) TSTR

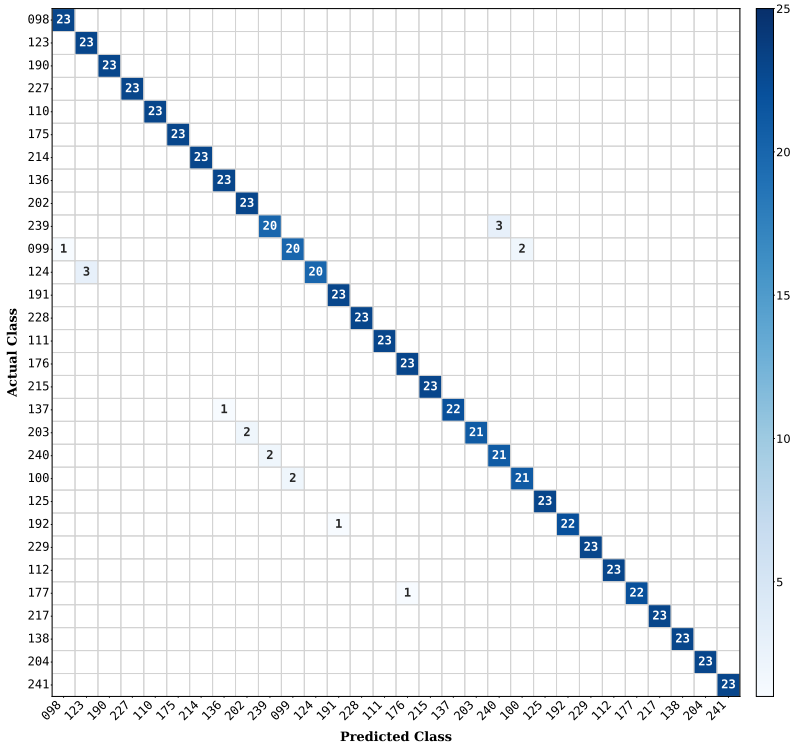


(d) TRTS

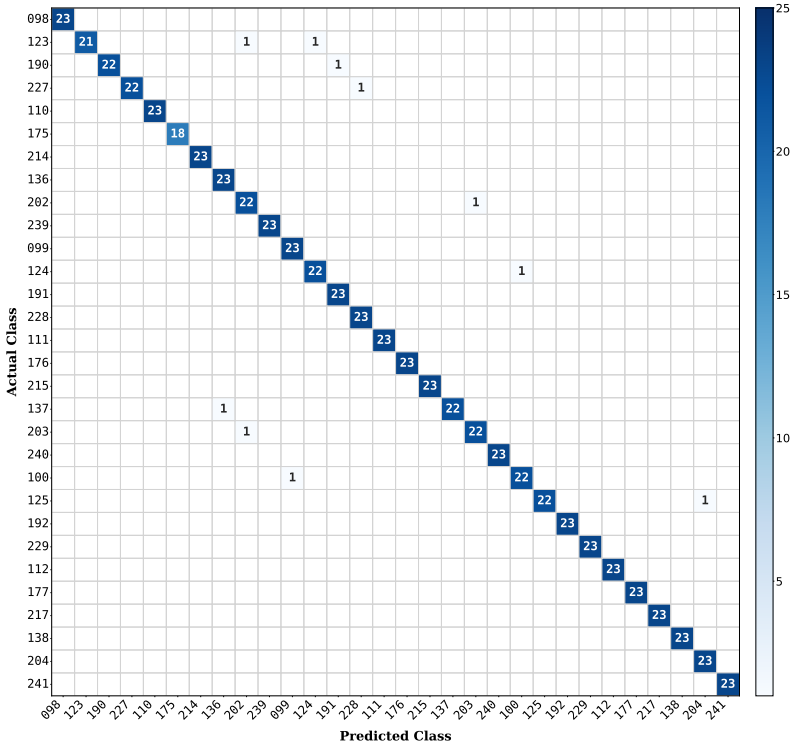
Figure 6.7: Confusion matrices for the four diagnostic evaluation scenarios (TRTR, TSTS, TSTR and TRTS) on the CWRU 12 kHz dataset.


Actual Class
Predicted Class
098 123 190 227 110 175 214 136 202 239 099 124 191 228 111 176 215 137 203 240 100 125 192 229 112 177 217 138 204 241
25
20
15
10
5

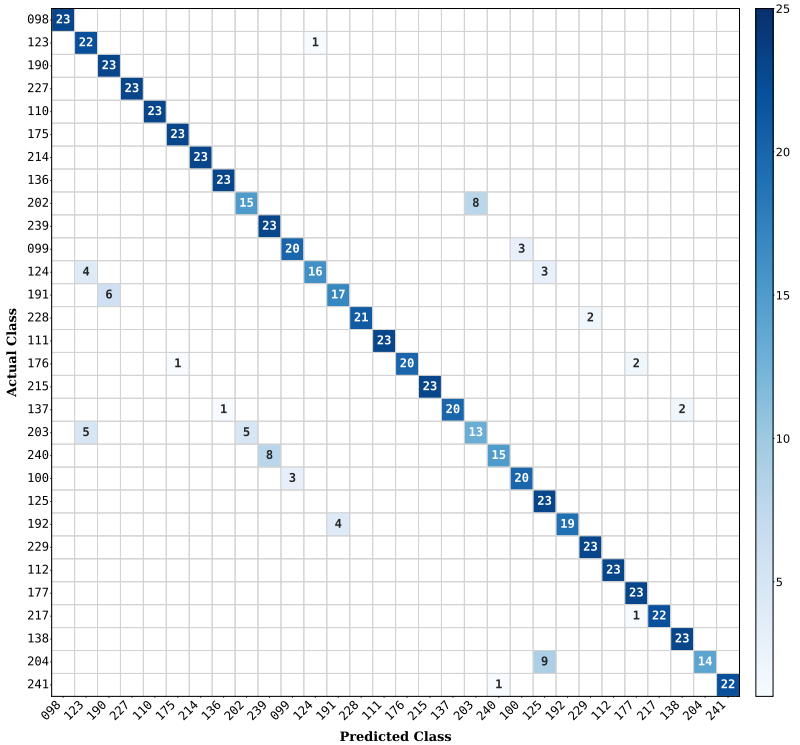
(a) TRTR


Actual Class
Predicted Class
098 123 190 227 110 175 214 136 202 239 099 124 191 228 111 176 215 137 203 240 100 125 192 229 112 177 217 138 204 241
25
20
15
10
5

(b) TSTS

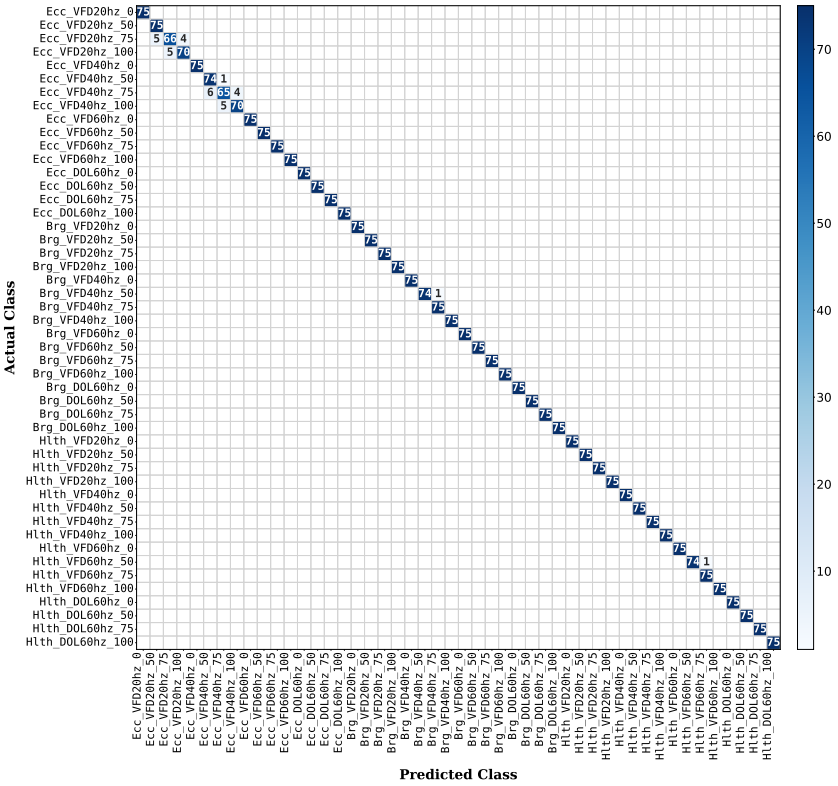


(c) TSTR

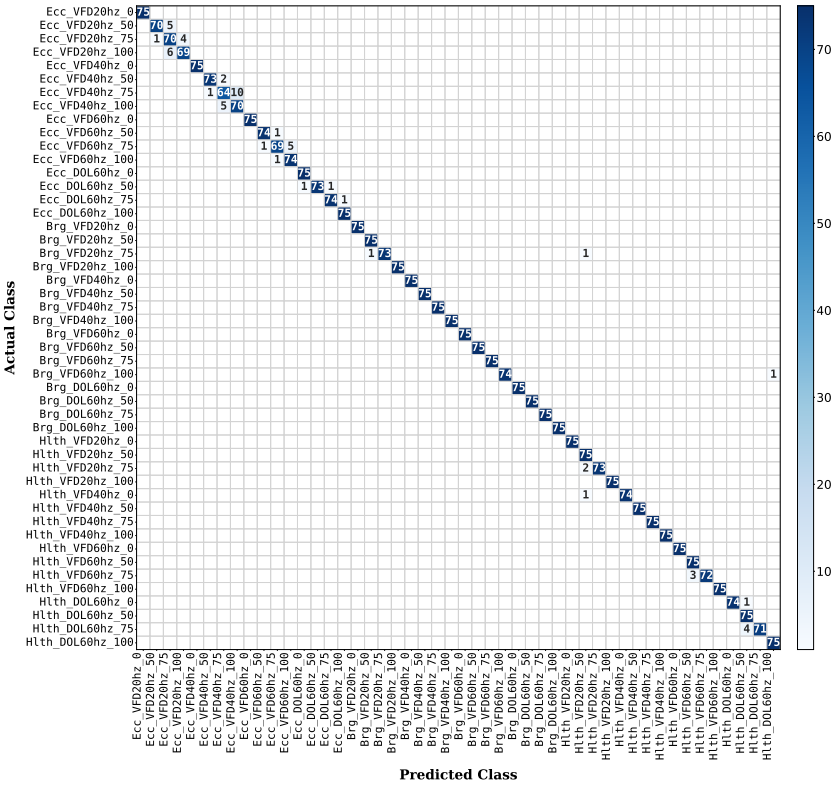


(d) TRTS

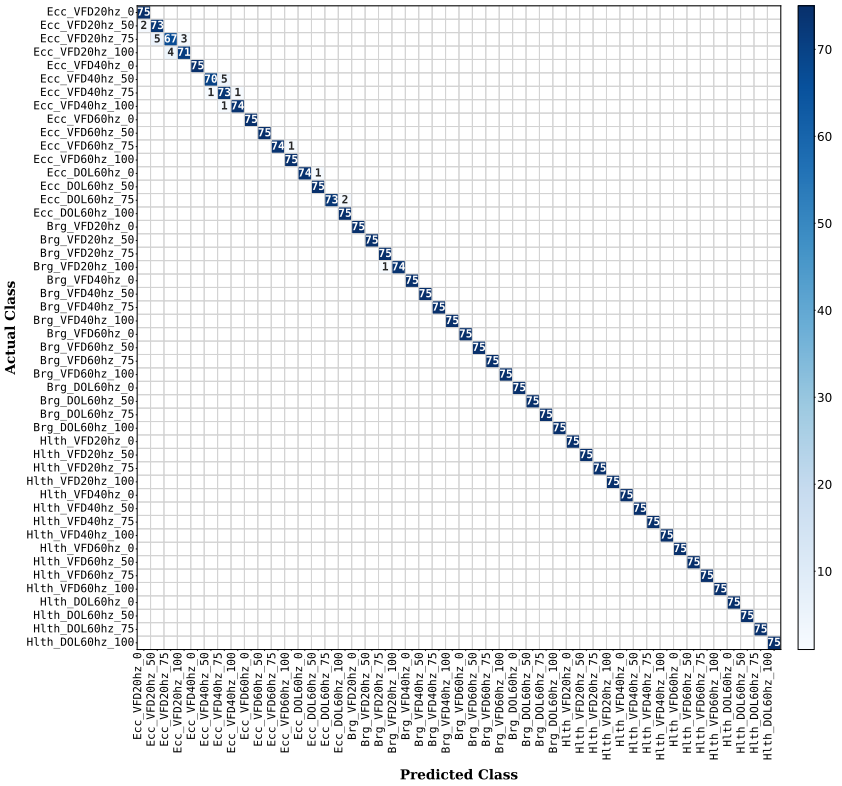
Figure 6.8: Confusion matrices for the four diagnostic evaluation scenarios (TRTR, TSTS, TSTR and TRTS) on the CWRU 48 kHz dataset.


Actual Class
Predicted Class
Ecc_VFD20hz_0
Ecc_VFD20hz_50
Ecc_VFD20hz_75
Ecc_VFD20hz_100
Ecc_VFD40hz_0
Ecc_VFD40hz_50
Ecc_VFD40hz_75
Ecc_VFD40hz_100
Ecc_VFD60hz_0
Ecc_VFD60hz_50
Ecc_VFD60hz_75
Ecc_VFD60hz_100
Ecc_DOL60hz_0
Ecc_DOL60hz_50
Ecc_DOL60hz_75
Ecc_DOL60hz_100
Brg_VFD20hz_0
Brg_VFD20hz_50
Brg_VFD20hz_75
Brg_VFD20hz_100
Brg_VFD40hz_0
Brg_VFD40hz_50
Brg_VFD40hz_75
Brg_VFD40hz_100
Brg_VFD60hz_0
Brg_VFD60hz_50
Brg_VFD60hz_75
Brg_VFD60hz_100
Brg_DOL60hz_0
Brg_DOL60hz_50
Brg_DOL60hz_75
Brg_DOL60hz_100
Hlth_VFD20hz_0
Hlth_VFD20hz_50
Hlth_VFD20hz_75
Hlth_VFD20hz_100
Hlth_VFD40hz_0
Hlth_VFD40hz_50
Hlth_VFD40hz_75
Hlth_VFD40hz_100
Hlth_VFD60hz_0
Hlth_VFD60hz_50
Hlth_VFD60hz_75
Hlth_VFD60hz_100
Hlth_DOL60hz_0
Hlth_DOL60hz_50
Hlth_DOL60hz_75
Hlth_DOL60hz_100
70
60
50
40
30
20
10

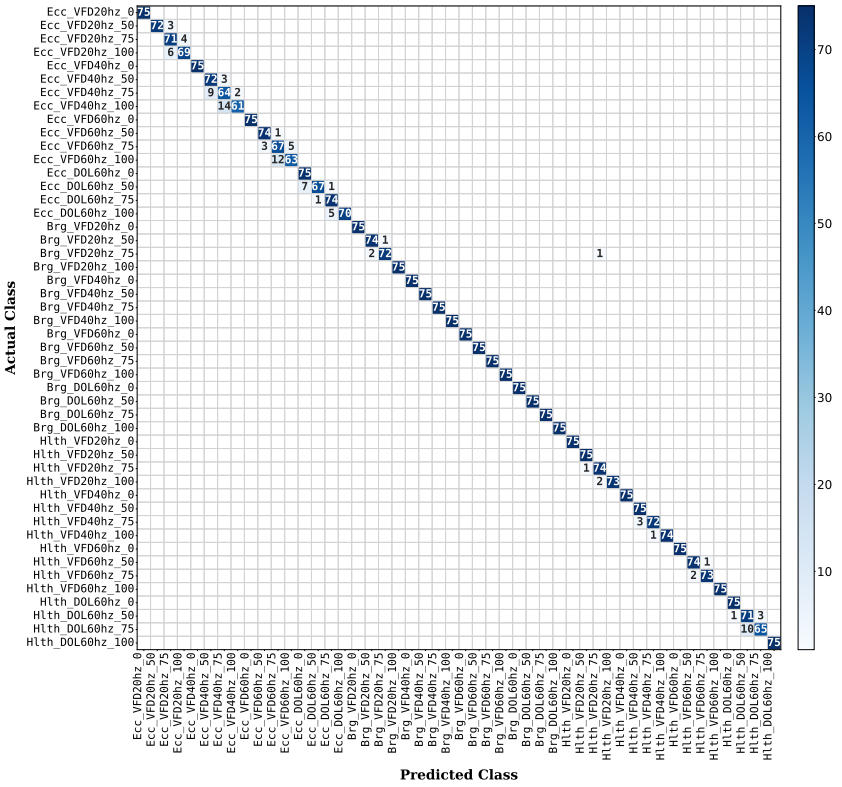
(a) TRTR


Actual Class
Predicted Class
Ecc_VFD20hz_0
Ecc_VFD20hz_50
Ecc_VFD20hz_75
Ecc_VFD20hz_100
Ecc_VFD40hz_0
Ecc_VFD40hz_50
Ecc_VFD40hz_75
Ecc_VFD40hz_100
Ecc_VFD60hz_0
Ecc_VFD60hz_50
Ecc_VFD60hz_75
Ecc_VFD60hz_100
Ecc_DOL60hz_0
Ecc_DOL60hz_50
Ecc_DOL60hz_75
Ecc_DOL60hz_100
Brg_VFD20hz_0
Brg_VFD20hz_50
Brg_VFD20hz_75
Brg_VFD20hz_100
Brg_VFD40hz_0
Brg_VFD40hz_50
Brg_VFD40hz_75
Brg_VFD40hz_100
Brg_VFD60hz_0
Brg_VFD60hz_50
Brg_VFD60hz_75
Brg_VFD60hz_100
Brg_DOL60hz_0
Brg_DOL60hz_50
Brg_DOL60hz_75
Brg_DOL60hz_100
Hlth_VFD20hz_0
Hlth_VFD20hz_50
Hlth_VFD20hz_75
Hlth_VFD20hz_100
Hlth_VFD40hz_0
Hlth_VFD40hz_50
Hlth_VFD40hz_75
Hlth_VFD40hz_100
Hlth_VFD60hz_0
Hlth_VFD60hz_50
Hlth_VFD60hz_75
Hlth_VFD60hz_100
Hlth_DOL60hz_0
Hlth_DOL60hz_50
Hlth_DOL60hz_75
Hlth_DOL60hz_100
70
60
50
40
30
20
10

(b) TSTS



(c) TSTR



(d) TRTS

Figure 6.9: Confusion matrices for the four diagnostic evaluation scenarios (TRTR, TSTS, TSTR and TRTS) on the EECSL 12 kHz dataset.

overall performance across all evaluated conditions proves that the generated signals are highly reliable for data augmentation in real-world industrial diagnostics, while simultaneously demonstrating the model's ability to reproduce the intrinsic diversity of the signal time windows. In the CWRU 12 kHz dataset, the main error concerns five samples of the 0.021-inch outer race fault under a 3 HP load being misclassified as its 2 HP counterpart, a specific classification error that was also present in the TRTR baseline. For the CWRU 48 kHz and EECSL datasets, discrepancies follow the same pattern and are numerically even lower than those of the empirical baseline. This alignment confirms that the decision boundaries learned by the classifier within the synthetic domain are equivalent to those required to discriminate the physical faults in the real domain. Notably, the few misclassifications that do occur are confined to adjacent load conditions, which produce only subtle differences in the vibration signal. The fact that the classifier trained exclusively on synthetic data successfully discriminates between different loads demonstrates that the generative model has also faithfully captured the fine-grained inter-load characteristics of the real signals.

In the TRTS scenario the highest performance was achieved on the EECSL dataset (97.11%), followed closely by the CWRU 12 kHz dataset (96.85%), with errors again restricted to adjacent load conditions, mirroring the TRTR baseline. The high-resolution CWRU 48 kHz dataset exhibits a slightly lower accuracy of 90.03%, a predictable drop caused by the higher sampling frequency. Specifically, to capture the same physical time interval, the model must generate sequences four times longer than those at 12 kHz. Given the substantial increase in generative complexity required to process such large windows, maintaining an accuracy above 90% remains a notable achievement that confirms the model's ability to successfully handle complex, long-term temporal dependencies.

Finally, the high values recorded in the TSTS scenario confirm the excellent separability of the generated classes, indicating that the CNN can trace clear decision boundaries within the synthetic domain as well.

6.4 t-SNE Visualisation of Distribution Alignment

The t-SNE algorithm was utilised to evaluate both the hand-crafted diagnostic feature space and the deep representation space extracted by the CNN, as detailed in Section 5.8.3.

Starting from the first, Figure 6.10a illustrates the resulting two-dimensional

map for the CWRU dataset. Specifically, the test set and an equivalent number of synthetic samples for each class are selected and grouped into macro-classes: baseline, inner race fault, outer race fault, and ball fault.

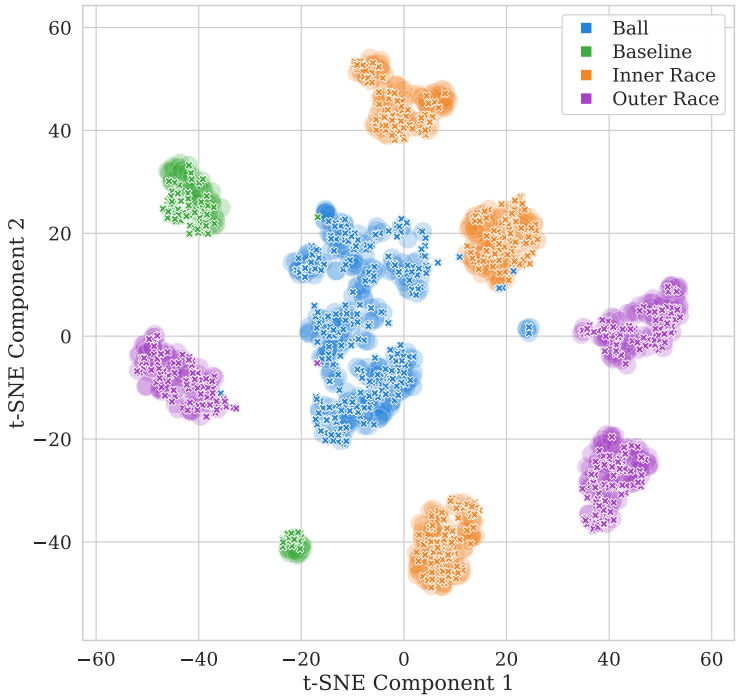
The figure reveals a strong overlap between the two distributions: the data points representing the synthetic samples (crosses) fall within the clusters formed by the real samples (circles) and cover the spatial extent of the real clusters without exhibiting isolated groups. Although a few minor outliers are observable especially for the ball class, and the baseline condition marginally deviates from the real data distribution compared to other classes, the overall multidimensional alignment remains highly satisfactory. This result suggests that the diffusion model successfully captures the joint distribution of the diagnostic features, replicating a statistical behaviour that is consistent with the real samples for the same classes.

Analysing the spatial topology, it is also evident that the inner race and outer race macro-classes naturally divide into three distinct sub-clusters. This separation faithfully reflects the three different fault severity conditions (defect diameters) present in the dataset, confirming the high discriminative power of the extracted features for raceway REB faults. Conversely, the samples belonging to the ball class converge into a single, contiguous macro-region. This behaviour is physically justified by the complex kinematics of rolling element defects, whose vibration signatures are less separable as fault severity changes.

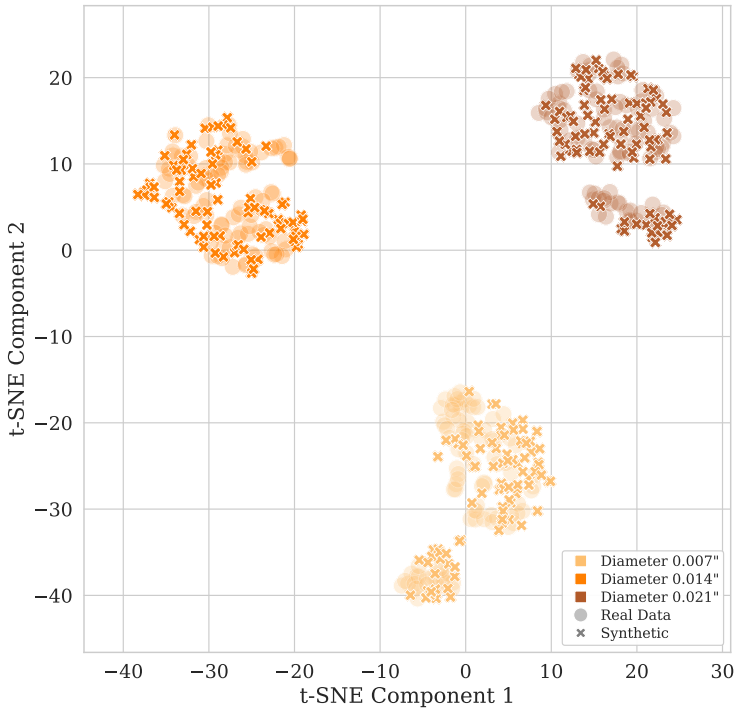
A more detailed analysis focused exclusively on the inner race class (Figure 6.10b) further corroborates these findings. By highlighting the macro-classes corresponding to the three severity levels, it is shown that the synthetic samples replicate the distribution of the real diagnostic features. Interestingly, when observing the influence of the four operating loads within each severity level, the data do not split into four isolated clusters, reflecting the minimal variations in the vibration signal induced by changes in motor load.

In a subsequent experiment, the t-SNE visualisation is extended to the deep feature space. Figures 6.11 and 6.12 correspond to the CWRU 12 kHz and EECSL datasets, respectively. For each operational condition, the test set and an equal number of synthetic samples are plotted (23 per class for the CWRU dataset and 75 for the EECSL dataset).

Both figures reveal a good alignment between the two domains, with almost no isolated synthetic data points (crosses). This indicates that the generated samples possess the same diagnostic signatures as their real counterparts. Furthermore, the intra-class distribution is well preserved since the synthetic points effectively span



(a) Macro-class mapping



(b) Severity analysis

Figure 6.10: t-SNE visualisation of the diagnostic feature space (CWRU 12 kHz dataset). Panel (a) shows the spatial overlap between real and synthetic data across macro-classes. Panel (b) details the varying severity levels of the inner race fault.

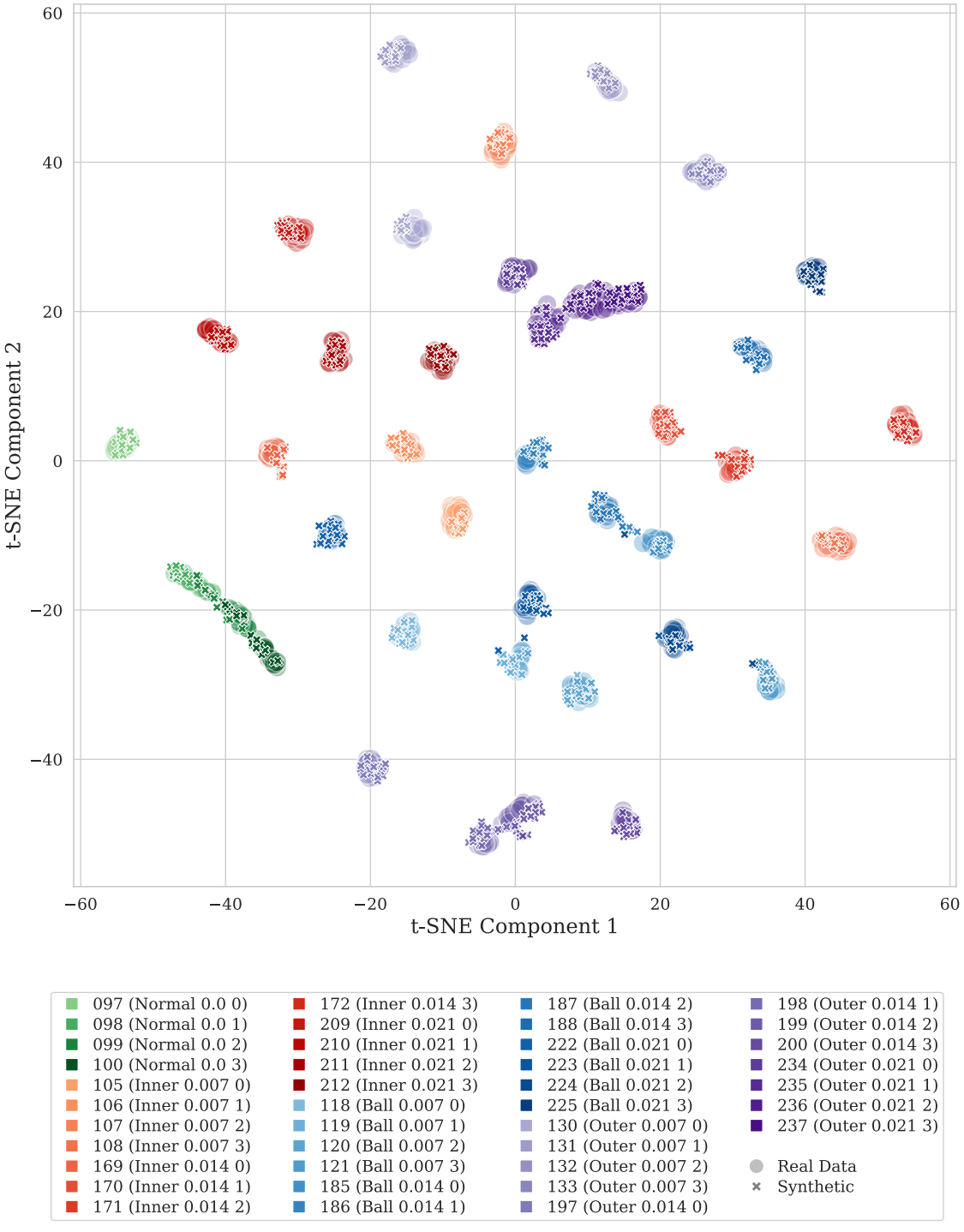


Figure 6.11: t-SNE visualisation of the deep representations extracted from the pre-trained CNN (CWRU 12 kHz, 23 samples per class).

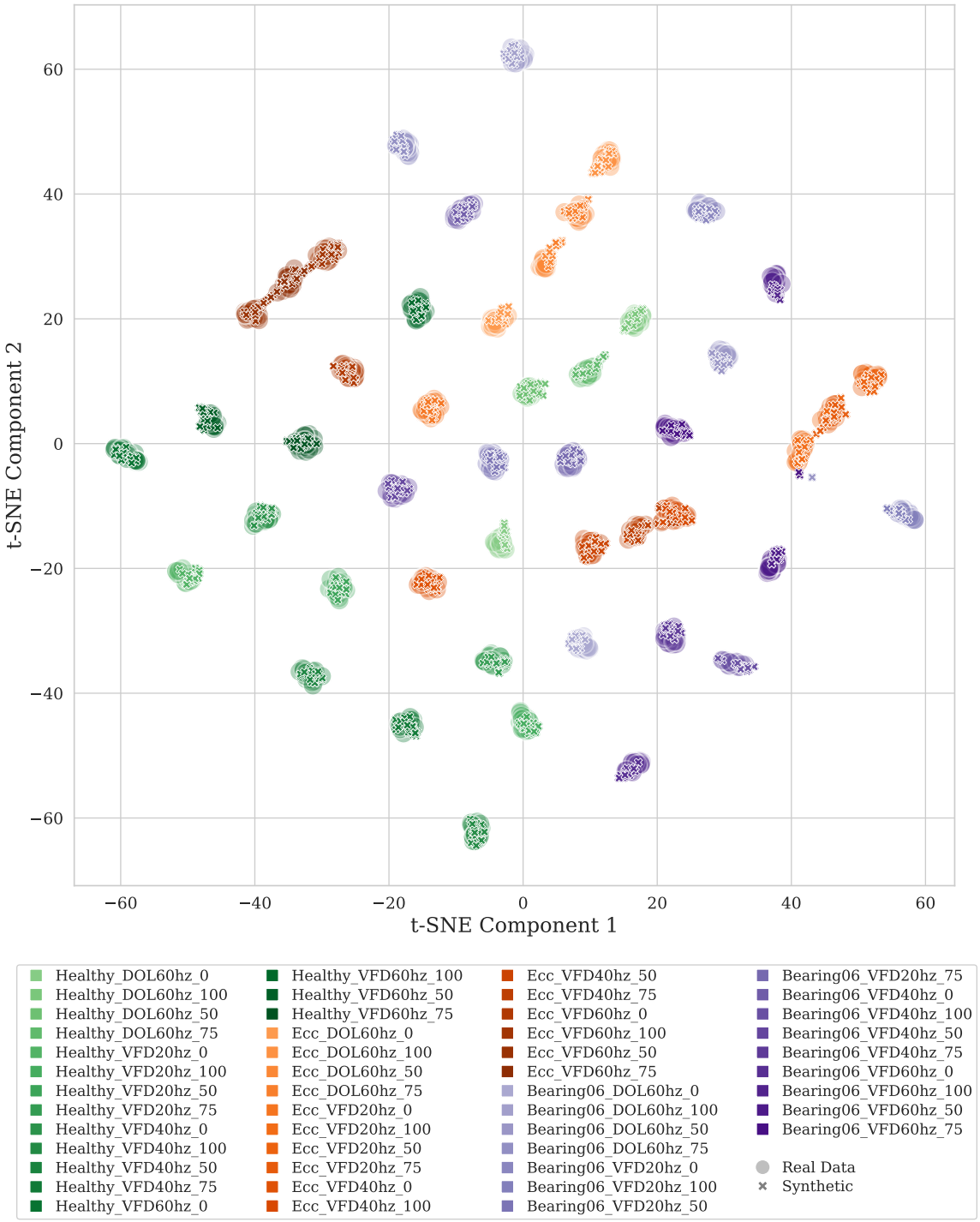


Figure 6.12: t-SNE visualisation of the deep representations extracted from the pre-trained CNN (EECSL, 75 samples per class).

the area of the real clusters rather than collapsing into a single point, confirming the model's generative diversity. Finally, these plots highlight the high discriminative power of the CNN: despite the high number of individual classes, the clusters for each specific condition are tightly confined and clearly separated from one another.

6.5 Memorisation Analysis Results

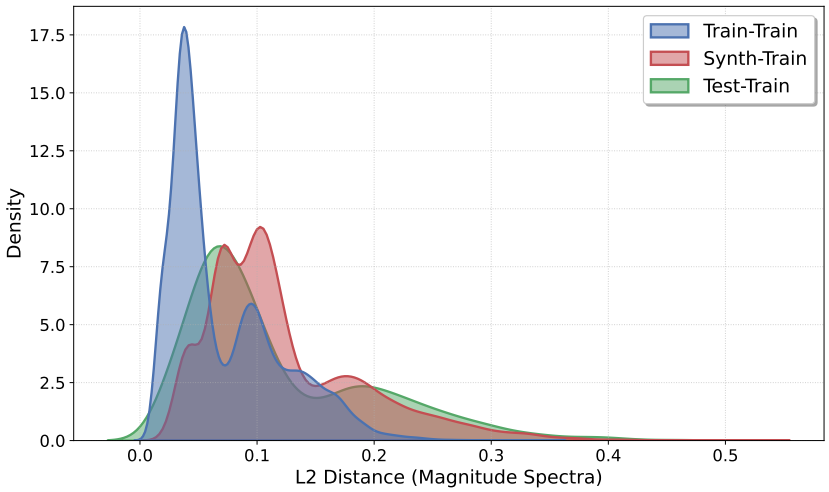
The results of the 1-NN distance analysis confirm that the generative model successfully avoids data replication. As illustrated in the global KDE plots (Figure 6.13), the distance distribution of the synthetic samples (red) does not exhibit a peak near zero. This confirms that the generative model does not simply memorise and reproduce training samples.

As expected, the intra-training distribution (Train-to-Train) exhibits a pronounced peak near zero. This is a direct consequence of the highly overlapping windowing strategy applied to the CWRU dataset, which features a higher degree of overlap compared to the EECSL dataset. Meanwhile, the overall distance distribution of the synthetic samples mirrors that of the unseen test data (Test-to-Train), suggesting that generated signals introduce a degree of novelty and natural variability consistent with the true underlying physical phenomenon.

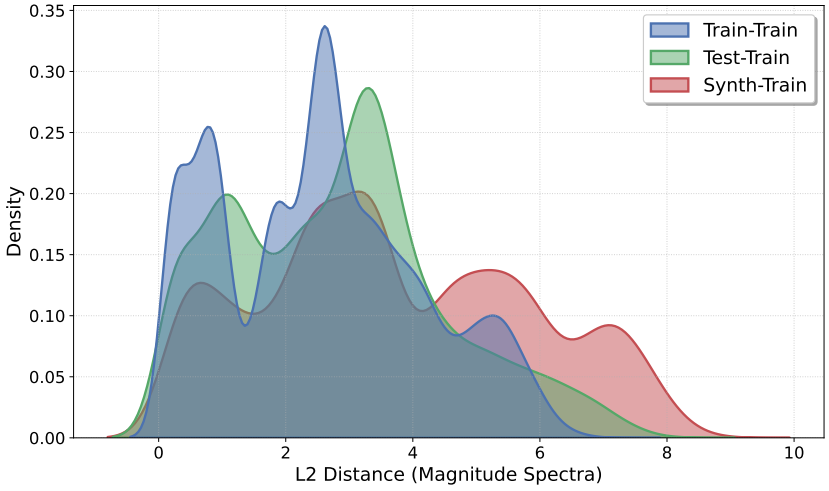
To robustly exclude the possibility of sporadic data replication, a sufficiently large set of synthetic samples, specifically 300 samples per class, was evaluated. Despite the numerical disparity between the synthetic and the real test samples per class, the probability density normalisation inherent to KDE ensures a fair comparison, independent of the number of instances in each set.

6.6 Zero-Shot Generation Results

To establish an initial proof of consistency in both the time and frequency domains for the zero-shot synthesised signals, a qualitative comparison is presented. Figures 6.14a and 6.14b illustrate the time-domain waveforms and magnitude spectra for two representative withheld conditions: the Inner Race fault (0.014 inches, 2 HP) from the CWRU dataset and the Outer Race fault (VFD 40 Hz, 75% load) from the EECSL dataset. A visual inspection reveals that the synthetic signals accurately replicate the real empirical waveforms in the time domain. In the frequency domain, the synthesised magnitude spectra remain within the frequency boundaries established by the real samples, successfully reproducing the spectral signature of the specific



(a) CWRU



(b) EECSL

Figure 6.13: Probability density distributions of the 1-NN frequency-domain Euclidean distances for the (a) CWRU and (b) EECSL datasets, estimated through KDE.

fault conditions.

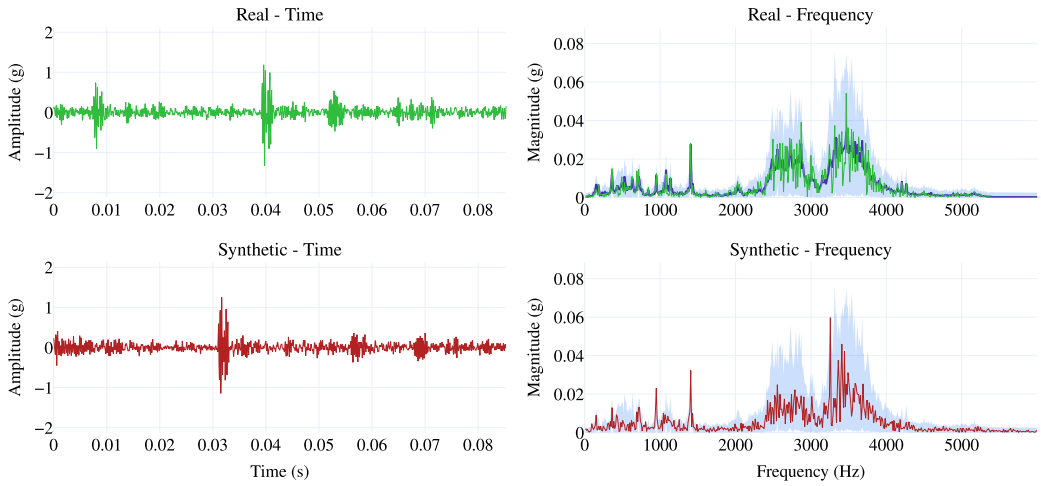
This visual assessment is statistically confirmed by the diagnostic features extracted from the signal windows and shown in strip plots (Figure 6.15). The alignment between the features extracted from the real and zero-shot synthetic signals is highly consistent across both datasets. For the EECSL dataset, minor deviations are observed in specific spectral features, in particular RMSF and RVF, likely due to the increased spectral complexity introduced by PWM harmonic distortion.

A fundamental evaluation, which represents a practical application of the proposed generative model, involves assessing the downstream classifier's performance when trained on synthetic data that includes zero-shot generated signals for the withheld target conditions.

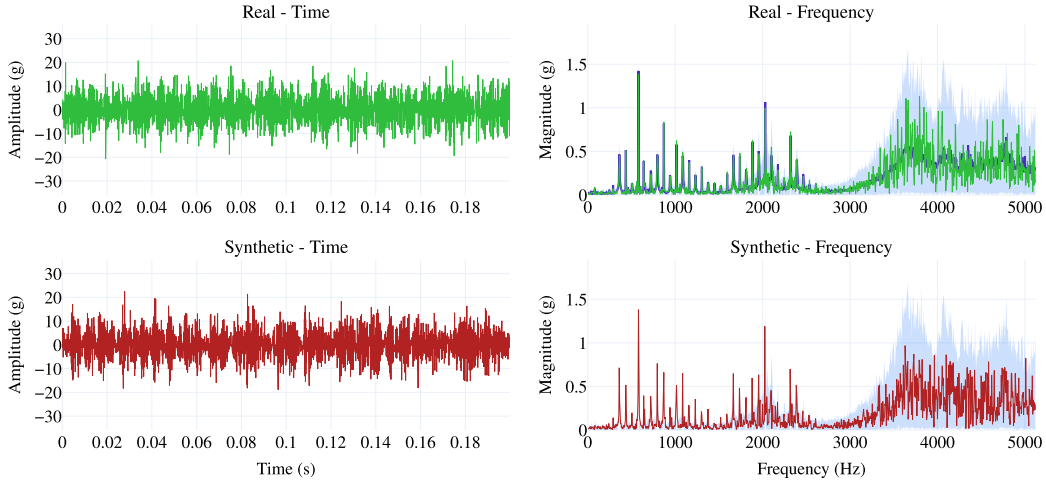
For the CWRU dataset, the overall classification accuracy reaches 96.63%, as detailed in the confusion matrix of Figure 6.16. The classification performance for the withheld conditions is strong: the 0.014-inch inner race fault (ID 171) and the 0.014-inch ball fault (ID 187) achieve perfect classification with no misclassifications, while the withheld outer race fault (ID 199) records only 2 out of 23 samples misclassified into the adjacent load category. It should be noted that the adoption of guidance scale $w = 2.0$, while useful to achieve high conditioning fidelity for the fault classes, slightly degrades classification performance for the healthy baseline signals, whose inherently low-energy vibration signatures make them more susceptible to distortion under high conditioning amplification.

Consistent results are achieved on the EECSL dataset, where the overall accuracy reaches 99%, as shown in the confusion matrix of Figure 6.17. For all three withheld conditions, healthy baseline, eccentricity fault, and bearing fault under VFD 40 Hz at 75% load, the classifier achieves perfect classification, correctly identifying all 75 real test samples per class. This result is significant given that the load variation from 50% to 100% produces only subtle differences in the vibration signature. The generative model successfully captures even these fine-grained inter-load characteristics, enabling the classifier to discriminate between adjacent loading conditions.

These results demonstrate the potential of diffusion-based generative models as a practical tool for data augmentation in industrial bearing diagnostics, reducing the need for exhaustive experimental campaigns.

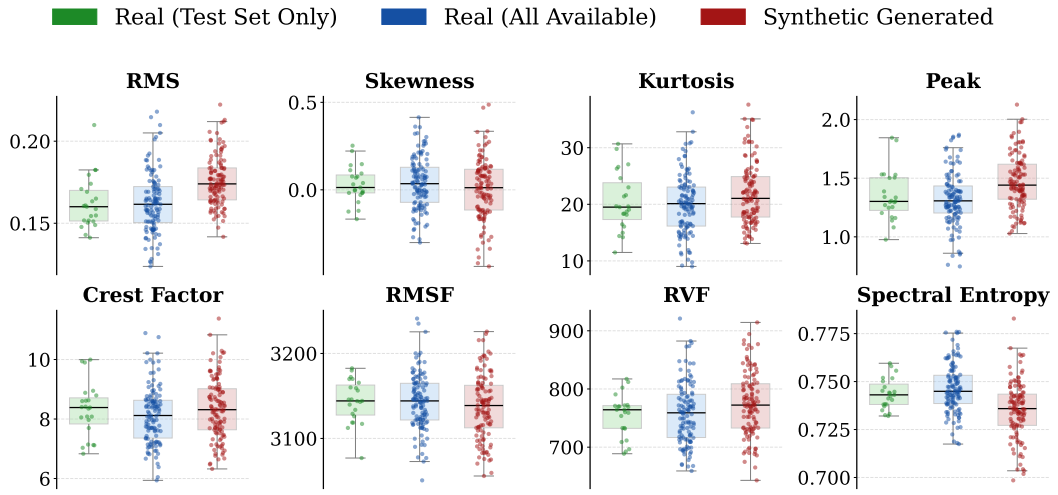


(a) Inner Race fault (0.014 inches, 2 HP Load)

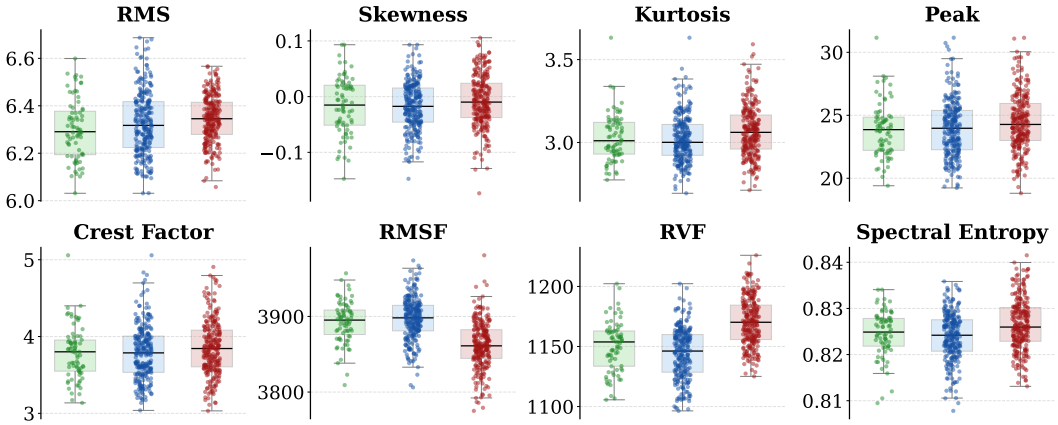


(b) Outer Race fault (VFD 40 Hz, 75% Load)

Figure 6.14: Time-domain waveforms and frequency-domain magnitude spectra comparing real signals and zero-shot generated samples for the (a) CWRU and (b) EECSL datasets.



(a) Inner Race fault (0.014 inches, 2 HP Load)



(b) Outer Race fault (VFD 40 Hz, 75% Load)

Figure 6.15: Comparison of real and synthetic zero-shot diagnostic feature distributions across the (a) CWRU and (b) EECSL datasets.

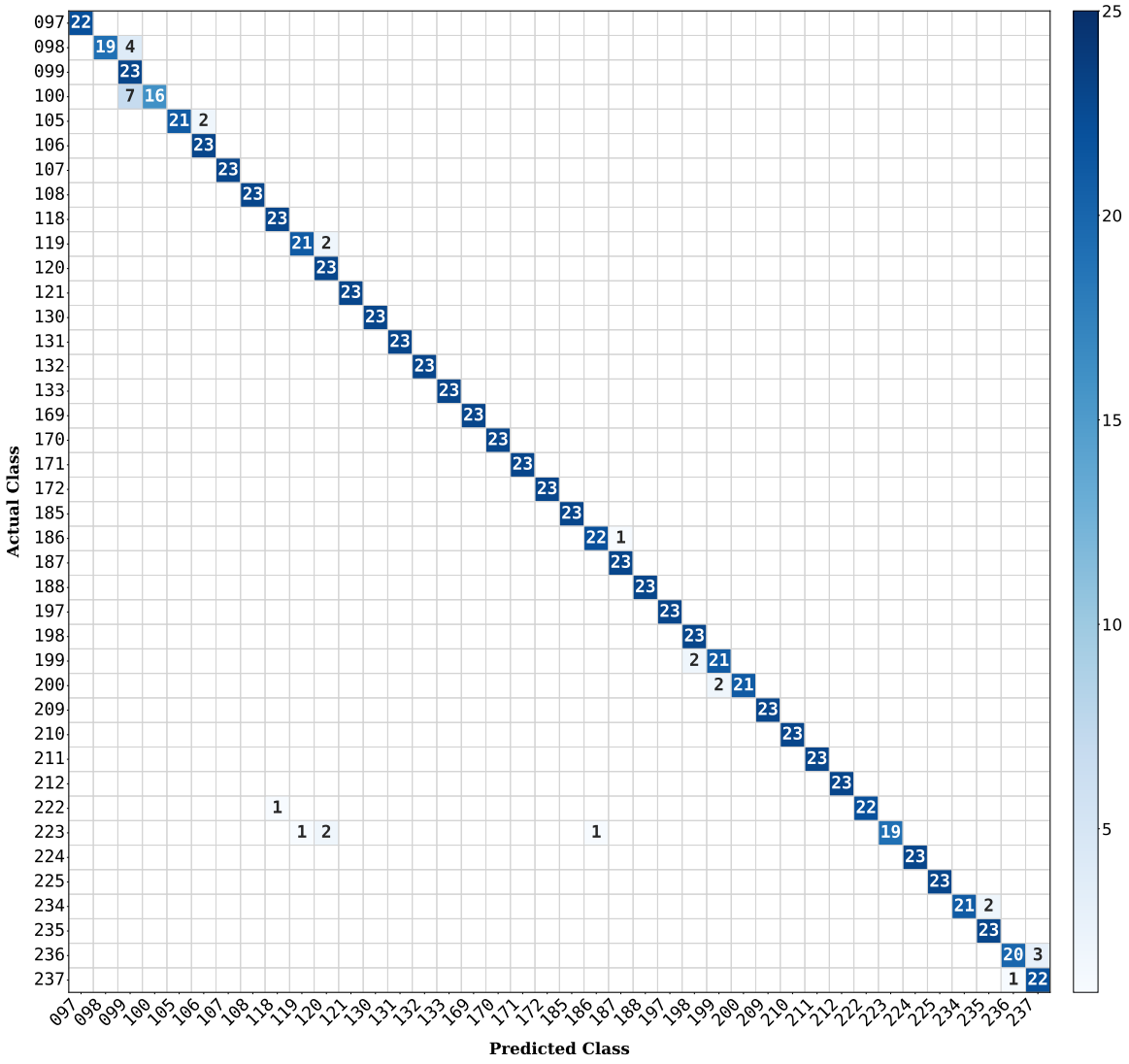


Figure 6.16: Confusion matrix of the classifier trained on zero-shot generated windows for the withheld conditions (CWRU).

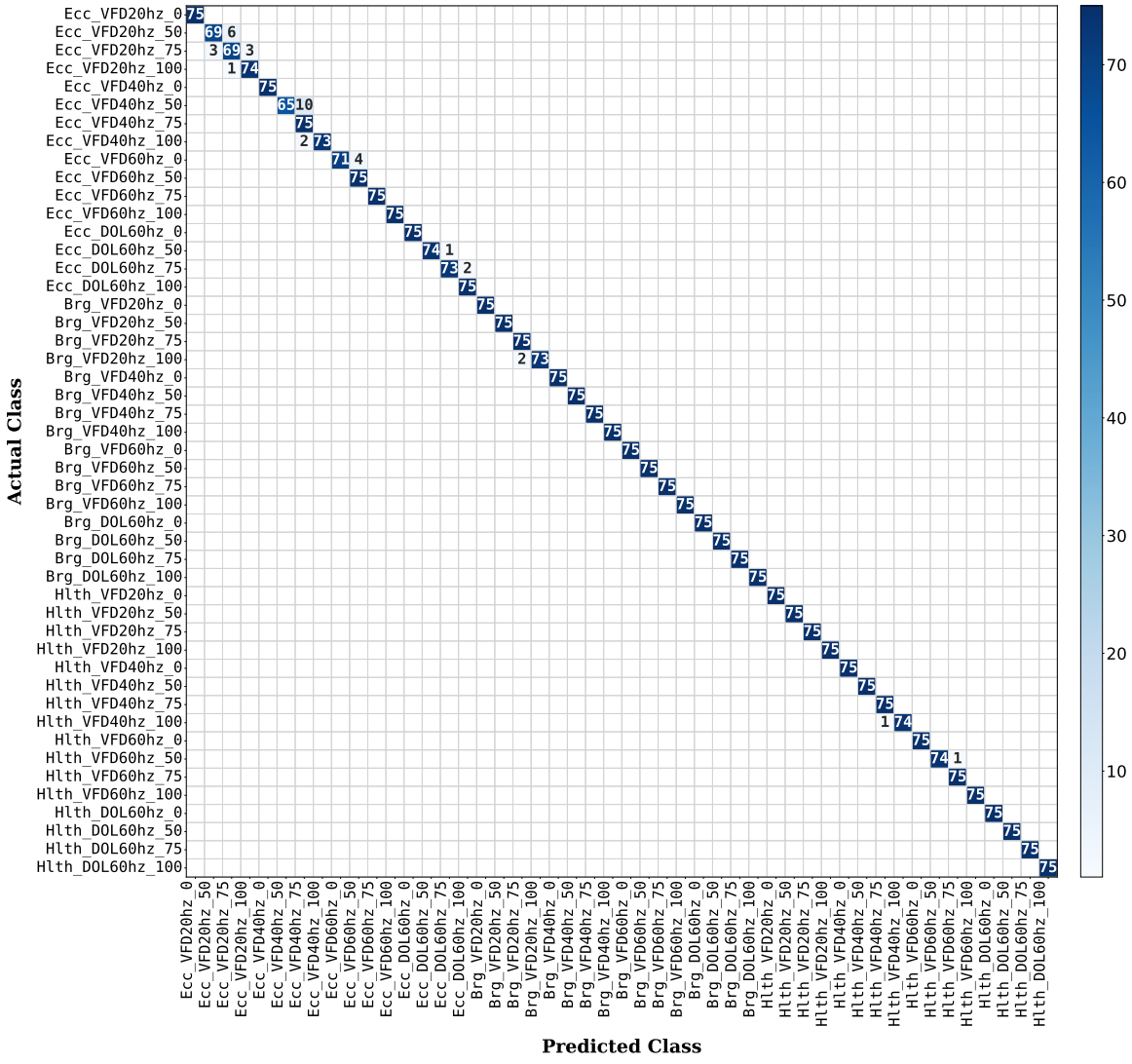


Figure 6.17: Confusion matrix of the classifier trained on zero-shot generated windows for the withheld conditions (EECSL).

Chapter 7

Conclusions

This thesis addressed the design and development of a novel 1D conditional diffusion framework capable of generating high-fidelity synthetic vibration signals directly in the time domain for rotating machinery diagnostics.

The adopted DDPM architecture mitigates the inherent limitations of alternative generative models: specifically, it avoids the mode collapse and training instability typical of GANs, while overcoming the severe oversmoothing of high-frequency transients observed in VAEs.

The success of the framework relies on key architectural choices, primarily the integration of local and global self-attention mechanisms across the U-Net, combined with a dual-stage AdaGN within each residual convolutional block to ensure robust conditioning. Furthermore, generation quality was enhanced by replacing the standard MSE loss with a hybrid loss function incorporating a spectral component, forcing the model to simultaneously respect both time- and frequency-domain characteristics of the target vibration signals.

Thanks to these structural design choices, the model proved capable of accurately generating vibration signals for localised bearing faults, conditioned on varying fault types, severities, and operating conditions including motor load and supply frequency, demonstrating robust performance across both DOL and VFD driving regimes. To validate its cross-domain versatility, the model was additionally trained and evaluated on electromechanical anomalies, successfully synthesising signals with the discriminative signature of dynamic eccentricity faults.

The most significant result of this work lies in the successful zero-shot generation of operating conditions never observed during training. This capability suggests that the model captured the underlying physical dynamics of the faults rather than merely memorising statistical patterns from the training data.

From an industrial standpoint, this framework enables the creation of comprehensive synthetic datasets for diagnostic applications without the need to physically operate machinery under every possible fault condition, load and speed regime, a process that is both economically costly and operationally hazardous. This provides a safe, cost-effective, and scalable solution to the pervasive problem of data scarcity and class imbalance in rotating machinery fault diagnosis.

7.1 Limitations

While the proposed conditional diffusion framework demonstrates high fidelity in synthetic vibration generation, some limitations should be acknowledged.

First, the framework generates fixed-length signal windows independently, without temporal coherence between consecutive segments. This design choice is fully adequate for the primary intended application, data augmentation for window-based real-time diagnostic tools, but would require architectural extensions to support applications demanding temporally continuous signal synthesis, such as the simulation of progressive fault degradation trajectories.

Additionally, the generation is currently restricted to isolated fault conditions, specifically localised bearing defects and dynamic eccentricity.

A further limitation concerns the evaluation of the zero-shot generation capability. In this study, the model was tested on new combinations of conditioning parameters, i.e., fault severity, motor load and supply frequency, that had been individually observed during training under distinct configurations. However, evaluating zero-shot extrapolation on completely unseen fault severities was hindered by the sparsity of the CWRU benchmark, which provides only three discrete severity levels of 0.007, 0.014, and 0.021 inches. This sparse sampling is insufficient for the model to learn a reliable continuous mapping along the severity variable. Furthermore, exploring denser alternative datasets would have substantially increased the computational burden, considering that training the proposed diffusion model already requires approximately 12 hours for 300 epochs on the selected CWRU configuration.

Finally, a structural limitation inherent to DDPMs is the high computational cost during inference. The reverse denoising process requires 1,000 sequential steps to generate each sample, making the framework significantly slower than single-pass generators such as GANs. This latency makes the direct use of the diffusion model unsuitable for real-time applications. This computational burden is further compounded by the integration of local and global self-attention mechanisms, which,

while crucial for signal fidelity, result in a model comprising approximately 33 million parameters.

7.2 Future Directions

To address the current limitations and advance the framework toward real-world industrial deployment, several directions for future research are proposed:

- **Expansion to novel fault typologies:** The model could be extended to generate vibration signals for a broader range of mechanical anomalies, including broken rotor bars, mechanical looseness, mass imbalance, and shaft misalignment. A particularly challenging and industrially relevant extension would be the synthesis of compound faults, where multiple defects occur simultaneously and their vibration signatures physically overlap and modulate each other.
- **Enhanced zero-shot evaluation:** The zero-shot generation capability should be further investigated on datasets with finer conditioning granularity. This would enable a more rigorous evaluation of the model's ability to generalise across entirely unobserved severity dimensions.
- **Validation in industrial environments:** An important open challenge is to validate the framework on signals acquired from real industrial machinery, verifying its ability to replicate the elevated noise floors, complex vibration transmission paths, and varying boundary conditions that characterise actual factory environments.
- **Unsupervised anomaly detection:** The proposed architecture shows high potential for deployment as an unsupervised anomaly detector. Initial trials demonstrated that, when trained exclusively on healthy baseline signals, the reconstruction error effectively serves as an anomaly score to detect the presence of a fault. In this configuration, the class conditioning mechanism and CFG are omitted, meaning the shift and scale parameters of the AdaGN layers depend solely on the diffusion timestep. The difference between the reconstructed and original signals yields a residual that isolates the anomaly. Future research should formally quantify the performance of this approach against standard baselines and identify optimal diagnostic metrics to classify the fault type directly from the isolated residual.

- **Inference acceleration:** Finally, future work should investigate replacing the standard Markovian reverse process with sampling acceleration techniques. It will be necessary to assess whether adopting approaches such as Denoising Diffusion Implicit Models (DDIM), Consistency Models, or Flow Matching compromises signal quality. If high fidelity is preserved, these methods could drastically reduce inference latency during the generation phase.

Bibliography

[1] Rajendra Akerkar and Priti Sajja. *Knowledge-based systems*. Jones & Bartlett Publishers, 2009.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[3] Jerome Antoni. Fast computation of the kurtogram for the detection of transient faults. *Mechanical systems and signal processing*, 21(1):108–124, 2007.

[4] Farhat Lamia Barsha and William Eberle. An in-depth review and analysis of mode collapse in generative adversarial networks. *Machine Learning*, 114(6):141, 2025.

[5] Jan-Hendrik Bastek, WaiChing Sun, and Dennis Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations*, volume 2025, pages 3360–3385, 2025.

[6] Byambasuren Battulga, Muhammad Faizan Shaikh, Jae Wook Chun, Sung Bong Park, Sangwook Shim, and Sang Bin Lee. Mems accelerometer and hall sensor-based identification of electrical and mechanical defects in induction motors and driven systems. *IEEE sensors journal*, 24(19):31104–31113, 2024.

[7] Andreas Binder and Annette Muetze. Scaling effects of inverter-induced bearing currents in ac machines. *IEEE Transactions on Industry Applications*, 44(3):769–776, 2008.

[8] Wahyu Caesarendra and Tegoeh Tjahjowidodo. A review of feature extraction methods in vibration-based condition monitoring and its application for degradation trend estimation of low-speed slew bearing. *Machines*, 5(4):21, 2017.

[9] Case Western Reserve University Bearing Data Center. Bearing data center seeded fault test data, 2024. https://engineering.case.edu/bearingdatacenter, Last Access 2026-04-25.

[10] Ronald E Crochiere and Lawrence R Rabiner. Interpolation and decimation of digital signals—a tutorial review. *Proceedings of the IEEE*, 69(3):300–331, 2005.

[11] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[13] Levent Eren, Turker Ince, and Serkan Kiranyaz. A generic intelligent bearing fault diagnosis system using compact adaptive 1d cnn classifier. *Journal of signal processing systems*, 91(2):179–189, 2019.

[14] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[15] Shreyas Gawde, Shruti Patil, Satish Kumar, Pooja Kamat, Ketan Kotecha, and Ajith Abraham. Multi-fault diagnosis of industrial rotating machines using data-driven approach: A review of two decades of research. *Engineering Applications of Artificial Intelligence*, 123:106139, 2023.

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[17] Augustine Gray and John Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391, 2003.

[18] Zhibin Guo, Lefei Xu, Yuhao Zheng, Jingsong Xie, and Tiantian Wang. Bearing fault diagnostic framework under unknown working conditions based on condition-guided diffusion model. *Measurement*, 242:115951, 2025.

[19] Kaixu Han, Wenhao Wang, and Jun Guo. Research on a bearing fault diagnosis method based on a cnn-lstm-gru model. *Machines*, 12(12):927, 2024.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Miao He and David He. Deep learning based approach for bearing fault diagnosis. *IEEE Transactions on Industry Applications*, 53(3):3057–3065, 2017.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[24] IEA-4E Electric Motor Systems Annex (EMSA). Policy brief #9 – electric motor systems: why are they important? Technical report, International Energy Agency (IEA) Technology Collaboration Programme, December 2025. Accessed: June 2026.

[25] Turker Ince, Serkan Kiranyaz, Levent Eren, Murat Askar, and Moncef Gabbouj. Real-time motor fault detection by 1-d convolutional neural networks. *IEEE Transactions on Industrial Electronics*, 63(11):7067–7075, 2016.

[26] Sanjeet Singh Khanuja and Harmeet Kaur Khanuja. Gan challenges and optimal solutions. *Int. Res. J. Eng. Technol.(IRJET)*, 8:836–840, 2021.

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[28] A Kiran and S Saravana Kumar. A comparative analysis of gan and vae based synthetic data generators for high dimensional, imbalanced tabular data. In *2023 2nd International Conference for Innovation in Technology (INOCON)*, pages 1–6. IEEE, 2023.

[29] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021.

[30] Serkan Kiranyaz, Adel Gastli, Lazhar Ben-Brahim, Nasser Al-Emadi, and Moncef Gabbouj. Real-time fault detection and identification for mmc using

1-d convolutional neural networks. *IEEE Transactions on Industrial Electronics*, 66(11):8760–8771, 2018.

[31] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[32] Korea University Electrical Energy Conversion System Lab (EECSL). Eecsl dataset, 2024. `https://home.eecs.korea.ac.kr/publications/database`, Last Access 2026-04-25.

[33] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

[34] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM internet measurement conference*, pages 464–483, 2020.

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[37] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[38] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[39] Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga,

et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[41] Yizhen Peng, Yu Wang, and Yimin Shao. A novel bearing imbalance fault-diagnosis method based on a wasserstein conditional generative adversarial network. *Measurement*, 192:110924, 2022.

[42] John G Proakis. *Digital signal processing: principles, algorithms, and applications, 4/E*. Pearson Education India, 2007.

[43] Robert B Randall and Jerome Antoni. Rolling element bearing diagnostics—a tutorial. *Mechanical systems and signal processing*, 25(2):485–520, 2011.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[45] Syahril Ramadhan Saufi, Zair Asrar Bin Ahmad, Mohd Salman Leong, and Meng Hee Lim. Challenges and opportunities of deep learning models for machinery fault detection and diagnosis: A review. *Ieee Access*, 7:122644–122662, 2019.

[46] Siyu Shao, Pu Wang, and Ruqiang Yan. Generative adversarial networks for data augmentation in machine fault diagnosis. *Computers in Industry*, 106:85–93, 2019.

[47] Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972, 2023.

[48] Wade A Smith and Robert B Randall. Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. *Mechanical systems and signal processing*, 64:100–131, 2015.

[49] Wade A. Smith and Robert B. Randall. Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. *Mechanical Systems and Signal Processing*, 64:100–131, 2015.

[50] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[51] Yuhta Takida, Wei-Hsiang Liao, Chieh-Hsin Lai, Toshimitsu Uesaka, Shusuke Takahashi, and Yuki Mitsufuji. Preventing oversmoothing in vae via generalized variance parameterization. *Neurocomputing*, 509:137–156, 2022.

[52] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

[53] Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE, 2020.

[54] Inc. The MathWorks. Rolling element bearing fault diagnosis using deep learning, 2026. Accessed: 2026-06-26.

[55] Olav Vaag Thorsen and Magnus Dalva. A survey of faults on induction motors in offshore oil industry, petrochemical industry, gas terminals, and oil refineries. *IEEE transactions on industry applications*, 31(5):1186–1196, 1995.

[56] Parishwad P Vaidyanathan. Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial. *Proceedings of the IEEE*, 78(1):56–93, 1990.

[57] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[59] Xin Wang, Hongkai Jiang, Zhenghong Wu, and Qiao Yang. Adaptive variational autoencoding generative adversarial networks for rolling bearing fault diagnosis. *Advanced Engineering Informatics*, 56:102027, 2023.

[60] Yuan Wei, Zhijun Xiao, Xiangyan Chen, Xiaohui Gu, and Kai-Uwe Schröder. A bearing fault data augmentation method based on hybrid-diversity loss

diffusion model and parameter transfer. *Reliability Engineering & System Safety*, 253:110567, 2025.

[61] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[62] Pengcheng Xia, Yixiang Huang, Zhiyu Tao, Chengliang Liu, and Jie Liu. A digital twin-enhanced semi-supervised framework for motor fault diagnosis based on phase-contrastive current dot pattern. *Reliability Engineering & System Safety*, 235:109256, 2023.

[63] Xuefang Xu, Xu Yang, Zijian Qiao, Pengfei Liang, Changbo He, and Peiming Shi. Multi-source domain adaptation using diffusion denoising for bearing fault diagnosis under variable working conditions. *Knowledge-Based Systems*, 302:112396, 2024.

[64] Ke Yan, Jianye Su, Jing Huang, and Yuchang Mo. Chiller fault diagnosis based on vae-enabled generative adversarial networks. *IEEE Transactions on Automation Science and Engineering*, 19(1):387–395, 2020.

[65] Renhe Yao, Hongkai Jiang, Xingqiu Li, and Jiping Cao. Bearing incipient fault feature extraction using adaptive period matching enhanced sparse representation. *Mechanical Systems and Signal Processing*, 166:108467, 2022.

[66] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[67] Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024.

[68] Shen Zhang, Shibo Zhang, Bingnan Wang, and Thomas G Habetler. Deep learning algorithms for bearing fault diagnostics—a comprehensive review. *IEEE access*, 8:29857–29881, 2020.

[69] Wei Zhang, Gaoliang Peng, and Chuanhao Li. Rolling element bearings fault intelligent diagnosis based on convolutional neural networks using raw sensing signal. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the Twelfth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Nov., 21-23, 2016, Kaohsiung, Taiwan, Volume 2*, pages 77–84. Springer, 2016.

[70] Wei Zhang, Gaoliang Peng, Chuanhao Li, Yuanhang Chen, and Zhujun Zhang. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors*, 17(2):425, 2017.

[71] Ke Zhao, Feng Jia, and Haidong Shao. A novel conditional weighting transfer wasserstein auto-encoder for rolling bearing fault diagnosis with multi-source domains. *Knowledge-Based Systems*, 262:110203, 2023.

[72] Pengcheng Zhao, Wei Zhang, Xiaoshan Cao, and Xiang Li. Denoising diffusion probabilistic model-enabled data augmentation method for intelligent machine fault diagnosis. *Engineering Applications of Artificial Intelligence*, 139:109520, 2025.

[73] Xuejun Zhao, Yong Qin, Changbo He, Limin Jia, and Linlin Kou. Rolling element bearing fault diagnosis under impulsive noise environment based on cyclic correntropy spectrum. *Entropy*, 21(1):50, 2019.

[74] Taisheng Zheng, Lei Song, Jianxing Wang, Wei Teng, Xiaoli Xu, and Chao Ma. Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings. *Measurement*, 158:107741, 2020.