Shuaib Olalekan Yusuf

Explainable Deep Learning for Structural Health Monitoring in Smart City Infrastructure

A Convolutional Autoencoder Approach with SHAP Analysis

School of Technology and Innovation
Smart Grid Solutions
Smart Cities and Communities

Vaasa 2026

UNIVERSITY OF VAASA
School of Technology and Innovation
Author: Shuaib Olalekan Yusuf
Title of the thesis: Explainable Deep Learning for Structural Health Monitoring in Smart City Infrastructure: A Convolutional Autoencoder Approach with SHAP Analysis
Degree: MSc in Technology, Master's Programme in Smart Energy
Degree Programme: Smart Cities and Communities
Supervisor: As. Professor Leonidas Akritidis, Professor Christos Tjortjis
Year: 2026 **Pages:** 91

ABSTRACT:

Anomaly detection by means of deep learning models has demonstrated strong performance in vibration-based structural health monitoring; however, its widespread adoption in smart city infrastructure management is constrained by the black-box nature of the models, flagging structural anomalies without providing a physically interpretable explanation to the engineers required to act on the detected anomaly. Although recent studies that applied one-class deep learning methods to the Z24 bridge benchmark established strong detection baselines, none have provided systematic SHAP-based spatial feature attribution across all measurement setups of the Z24 Progressive Damage Test (PDT) dataset to interpret the physical meaning of unsupervised anomaly scores. This thesis demonstrates that a pipeline integrating a phase-invariant Conv1D autoencoder with a SHAP explainability framework enables reliable anomaly detection on the Z24 bridge and provides a quantifiable empirical settlement threshold below which any structural attribution is unreliable for progressive pier settlement. The autoencoder was trained on a log-PSD representation of healthy ambient vibration data from five reference channels and validated against 4 settlement scenarios out of the 17 PDT scenarios. SHAP feature attributions were computed using a validated XGBoost surrogate trained on 92 physically named spectral features, and TreeSHAP was applied to attribute each anomaly score to specific vibration characteristics. The AUROC exceeded 0.95 across all nine spatial measurement zones (mean = 0.987), and the SHAP attribution identified the inter-sensor transmissibility ratio R3V/R2V peak frequency as the dominant feature, with the lateral mid-span channel R2L accumulating the highest importance across all independently trained models, and most importantly, reliable SHAP explanations were only achievable from 80 mm settlement onward, below which surrogate fidelity collapsed. These findings reframe the objective of deep learning SHM systems from maximising detection accuracy alone to simultaneously enabling physically interpretable explanations.

KEYWORDS: Structural Health Monitoring, Explainable Artificial Intelligence, Convolutional Autoencoder, SHAP, Anomaly Detection, Z24 Bridge Benchmark, Smart City Infrastructure

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my scholarly work.

During the development of the models in this thesis, Anthropic Clause AI, integrated into the Visual Studio Code IDE, was used to augment the coding process for previously known and practised concepts, enhancing code modularity, reusability, and readability through clear annotations. The output was verified against the standard coding procedure and the documentation and white papers for each library involved, which were also mostly cited in this study. ChatGPT was also prompted to create Figure 3.6 showing the autoencoder architecture of the developed model.

The author of the thesis takes full responsibility for the content of the thesis and confirms that all analysis, interpretations, and conclusions were developed by the author

Contents

Figures

Tables

Abbreviations

AE	Autoencoder
AI	Artificial Intelligence
ANN	Artificial Neural Network
ARX	Auto-Regressive eXogenous model
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
AVT	Ambient Vibration Test / Ambient Vibration Testing
Conv1D	One-Dimensional Convolutional (layer/neural network)
DFT	Discrete Fourier Transform
EDA	Exploratory Data Analysis
EMS	Environmental Management System
EOV	Environmental and Operational Variability
F1	F1 Score (harmonic mean of Precision and Recall)
FAR	False Alarm Rate
FDD	Frequency Domain Decomposition
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FVT	Forced Vibration Test / Forced Vibration Testing
ICT	Information and Communication Technology
IDE	Integrated Development Environment
IoT	Internet of Things
LIME	Local Interpretable Model-agnostic Explanations
log-PSD	Logarithm of Power Spectral Density
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
OCSVM	One-Class Support Vector Machine

OMA	Operational Modal Analysis
PCA	Principal Component Analysis
PDT	Progressive Damage Test
PSD	Power Spectral Density
R1V	Reference Channel 1 – Vertical (Koppigen abutment side)
R2L	Reference Channel 2 – Lateral (mid-span)
R2T	Reference Channel 2 – Transverse (mid-span)
R2V	Reference Channel 2 – Vertical (mid-span)
R3V	Reference Channel 3 – Vertical (Utzenstorf abutment side)
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
RQ1	Research Question 1
RQ2	Research Question 2
RQ3	Research Question 3
SHAP	SHapley Additive exPlanations
SHM	Structural Health Monitoring
SIMCES	System Identification to Monitor Civil Engineering Structures
SSI	Stochastic Subspace Identification
SVD	Singular Value Decomposition
TCN-GAT	Temporal Convolutional Network – Graph Attention Network
TN	True Negative
TP	True Positive
TPR	True Positive Rate
VAE	Variational Autoencoder
VAE-OCSVM	Variational Autoencoder – One-Class Support Vector Machine
WSSN	Wireless Smart Sensor Networks
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

1 Introduction

This chapter presents the general overview of this thesis. It describes and justifies the problem in the context of this work, explains why some existing solutions fall short, and states precisely what this thesis does about it.

1.1 Background and Motivation

The motivation for this research can be summed up into two headings: describing the role of ICT systems in infrastructure monitoring and the problem of infrastructure aging.

1.1.1 Smart cities and the role of ICT in infrastructure monitoring

Smart city infrastructure monitoring systems leverage ICT systems with the use of connected sensors for real-time data collection, processing, and decision-making across diverse physical infrastructures, including mobility systems, energy grids, water networks, and civil structures (Sivasuriyan et al. 2026; Aktan et al. 2024; Zeng, Pang, and Tang 2024). By providing continuous insights into the condition of critical assets, these systems support urban resilience, thereby enabling the early detection of damage and deterioration, helping to prevent catastrophic failures, extending service life, and reducing maintenance costs (Aktan et al. 2024; Yu et al. 2024). Infrastructures, such as bridges, which carry traffic, connect cities, and support utility conduits, among other functions, are central to mobility, safety, and economic productivity, and their failure can ripple across the entire urban system (Salave 2025; Zinno et al. 2022). Integrating structural health monitoring (SHM) into the broader cyber-physical and digital twin platforms enhances a city's ability to manage risks, absorb shocks, and recover more quickly from climate-related hazards (Aktan et al. 2024; Malekloo et al. 2022).

Recent advances in wireless smart sensor networks (WSSN), low-power IoT nodes, and cloud/edge computing have transformed what is technically feasible in infrastructure monitoring. The use of long-life batteries, energy harvesting, and wireless communication now supports continuous, fine-grained data collection from civil structures, which

was impractical about a decade ago due to power, cabling access, and cost constraints (Noel et al. 2017; Yu et al. 2024). These ICT-enabled systems represent a transition from periodic manual and visual inspections to automated, real-time monitoring integrated into smart-city management systems (Javadinasab Hormozabad, Gutierrez Soto, and Adeli 2021; Malekloo et al. 2022). This transition has been instrumental in improving the accuracy and timeliness of condition assessment, thereby enabling infrastructures to become more self-diagnosing, adaptive, and resilient (Javadinasab Hormozabad et al. 2021). Figure 1.1 below shows the typical composition of an SHM system for infrastructure.

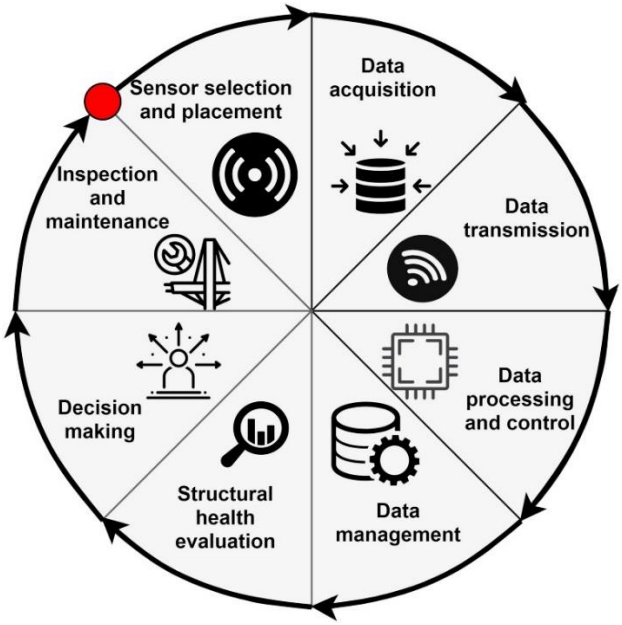


Figure 1.1: Typical components of infrastructure SHM systems (Malekloo et al. 2022)

1.1.2 The aging infrastructure problem

As previously established, civil infrastructures, such as bridges, are the backbone of modern society, but these critical assets face challenges of aging and deterioration (Benfenati et al. 2025; Favarelli and Giorgetti 2021; Jiang et al. 2023). Across Western countries, a rising number of bridges that have exceeded their design life and need special monitoring, with thousands of key bridges across France, Italy, and Germany identified as either being in critical condition or requiring restoration (Favarelli and Giorgetti 2021; Giglioni

et al. 2023). This situation prompted governments to revisit policies and technical codes governing infrastructure safety. New Italian guidelines, for example, explicitly recommend the installation of permanent SHM systems for prestressed concrete bridges that are older than 40 years, having significant spans or difficult access (Matos et al. 2023).
Traditionally, authorities have relied on routine manual inspections, in which a team of human inspectors conducts periodic visual surveys to assess the structural condition of the bridge. These conventional methods are costly, time-consuming, labour-intensive, infrequent, and limited by the inspector's expertise and resources, as well as by the ability to detect internal or gradual structural deterioration before these defects become critical to the bridge's health (Faris, Zayed, and Fares 2025; Kaewunruen et al. 2021). As infrastructure continues to age, these limitations increase the risk of undetected deterioration and catastrophic failures, which can result in significant human and economic losses (Salave 2025). For this reason, many reviewers advocate for non-destructive, continuous, data-driven evaluation, based on SHM methods, and AI-assisted imaging to bridge this gap (Faris et al. 2025; Giglioni et al. 2023; Jiang et al. 2023; Salave 2025). Early detection of degradation and structural anomalies allows for targeted interventions, reduces the need for more comprehensive repairs, optimizes maintenance costs, extends service life, and improves public safety (Benfenati et al. 2025). The collapse of the Ponte Morandi bridge in Genoa, Italy, in 2018, which tragically claimed 43 lives, illustrates the consequence of undetected deterioration and motivated the adoption of smart SHM technologies to monitor the health of the replacement bridge built in 2020 (Zinno et al. 2022).

1.2 Problem Statement

Vibration-based SHM is an effective strategy for monitoring bridges because damage essentially alters the dynamic properties and vibration response of the structure (Benfenati et al. 2025; Dabbous et al. 2024). Even though these systems are very promising, three key problems have slowed their practical deployment and effectiveness.
The first problem is detection reliability. Bridges are subjected to various natural vibrations caused by environmental factors. Temperature alone, for instance, can shift the

natural frequencies of a healthy bridge by 14–18% annually, exceeding the changes caused by moderate structural damage (Peeters and De Roeck 2001). So, distinguishing between a genuine structural anomaly and this normal environmental variation can be challenging. Deep learning approaches, such as autoencoders trained exclusively on healthy-state data that encompasses environmental variation, have shown promise in learning complex multivariate patterns in vibration signals without labelled data (Giglioni et al. 2023). However, the choice of whether to represent the input in the raw time-domain or in the frequency-domain (spectra) determines whether such models can learn stable, generalizable patterns or fail due to the variability in the signal phase across the measurement window. To elaborate, recording two time-domain measurements of the same healthy bridge might appear completely different when presented to a neural network. This is because, even though they are both vibrating at the same frequency, say 3.9 Hz (cycles per second), phase variability causes them to appear dissimilar in the time domain despite identical modal content. This causes these two structurally equivalent signals to be "*out of phase*." This phase-random problem is the reason why naïve autoencoder approaches on raw time-domain signals often fail.

The second problem is the explainability of the detection. Black-box anomaly detection using deep learning models is usually not operationally useful. This is because they provide a single anomaly score (say, "*anomaly score = 0.27*"), which a bridge maintenance engineer cannot act on. The detection can only be translated into maintenance decisions if engineers and decision-makers can understand *what* changed, *where* the change originated from (i.e., *which* sensor recorded it), and *which* known structural failure mode the change is consistent with. The absence of this kind of physically interpretable explanation makes it difficult to determine whether the model is detecting a genuine structural anomaly or merely accidental data patterns (statistically significant to the model but physically irrelevant to the bridge's health), leading to a false alarm. Throughout this thesis, the term "physically interpretable" refers to a quantity that maps directly to a measurable property of the structural system, such as modal frequency, rather than to noise or ambient excitation variability.

The third problem concerns the practical one-class constraints of the real-life bridge SHM dataset. This is because damaged bridge states are rare, as the vast majority of a bridge's lifetime is spent in healthy operation, thereby leading to class imbalance and ruling out the use of a supervised classification approach. Also, controlled-damage states are rare and expensive to produce. They are only available in benchmark datasets. Therefore, the use of anomaly detection models that are trained solely on healthy-state data is not just a methodological preference but a critical operational requirement.

This thesis addresses all of the three stated problems through an integrated pipeline applied to the Z24 bridge benchmark dataset, which is the most widely used controlled-damage dataset in SHM research.

1.3 Aim and Research Questions

The main aim of this thesis is to develop and evaluate an explainable, one-class deep learning end-to-end pipeline for vibration-based structural health monitoring of smart city infrastructure, that can reliably detect structural anomalies under the constraint of healthy-only training data, and provide physically interpretable explanations of what has changed and where the change occurred.

The work is validated using the Z24 bridge progressive damage benchmark dataset, which simulates 17 controlled damage scenarios, including different levels of pier settlements on the Koppigen pier. To achieve the stated aim, the following three research questions were addressed:

- **RQ1.** Can a Conv1D autoencoder utilizing frequency-domain of the ambient vibration data reliably detect Koppigen pier settlement across all spatial measurement zones of the Z24 benchmark?
- **RQ2.** Which vibration features most strongly explain the autoencoder's anomaly scores, and do they correspond to physically interpretable structural phenomena?
- **RQ3.** How do SHAP-attributed feature importances evolve as Koppigen pier settlement progresses from 20 mm to 95 mm, and does this progression reveal a quantifiable threshold for damage severity?

1.4 Scope of Thesis

To make the results of this work measurable and facilitate the evaluation, this thesis is limited to the following scopes:

1. This research focuses exclusively on the historical Z24 bridge benchmark dataset, and not real-time operational monitoring data.
2. From the Z24 dataset, this study focuses mainly on the Progressive Damage Test (PDT) data, which is a short-term, controlled damage data, and excludes the long-term monitoring Environmental Management System (EMS) data.
3. In the PDT data, the models utilize only the Ambient Vibration Test (AVT) mode data and exclude the Forced Vibration Test (FVT) mode data.
4. Aside from the data preprocessing and exploratory data analysis (EDA), the pipeline focuses on only five reference channels, which are present in all setups, for the anomaly detection and explainability models, and not the roving channels.
5. The post-hoc explanations rely solely on SHAP analysis, which investigates only pier settlement scenarios (S03 – S06) and not all 17 PDT scenarios.

1.5 Research Contributions

This thesis makes the following contributions to the field of structural health monitoring:

1. A phase-invariant one-class anomaly detection pipeline for bridge vibration monitoring, based on a Conv1D autoencoder trained on log-PSD representations of ambient acceleration data.
2. A feature-based SHAP explainability framework for bridge SHM that translates autoencoder reconstruction errors into physically named spectral attributions through a validated XGBoost surrogate, enabling structural engineers to identify which vibration characteristics changed and at which sensor locations.
3. An experimental characterization of the Z24 bridge PDT benchmark under a one-class constraint across all nine spatial measurement setups, providing SHAP-

based spatial analysis of pier settlement detection using only reference channel measurements.

4. A surrogate fidelity analysis across the full settlement progression of the Z24 pier damage sequence, providing a quantified lower bound on damage severity below which SHAP-based structural explanation becomes unreliable for this class of progressive pier damage.

1.6 Structure of Thesis

This thesis is organized into six chapters, each of which contributes to the development of the integrated pipeline employed in anomaly detection using a Convolutional 1D autoencoder, explainability using SHAP analysis, and model evaluations.

Chapter 2 established the theoretical foundation of each of the components of the proposed pipeline by reviewing the literature on vibration-based structural health monitoring and giving a review of SHM methods, data-driven anomaly detection, and explainable AI.

Chapter 3 describes and presents the methodology in detail, covering the dataset preparation, signal processing, autoencoder architecture and training, anomaly scoring and thresholds, and the SHAP surrogate explainability framework.

Chapter 4 presents the experimental results, which include the autoencoder training convergence, the performance of the detection model across the AVT setup zones, the attribution of SHAP features for primary benchmarking, multi-setup consistency, and the progression of the settlement severity.

Chapter 5 discusses the key findings of this thesis. It also compares the results with previous work on the Z24 Bridge dataset, addresses the research questions, and discusses the study's limitations.

Chapter 6 concludes the thesis by giving a summary of the contributions and key findings, limitations, and direction for future research.

2 Literature Review

Building on the motivation established in Chapter 1, this chapter reviews the academic literature and establishes the theoretical foundation for the concepts underpinning the methodology of this thesis. It frames the problem statement of this study within a broader academic context, reviews existing knowledge, and highlights gaps that this study aims to address.

The chapter begins with an overview of structural health monitoring principles and methods in Section 2.1, covering vibration-based methods, operational modal analysis, and environmental and operational variability. Section 2.2 reviews data-driven from classical statistical methods to modern machine learning and deep learning approaches. The section focuses on unsupervised one-class learning and its relevance for bridge monitoring. Section 2.3 examines explainable artificial intelligence, its categories, and its application in SHM. Section 2.4 presents a brief overview of the Z24 progressive damage test and provides a summary of previous work on the Z24 bridge benchmark. Section 2.5 concludes with a summary of the review and the identified gaps.

2.1 SHM Principles and Methods

Many studies have attributed different definitions to structural health monitoring (SHM), but they all sum up to the periodic or continuous measurement and analysis of a structure's response using in-situ sensing systems to assess its condition, detect and characterize damage, thereby enhancing its performance and safety-based decision-making (Malekloo et al. 2022; Plevris and Papazafeiropoulos 2024; Sivasuriyan et al. 2026). This assessment process has been widely classified into four progressive levels: (1) detection of damage existence, (2) localization of the damage's position, (3) classification and quantification of the damage type and severity, (4) evaluation of the structural risk and prognosis of the structure's remaining useful life (Malekloo et al. 2022; Peeters and De Roeck 2001; Rytter 1993; Zinno et al. 2022). Following this structured categorization, the accurate execution of higher-level assessments directly relies on the success and information gathered from the preceding level (Malekloo et al. 2022).

2.1.1 Vibration-based SHM

Among the various SHM approaches, vibration-based methods have emerged as widely adopted and robust for structural assessment. These methods are highly prominent, relying on the principle that structural damage alters the system's dynamic properties, which can be observed through variations in its vibration characteristics (Javadinasab Hormozabad et al. 2021; Malekloo et al. 2022; Zhang and Zhou 2023). When a structure is represented as a system with multiple degrees of freedom, its dynamic characteristics can be described using specific modal parameters, such as natural frequencies, modal shapes, and modal damping ratios (Rebenciuc, Bibic, and Toma 2021). These modal parameters are directly affected by structural degradation or physical damage, making them effective indicators of infrastructure health throughout its service life (Peeters and De Roeck 2001; Rebenciuc et al. 2021).

2.1.2 Operational modal analysis and frequency-domain methods

In practical field applications, vibration-based SHM relies on Operational Modal Analysis (OMA) to continuously monitor the condition of a structure using ambient vibration testing (AVT) data, thereby avoiding the need for controlled artificial excitation or forced vibration testing (FVT). This is because on large-scale civil structures like bridges, FVT is often unfeasible, costly, and disruptive to normal traffic operations (Brincker and Ventura 2015; Favarelli and Giorgetti 2021). According to Brincker (Brincker and Ventura 2015), OMA refers to the process of mapping the dynamic properties of an elastic system by identifying its unique natural vibration modes under normal operating conditions. One key aspect of this method is transforming raw time-domain acceleration signals into the frequency domain using methods such as the Power Spectral Density (PSD) function (Fang, Li, and Li 2025). PSD provides a stable representation of how the signal’s power is distributed across different frequencies, with peaks occurring at the structure's resonant frequencies (Zhang and Zhou 2023).

Through analysing these frequency-domain representations, engineers can extract important modal parameters, such as natural frequencies, modal shapes, and damping ratios, which serve as foundational damage-sensitive indicators for SHM. Equation (1) below shows the conceptual definition of PSD.

$$S_{xx}(f) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\left| X_{T(f)} \right|^2 \right] \quad (1)$$

where $S_{xx}(f)$ is the power spectral density of signal $x(t)$ at frequency f, $X_T(f)$ is the Fourier transform of $x(t)$ over a finite time window of duration T, and $\mathbb{E}[\cdot]$ denotes the expectation operator.

To extract these modal parameters from AVT data, various system identification techniques have been developed, some of which are listed below:

- **Peak-Picking:** This fundamental and intuitive frequency domain method involves identifying natural frequencies by directly selecting the most prominent peaks in the PSD plot (Brincker and Ventura 2015). The mode shapes are subsequently determined by the relative values of the transfer functions at these frequencies. As simple as the method is, it struggles with highly damped structures and poorly defined peaks, as well as closely spaced modes (Noel et al. 2017).
- **Frequency Domain Decomposition (FDD):** This method performs Singular Value Decomposition (SVD) on the spectral density matrix formed by multi-channel sensor data. By decomposing the matrix, physical response can be separated from noise, with the dominant singular values indicating the natural frequencies and the corresponding singular vectors providing the mode shape (Brincker and Ventura 2015; Noel et al. 2017; Okur, Altunişik, and Kalkan Okur 2025).
- **Stochastic Subspace Identification (SSI):** This method, unlike peak-picking and FDD, is a robust and parametric time-domain procedure. SSI constructs state-space models directly from output-only accelerometer measurements, without prior knowledge of the structure, to accurately estimate all modal parameters, including damping ratios, which are difficult to obtain via pure frequency-domain methods (Favarelli and Giorgetti 2021; Peeters and De Roeck 2001).

Despite the adoption of these modal parameter extraction techniques, they face two fundamental limitations. First, the estimates they produce are very sensitive to

Environmental and operational variability (EOV) (Pakzad and Masoodi 2025; Peeters and De Roeck 2001; Sohn 2006). Second, they often require the manual intervention of experts for post-processing (Brincker and Ventura 2015; Yang and Huang 2024). Peak-picking is subjective and prone to human error, and advanced SSI can also generate noise modes alongside the true physical mode, requiring an expert to analyse and select the correct parameters (Favarelli and Giorgetti 2021). These limitations motivate the adoption of data-driven deep learning methods in modern SHM (Yang and Huang 2024).

2.1.3 Environmental and operational variability (EOV)

EOV is considered one of the primary obstacles to the deployment of SHM systems (Malekloo et al. 2022; Sohn 2006; Zinno et al. 2022). This is because in practical, real-world environments, structures are subjected to continuously varying conditions that change their dynamic response. These ambient variations can easily mask subtle changes caused by actual structural damage in the system. Temperature, for instance, is one of the dominant environmental factors affecting a bridge's dynamic properties (Peeters and De Roeck 2001; Sohn 2006). Peeters and De Roeck (2001) observed that the normal environmental temperature changes caused a shift by about 14 – 18% in the natural frequencies of the Z24 bridge annually, which are much larger than the frequency changes induced by a typical early-stage structural damage (3 – 8%). A bilinear relationship was also reported between temperature and natural frequencies, allowing frozen asphalt to contribute to the bridge's stiffness and to increase its natural frequencies at temperatures below 0°C, but not at warmer temperatures above 0°C (Peeters and De Roeck 2001; Reynders and Roeck 2008; Sohn 2006).

On the other hand, operational variables, such as traffic loading effects, also introduce extra mass to the bridge's deck, thereby causing a measurable decrease in the structure's natural frequencies and higher damping ratios. These introduce a layer of non-stationary variability into the monitoring data (Sohn 2006). A successful SHM system is therefore required to separate changes related to actual damage from those induced by these different EOVs. This can be achieved by explicitly modelling or removing the EOV using various methods, including data normalization, environmental clustering, statistical pattern

recognition and machine learning approaches to enhance the system's reliability and prevent triggering false alarms (Eltouny, Gomaa, and Liang 2023; Hassan Daneshvar and Sarmadi 2022; Pakzad and Masoodi 2025; Sohn 2006).

While many studies have extensively addressed the use of EOVs in long-term monitoring, short-term PDT measurement conducted within a narrow seasonal window can suppress seasonal variations and enable a controlled damage-detection study without interference from environmental variables.

2.2 Data-Driven Anomaly Detection for SHM

From the viewpoint of SHM, anomaly detection is a foundational data-driven framework employed in identifying patterns in the structural response of a system that do not conform to expected normal behaviour (An and Cho 2015; Chandola, Banerjee, and Kumar 2009). The primary objective of this framework is to process data from continuous monitoring systems to uncover structural defects and track changes that may indicate damage (Hassan Daneshvar and Sarmadi 2022). Since real-world deployment of SHM is constrained by the lack of labelled data for damaged structures, data-driven anomaly detection in SHM must operate under a "one-class constraint," relying on unsupervised machine learning methods trained on healthy baseline data to establish a boundary of normality (Nesackon Abraham et al. 2026).

2.2.1 Classical statistical methods to machine learning

Early classical approaches for anomaly detection in SHM relied on the use of statistical distance measures, such as Mahalanobis distance, or fitting simple probability distributions to baseline data (Eltouny et al. 2023; Hassan Daneshvar and Sarmadi 2022). But over time, more advanced unsupervised machine learning approaches were adopted to handle the complexities of multi-sensor data, with algorithms such as k-means clustering, Principal Component Analysis (PCA), and Local Outlier Factor (LOF) providing enhanced mechanisms for feature extraction, dimensionality reduction, and density estimation (Bayane, Leander, and Karoumi 2024; Hassan Daneshvar and Sarmadi 2022).

They, however, exhibit notable limitations when applied to high-dimensional and highly variable data generated by modern continuous monitoring systems. One such challenge is the influence of EOVs. Classical PCA are often restricted by their linearity assumptions and struggles with the non-linear interactions and distortions introduced by EOVs (Nesackon Abraham et al. 2026). Additionally, one-class classifiers such as OCSVM and traditional distance-based classifiers often lack the capacity to distinguish environmental noise from actual damage. Their success often relies on manually engineered features, making them less robust and highly prone to false alarms in practical deployment (Pakzad and Masoodi 2025). Deep learning approaches have piqued researchers' interest and show promise for overcoming the limitations of classical methods.

2.2.2 Deep learning for SHM

Early applications of deep learning used Multilayer Perceptrons (MLPs) and standard Artificial Neural Networks (ANNs). However, the field quickly adopted Convolutional Neural Networks (CNNs) to handle the complex, high-dimensional data typical of civil infrastructure, with 1D and 2D CNNs proving very effective at processing vibration signals and time-frequency representations, autonomously learning spatial and damage-sensitive features without manual engineering (Santaniello and Russo 2023). Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, were subsequently integrated into SHM pipelines because continuous monitoring of structures produced time-series data. The specialized gating mechanism of LSTMs equipped them with the ability to learn long-term temporal dependencies in structural responses, mitigating the vanishing gradient problem common in standard RNNs (Santos-Vila et al. 2024).
In recent studies, Foundation Models and Transformers, which use self-attention mechanisms to learn highly generalizable representations across diverse datasets, have been introduced to SHM (Benfenati et al. 2025). Additionally, other hybrid architectures, such as Temporal Convolutional Networks combined with Graph Attention Networks (TCN-GAT), designed to model both the temporal dynamics of vibration signals and the spatial topological relationships of complex multi-sensor networks, have emerged in the field (Ni, Jin, and Hu 2025).

2.2.3 Reconstruction-based anomaly detection

Despite the success of various supervised deep learning classifiers, a shortage of labelled data for damaged structural states has driven the adoption of unsupervised, reconstruction-based anomaly detection frameworks (Giglioni et al. 2023; Santos-Vila et al. 2024). This approach relies on the autoencoder, which is a concept pioneered by Hinton and Salakhutdinov as a nonlinear generalization of PCA for dimensionality reduction (Hinton and Salakhutdinov 2006).

An autoencoder is a specific type of neural network that is trained to attempt to copy (reconstruct) its input data to its output (Goodfellow, Bengio, and Courville 2016). It does that by compressing a high-dimensional input feature set into a lower-dimensional latent representation. Structurally, it consists of an encoder layer, a bottleneck (hidden or code) layer, and a decoder layer. The primary goal of an autoencoder is to minimize reconstruction error, which is the difference between the original input and the reconstructed output. By forcing the data to pass through the bottleneck layer, the network autonomously learns to preserve the most essential and meaningful features in the data (Giglioni et al. 2023; Okur et al. 2025). Figure 2.1 below illustrates an autoencoder architecture.

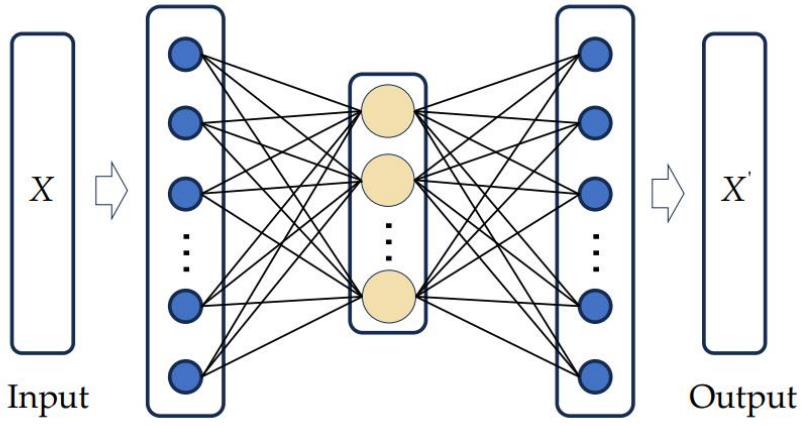


Figure 2.1: Architecture of an autoencoder (Ni et al. 2025)

In this reconstruction-based anomaly detection framework, the autoencoder is trained exclusively on healthy structural data during the normal operational state. Since the objective of the training is to minimize the reconstruction loss, it forces the network to

learn the underlying patterns and dynamic characteristics of the undamaged structure (Giglioni et al. 2023; Santos-Vila et al. 2024). During testing and validation, or the monitoring phase, new sensor data is fed into the trained model. The autoencoder will successfully reconstruct the input with minimal error if the structure remains healthy, but if structural damage has occurred, the incoming data will deviate from the learned healthy distribution, causing the decoder to fail and resulting in a significantly higher reconstruction error (Favarelli and Giorgetti 2021; Okur et al. 2025). This reconstruction error serves as a robust, damage-sensitive feature that triggers anomalies when it surpasses a statistically defined threshold (Giglioni et al. 2023).

2.3 Explainable Artificial Intelligence

The adoption of machine learning and deep learning methods, such as the previously discussed autoencoder, has enhanced predictive modelling across many disciplines. The exceptional predictive capabilities of these models, however, come at the cost of transparency, given the "*black box*" nature of their architecture. Their internal decision-making logic is obscured by millions of parameters and nonlinear transformations, making it incomprehensible to a human operator how they mathematically map inputs to outputs (Adadi and Berrada 2018; Tang et al. 2025). The emergence of Explainable AI (XAI) as a research field aims to bridge this transparency gap by developing interpretability tools that demystify black-box algorithms without sacrificing model predictive accuracy, thereby enhancing trust, accountability, and safety in the decision-making process (Adadi and Berrada 2018).

2.3.1 Categories of XAI methods

The different methodologies of XAI can be broadly classified across three main categories:

I. **Model-specific vs. Model-agnostic:** Model-specific methods are restricted only to certain algorithmic architectures. For instance, gradient-based methods like GradCAM or Layer-Wise Relevance Propagation are exclusively designed to

interrogate the internal activations and layers of neural networks. While model-agnostic methods separate the explanation from the prediction model and can be applied to any machine learning algorithm (Adadi and Berrada 2018; Lundberg and Lee 2017).

II. **Local vs. Global:** The scope of interpretability dictates the focus of the explanation. Local methods explain individual, specific predictions to justify why a model made a particular decision for a single instance. Global methods, on the other hand, aim to explain the model's overall logic and general behaviour across an entire population (Adadi and Berrada 2018).

III. **Ante-hoc vs. Post-hoc:** Ante-hoc (or intrinsic) methods build interpretability directly into the model's architecture by design, using algorithms like decision trees or linear regression. While post-hoc methods add an explanatory layer to an already-trained "*black-box*" model (Adadi and Berrada 2018).

2.3.2 Theoretical foundations of SHAP

Shapley Additive exPlanation (SHAP) has a theoretical foundation in classical cooperative game theory. This model-agnostic framework involves the input features of a machine learning model acting as "players" in a game, forming a "coalition" to achieve the model's final prediction ("*the payout*"). The main objective of this framework is to fairly distribute the prediction's outcome among the features by calculating their marginal contribution to the model's prediction (Lundberg and Lee 2017). Equation (1) below shows the attribution for each feature, determined by the Shapley value formula.

$$\phi_i(f) = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!\ (|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} [f(S \cup \{i\}) - f(S)] \qquad (2)$$

where $\phi_i(f)$ is the Shapley value (SHAP attribution) assigned to feature i, $\mathcal{F}$ is the full set of features, S is a subset of features that does not include feature i, $|S|$ is the number of features in subset S, $|\mathcal{F}|$ is the total number of features, $f(S)$ is the model prediction using only the features in S, $f(S \cup \{i\})$ is the model prediction when feature i is added to subset S, and $\frac{|S|!\ (|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!}$ is a weighting factor that accounts for all possible

orderings in which feature i could be introduced into the coalition. This equation calculated a weighted average of all possible differences in the model's output when feature i is included versus when it is not (Lundberg and Lee 2017).

There are four axioms satisfied by SHAP, based on game theory, that mathematically guarantee a unique and fair distribution of feature attributions:

1. **Efficiency (Local Accuracy):** This guarantees that the feature attributions sum exactly to the difference between the model prediction and the baseline expected prediction.

$$\sum_{i=1}^{|\mathcal{F}|} \phi_i = f(x) - \mathbb{E}[f(x)] \quad (3)$$

2. **Symmetry:** This ensures that if two features contribute equally to all possible coalitions, they receive the exact same attribution value.
3. **Nullity (Missingness):** This guarantees that if a feature is missing or has no effect on the model prediction in any context, it is assigned an attribution of zero ($\phi_i = 0$).
4. **Additivity:** This ensures that if a model can be expressed as the sum of two value functions f and g, the Shapley attribution for feature i in the combined model equals the sum of its attributions in each individual model separately. That is:

$$\phi_i(f + g) = \phi_i(f) + \phi_i(g) \quad (4)$$

2.3.2.1 KernelSHAP and TreeSHAP

Since calculating exact Shapley values requires training a model on all possible combinations of features, which can be computationally expensive for datasets with many variables, KernelSHAP and TreeSHAP serve as approximation methods to address this issue.

KernelSHAP is used to estimate the feature importance for any machine learning model by sampling data combinations using weighted linear regression, while TreeSHAP, specialized and highly efficient for tree-based models(such as Random Forests or XGBoost), exploits the internal structure of trees and computes the exact Shapley values

deterministically in a fraction of the polynomial time, thereby avoiding slow approximations unlike KernelSHAP (Lundberg and Lee 2017).

2.3.2.2 LIME, Integrated Gradients, DeepLIFT, as an alternative to SHAP

A very common post-hoc alternative to SHAP is LIME (Local Interpretable Model-agnostic Explanations). LIME explains individual predictions by learning a sparse, interpretable linear model locally around the prediction of interest (Ribeiro, Singh, and Guestrin 2016). However, it relies heavily on heuristically chosen parameters (such as its loss function and weighting kernel), resulting in approximate local explanations with no theoretical consistency guarantees. For this reason, LIME can violate local accuracy and consistency, sometimes leading to unintuitive behaviours (Lundberg and Lee 2017).
Other approaches, like Integrated Gradients and DeepLIFT, attribute feature importance by comparing the activation of neurons against a baseline reference value, then recursively passing multipliers backwards through a network to determine the effect of each input. However, grounded in classical cooperative game theory, SHAP unifies six existing feature attribution methods, including LIME and DeepLIFT, under a single framework, being the most rigorous and robust choice for interpreting complex machine learning models (Lundberg and Lee 2017).

2.3.3 Surrogate models for explainability

Although SHAP provides robust feature attribution, explaining the internal logic of a complex model on raw input spaces often yields attributions that are uninterpretable to humans. In such a scenario, a surrogate is needed. A surrogate model is a simpler, interpretable model (such as linear regression or a decision tree) that is trained to approximate the prediction of a black-box model (Adadi and Berrada 2018). It is often needed when the input space of the primary model is complex or high-dimensional. A surrogate learns to approximate the black-box model's outputs from a more human-comprehensible, lower-dimensional set of features (Adadi and Berrada 2018; Ribeiro et al. 2016). For a surrogate model to be considered valid and reliable, it must satisfy the fidelity criterion.

This implies that the surrogate must accurately represent the black-box model's behaviour (Ribeiro et al. 2016).

2.3.4 XAI in structural health monitoring

SHAP has been applied successfully in recent SHM literature to interpret complex supervised learning models. For instance, Chen et al. (Chen et al. 2024) successfully combine the highly scalable XGBoost with SHAP to accurately predict and explain the load-deformation correlation of a long-span suspension bridge. Using SHAP values, they explicitly quantify the contributions of temperature, wind, and traffic load to bridge deflection and demonstrate that temperature dominates the structural deformation. However, integrating SHAP and other XAI frameworks into unsupervised one-class anomaly detection models remains underexplored. Addressing this gap is key to the future of smart city infrastructure, as it enhances engineers' trust in an autonomous system that flags anomalies without relying on prior damage states.

2.4 The Z24 Bridge Benchmark

The Z24 bridge was a post-tensioned two-cell concrete box-girder highway overpass in Switzerland, demolished in 1998 to accommodate a new railway. Before its demolition, the bridge served as the primary test subject for the European Brite-EuRam project SIMCES, which conducted a full-scale monitoring campaign (Maeck and De Roeck 2003; Reynders and Roeck 2008). The measurement campaign was divided into two main phases, consisting of about a year of continuous environmental monitoring to capture the effects of seasonal variations (especially freezing temperatures), followed by a short-term PDT campaign lasting for about a month (Peeters and De Roeck 2001; Reynders and Roeck 2008). A more detailed description of the measurement campaign is given in Section 3.1.

2.4.1 Significance as a benchmark

The Z24 bridge dataset is a standard benchmark for SHM, widely recognised as the gold standard for different SHM methods. Many traditional SHM models often perform well on simulated data and control laboratory experiments, but they most frequently fail in real-life deployment due to EOVs (Dabbous et al. 2024). This is because it is impractical and economically inefficient to intentionally damage operational civil infrastructure, such as a bridge, making the Z24 campaign very rare (Buckley, Ghosh, and Pakrashi 2023). The dataset remains publicly available through KU Leuven and continues to be used extensively for validating emerging techniques for anomaly detection and SHM (Z24 Bridge benchmark n.d.).

2.4.2 Key prior methods applied

Historically, the analyses of the Z24 dataset relied on OMA, with early studies utilizing SSI and ARX to extract modal parameters and remove EOVs (Maeck and De Roeck 2003; Peeters and De Roeck 2001). As the SHM field evolved, the dataset remained a testbed for the different data-driven feature selection and extraction frameworks (Buckley et al. 2023). Recent studies have applied a range of supervised, multi-class, and one-class learning approaches to this dataset, including autoencoder-based reconstruction methods, hybrid VAE frameworks, TCN-GAT architectures, edge-deployed convolutional classifiers, zero-shot streaming LSTM networks, information-based density anomaly detectors with semi-parametric extreme value theory, and ensemble fusion frameworks combining linear PCA and non-linear autoencoders (Dabbous et al. 2024; Giglioni et al. 2023; Hassan Daneshvar and Sarmadi 2022; Mehrjoo et al. 2025; Nesackon Abraham et al. 2026; Ni et al. 2025; Pakzad and Masoodi 2025). All these studies establish strong detection baselines but do not provide post-hoc feature-level attribution of the anomaly scores, which is the gap this thesis addresses.

2.5 Summary and Research Gap

Despite the good progress made across different studies applying deep learning to the Z24 bridge benchmark, existing data-driven AI studies predominantly exhibit limitations in interpretability (Pakzad and Masoodi 2025). The different one-class models and auto-encoders can successfully detect anomalies under severe EOVs, but they operate as black boxes without explaining the physical or structural reasoning behind their outputs (Tang et al. 2025). To the best of the author’s knowledge, no prior study has systematically provided a SHAP-based spatial feature attribution across all measurement setups of the Z24 PDT dataset to interpret the physical meaning of unsupervised anomaly scores.

This thesis addresses this gap by integrating SHAP analysis into an unsupervised Conv1D autoencoder framework. Specifically, this research aims to determine whether a Conv1D autoencoder utilizing frequency-domain ambient vibration data can reliably detect the Koppigen pier settlement across all spatial measurement zones of the Z24 benchmark (RQ1). Additionally, it investigates which vibration features most strongly explain the autoencoder's anomaly scores using XAI and verifies their physical interpretability (RQ2). Finally, the study examines how SHAP-attributed feature importances evolve as the Koppigen pier settlement progresses from 20 mm to 95 mm to establish a quantifiable threshold for damage severity (RQ3).

3 Methodology

This chapter presents the methodology used to build the integrated pipeline of the proposed framework, with each one building on the previous one, forming a logical chain that links the raw data to the interpretable explanation. The chapter is organized as follows: Section 3.1 gives a description of the dataset and preprocessing decisions; Section 3.2 explains the signal processing in detail; Section 3.3 presents the details of the autoencoder architecture and training; Section 3.4 defines and outlines the anomaly scoring and the model evaluation protocol; and finally, Section 3.5 gives a description of the SHAP explainability pipeline. Figure 3.1 below provides an overview of the entire methodology.

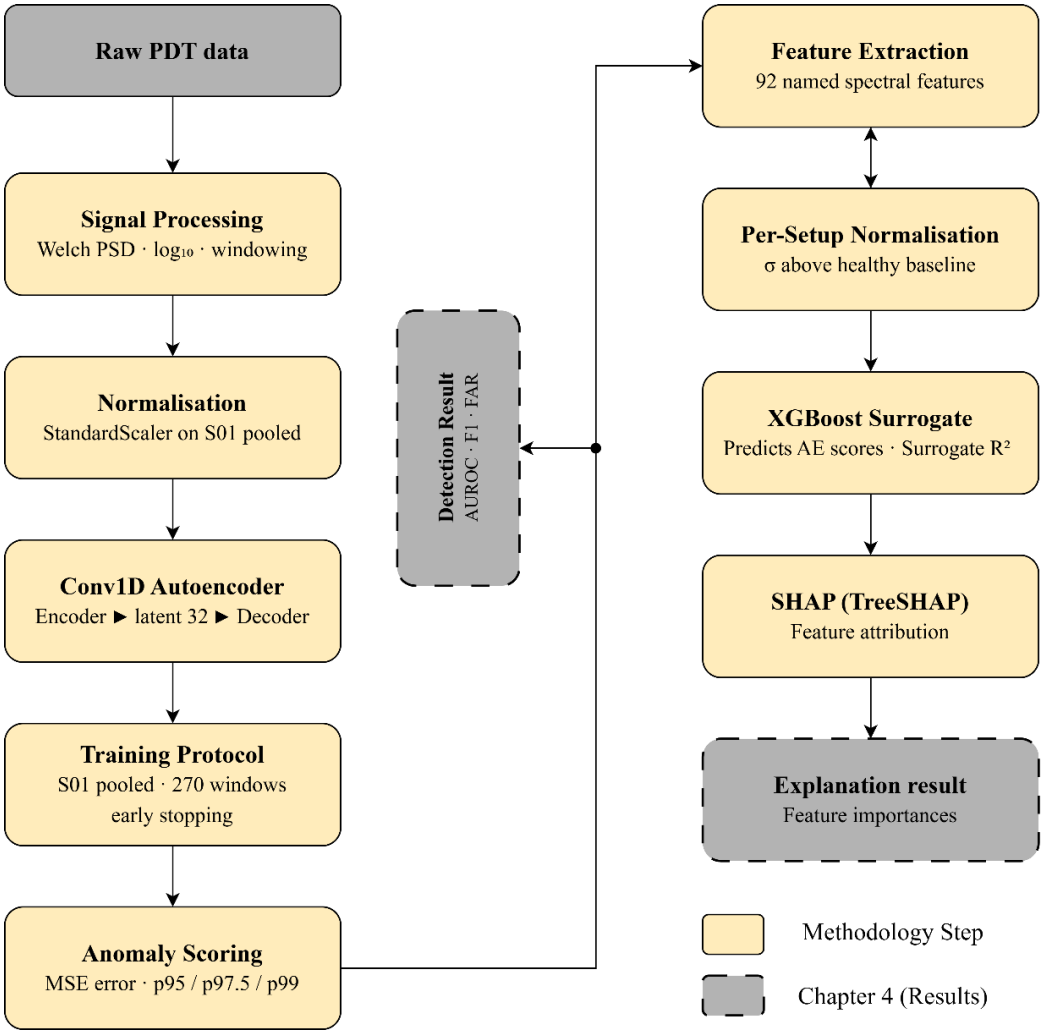


Figure 3.1: Thesis methodology flowchart

3.1 Dataset Description

The Z24 bridge benchmark dataset used for this study was obtained from the KU Leuven Structural Mechanics Section (Z24 Bridge benchmark n.d.), under the terms of their non-commercial research agreement. The description of the bridge, the System Identification to Monitor Civil Engineering Structures (SIMCES) project, and the different PDT scenarios are presented in the following subsections.

3.1.1 The Z24 Bridge and the SIMCES Campaign

The Z24 bridge was located in the canton of Bern near Solothurn, Switzerland, connecting the villages of Koppigen and Utzenstorf and crossing the A1 highway between Bern and Zurich. This bridge was a prestressed concrete highway bridge that served as a full-scale testbed for vibration-based damage detection. In order to study the effects of stiffness and cracking on eigenfrequencies (natural frequencies), damping, and mode shapes, the bridge was progressively damaged and monitored (Maeck and De Roeck 2003). Because it is a real bridge with controlled, documented damage scenarios and multiple full-modal measurements before and after each damage step, it has become a widely used benchmark for validating SHM and anomaly-detection methods (Krämer, De Smet, and De Roeck 1999; Maeck and De Roeck 2003). The dataset covers both undamaged and various damaged states, making it ideal for assessing the vibration-based techniques in realistic conditions.

The tests were conducted within the EU Brite-EuRam SIMCES project (BE-3157), and the Swiss Federal Office of Roads sponsored a comprehensive test program on Z24 between 1997 and 1998. There are two main measurement campaigns involved in the project. The first is a long-term environmental monitoring system (EMS) that recorded the influence of environmental variables (such as temperature, wind, and rain) on the bridge's dynamic characteristics. The EMS demonstrated that temperature and other environmental factors strongly affect the bridge's natural frequencies and complicate damage detection (Peeters and De Roeck 2001). The second measurement campaign is a series of Progressive Damage Tests (PDT), which subjected the bridge to 17 incrementally

severe damage scenarios while vibration measurements were recorded across nine measurement setups. Within each setup, measurements were collected in two modes:

1. **Ambient Vibration Testing (AVT):** In this mode, the bridge's natural response to ambient excitation from traffic and wind was recorded passively without any artificial input.
2. **Forced Vibration Testing (FVT):** In this mode, two shakers were mounted on the bridge deck, and controlled harmonic excitation was imposed to produce a more deterministic structural response.

This thesis focused on the PDT campaign, as it provides measurements of controlled damage evolution under consistent summer conditions. This choice of PDT over EMS is justified because it allows changes in dynamic response to be attributed mainly to damage rather than to seasonal environmental variability, thereby avoiding the need for complex environmental normalization while still using a full-scale bridge benchmark (Maeck and De Roeck 2003; Peeters and De Roeck 2001). Figure 3.2 below presents the elevation and plan drawings of the Z24 bridge, showing the Koppigen pier (the damaged/settled pier) and reference channel locations.

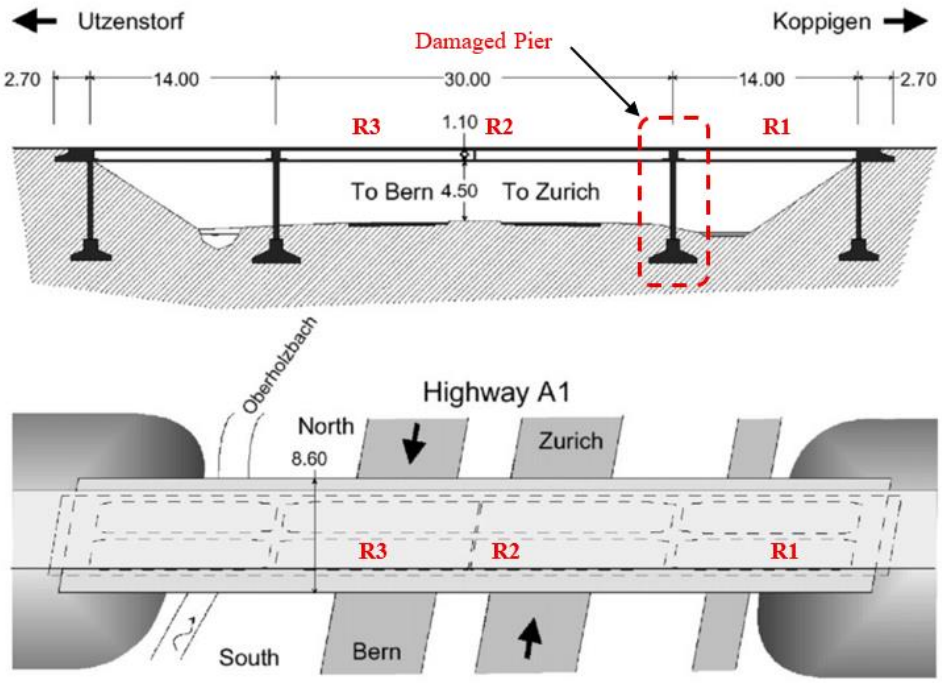


Figure 3.2: A schematic drawing of the Z24 bridge showing reference channels and the damaged pier. Adapted from **(Peeters and De Roeck 2001)**

3.1.2 PDT Scenarios

The PDT campaign ran from 4 August to 9 September 1998, spanning 36 days. Each scenario was surveyed using a roving grid of nine measurement setups of accelerometer arrays, as shown in Figure 3.3. Table 3.1 below summarises all 17 scenarios with their measurement dates, structural conditions, and classification.

Scenario 01 serves as the undamaged, healthy baseline against which all anomaly detection is performed. Scenarios 03 to 06 represent the settlement progression, which is the primary focus of this thesis. To simulate progressive foundation settlement, the Koppigen pier was incrementally lowered by 20 mm, 40 mm, 80 mm, and 95 mm, respectively.

Scenario 05 (80 mm Koppigen pier settlement), measured on 17 August 1998, is adopted as the primary benchmark scenario throughout this work. This scenario was selected as the primary benchmark because SHM aims to detect damage reliably at moderate severity levels, not only at the structural limit. The 80 mm case presents a more realistic and challenging detection scenario than the maximum settlement (95 mm), and it is the benchmark adopted in recent Z24 studies, such as (Dabbous et al. 2024), (Buckley et al. 2023), and (Abdrabo 2024). This choice is also consistent with the work of Peeters & De Roeck (Peeters and De Roeck 2001), which demonstrated that only 80mm settlement produced a statistically significant deviation in the first bending mode after removing environmental effects. Earlier settlement levels (20 mm and 40 mm) are still within the 95% confidence interval bounds, and 80 mm is the first operationally meaningful damage state.

Scenario 08, measured on 20 August 1998 after the structural restoration from the settlement cycle, serves as a second healthy reference for calibrating the threshold. This is discussed further in Section 3.1.3. Scenarios 02 and 07 represent transitional states which are associated with the installation, adjustment, and removal of the settlement apparatus. Because their classifications didn’t reflect a healthy or damaged state, they are excluded from damage detection modelling. Scenarios 09 to 17 cover a variety of additional damage types, including concrete spalling, anchor head failures, and post-

tensioning tendon ruptures, providing more experimental context. They are not included in the scope of the SHAP analysis presented in Chapter 4.

Table 3.1: PDT scenario summary

Scenario	Date (1998)	Damage description	Classification
S01	4 Aug	Undamaged baseline	Healthy
S02	9 Aug	Settlement system installed	Transitional
S03	10 Aug	Koppigen pier settlement 20 mm	Damaged
S04	12 Aug	Koppigen pier settlement 40 mm	Damaged
S05	17 Aug	Koppigen pier settlement 80 mm	Damaged
S06	18 Aug	Koppigen pier settlement 95 mm	Damaged
S07	19 Aug	Pier lifted and tilted	Transitional
S08	20 Aug	Structural restoration (new ref.)	Healthy
S09	25 Aug	Concrete spalling 12 m²	Damaged
S10	26 Aug	Concrete spalling 24 m²	Damaged
S11	27 Aug	Landslide at abutment	Damaged
S12	31 Aug	Concrete hinge introduced	Damaged
S13	2 Sep	2 of 16 anchor heads failed	Damaged
S14	3 Sep	4 of 16 anchor heads failed	Damaged
S15	7 Sep	2 of 16 tendons ruptured	Damaged
S16	8 Sep	4 of 16 tendons ruptured	Damaged
S17	9 Sep	6 of 16 tendons ruptured	Damaged

This thesis uses only AVT data throughout. The reason for this choice is that AVT is a good representation of how a permanently deployed monitoring system would function, because it monitors the bridge's response to ambient conditions and does not require any active excitation hardware. FVT recordings, on the other hand, include the shaker driving-point channels (DP1V, DP2V) whose amplitudes are dominated by the imposed force rather than the bridge's structural response. Additionally, the Power Spectral Density (PSD) input representation adopted in this work, which is described in Section 3.2, is

most physically meaningful for ambient response data. Figure 3.3 below shows the different measurement setups on the bridge's deck and the roving accelerometer grid.

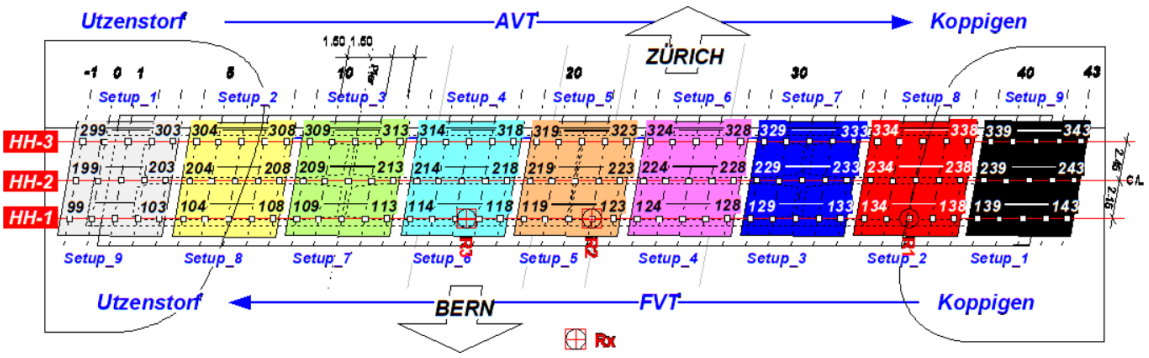


Figure 3.3: Plan view of the Z24 bridge showing different setups and roving grid **(Mehrjoo et al. 2025)**.

Across all nine setups on the deck and all 17 scenarios, five channels remain physically fixed in position throughout the entire campaign. These sensors were not relocated between setups and are referred to as the reference channels (Abdrabo 2024). Their positions are shown in Figure 3.2 and Figure 3.3 above. The fact that they are present in every measurement file makes spatial comparison possible across the roving grid. The five reference channels are listed below:

1. **R3V**: vertical accelerometer, positioned on the deck just after the Utzenstorf pier into the main span
2. **R2L**: longitudinal accelerometer at mid main-span, oriented parallel to bridge axis
3. **R2T**: transverse accelerometer at mid main-span, oriented perpendicular to bridge axis
4. **R2V**: vertical accelerometer at the middle of the main span
5. **R1V**: vertical accelerometer in the Koppigen half of the deck

The R2 cluster (R2L, R2T, R2V) occupies the same mid-span cross-section and therefore captures the multidirectional vibration response at the mid-span. These five channels constitute the complete input feature set for both the autoencoder and the SHAP

explainability analysis. All other, roving deck channels and pier channels are excluded from modelling.

3.1.3 Dataset Observations

During exploratory data analysis (EDA), the AVT measurement file for Scenario 01, Setup 07 (S01setup07), was found to contain 32 accelerometer channels, instead of the 33 found in the other Scenario 01 setups. But since all five reference channels (R1V, R2L, R2T, R2V, R3V) are present and complete in this file, and the modelling pipeline for this thesis uses only the reference channel subset, this missing data anomaly has no effect on the training data, the anomaly scoring, or the SHAP analysis.

Additionally, Scenario 08, which is the restored condition after the series of settlements, was used alongside Scenario 01 when computing the anomaly detection thresholds described in Section 3.4. To justify this use, the spectral similarity between Scenario 01 and Scenario 08 was verified during the EDA. As shown in Figure A.1 (see Appendix), the Mode 1 peak frequency of R1V, R2T, R2V, and R3V is identical in both scenarios. Channel R2L, on the other hand, shows a Mode 1 peak frequency offset of about 0.5 Hz between the two scenarios, which could be attributed to the temperature difference between 4 August (undamaged or healthy state) and 20 August (restored state) rather than any residual structural change. This is because it falls within the 14–18% annual seasonal range that was documented by Peeters & De Roeck (Peeters and De Roeck 2001). For this reason, Scenario 08 is treated as structurally equivalent to Scenario 01 for the purpose of threshold calibration.

3.2 Signal Preprocessing and Processing

For the AVT data, measurements are recorded for each setup in the different scenarios. Each file contains approximately 65,530 samples per channel at a sampling frequency of 100 Hz, which represents about 655 seconds (10.9 minutes) of continuous ambient vibration recording. To avoid errors during processing, the actual shape of each array was

read directly from its file when loading, rather than assuming a fixed sample count. The data processing sequence is described in the subsections below.

3.2.1 Time Domain Acceleration Data to Frequency Domain

As mentioned in the problem statement in Section 1.2, working directly with raw time-domain signals for ambient vibration, which is driven by different stochastic (random) excitations, is unsuitable for training a naïve autoencoder anomaly detector. Additionally, many studies, including (Abdrabo 2024), demonstrated that frequency-domain features outperform time-domain features for Z24 anomaly detection. For these reasons, this work is to represent each recording in the frequency domain using the Power Spectral Density (PSD). The PSD was estimated using Welch's method, which divides the given recording into overlapping segments, computes the discrete Fourier transform of each segment, and averages the squared magnitudes across segments. By averaging, the variance of the spectral estimate is reduced compared to when a single Fourier transform is applied to the full recording (Welch 1967). A segment length of $nperseg = 1024$ samples was used throughout, yielding a frequency resolution of 100 Hz/1024 = 0.098 Hz (approximately 0.1 Hz). This resolution is sufficient to distinguish the four Z24 natural frequencies at approximately 3.9 Hz, 4.9 Hz, 9.8 Hz, and 10.5 Hz, which were confirmed in the EDA phase and are consistent with articles on the Z24 bridge, especially (Peeters and De Roeck 2001). Equations (5) and (6) show the formula for the discrete Fourier transform (DFT), for converting a time-domain quantity to the frequency domain, and the Welch PSD estimator, respectively.

$$X[k] = \sum_{n=0}^{N-1} x[n]\ e^{-j2\pi kn/N}, k = 0,1, \dots, N-1 \qquad (5)$$

where $x[n]$ is the n-th discrete time-domain sample, N is the total number of samples in the window, $k$is the frequency bin index, $X[k]$is the complex-valued frequency-domain representation at bin k, and $j = \sqrt{-1}$ is the imaginary unit.

$$\hat{S}_{xx}(f_k) = \frac{1}{K \cdot U \cdot f_s} \sum_{i=0}^{K-1} \left| \sum_{n=0}^{L-1} x_{i[n]} \cdot w[n] \cdot e^{-j\frac{2\pi kn}{L}} \right|^2 \qquad (6)$$

where $x_i[n]$ is the n-th sample of the i-th overlapping segment, $L = 1024$ is the segment length ($nperseg$), K is the number of segments, $w[n]$ is the Hann window function, $U = \frac{1}{L}\sum_{n=0}^{L-1} w^2\,[n]$is the window power normalisation factor, $f_s = 100$ Hz is the sampling frequency, and $f_k = k \cdot f_s/L$ is the physical frequency in Hz corresponding to bin k. In this thesis, only bins corresponding to $f_k \in [0{,}30]$ Hz are retained, yielding 308 frequency bins per channel (Welch 1967).

The PSD is then transformed using $log_{10}\,(PSD + 10^{-30}\,)$. Without applying the log transform to the PSD, the vibration energy values range from 0.000001 to 1.0 (6 orders of magnitude), but with the log transform, the values range from -6 to 0, which is much more manageable, and also makes the distribution more Gaussian (bell-shaped), which deep learning models find easier to work with. The addition of the 10^{-30} constant prevents log (0) = – ∞ (negative infinity) when the PSD is exactly zero. This combined representation is hereafter referred to as the *log-PSD*. Figure 3.4 below illustrates the phase problem with a display range limited to 0–12 Hz.

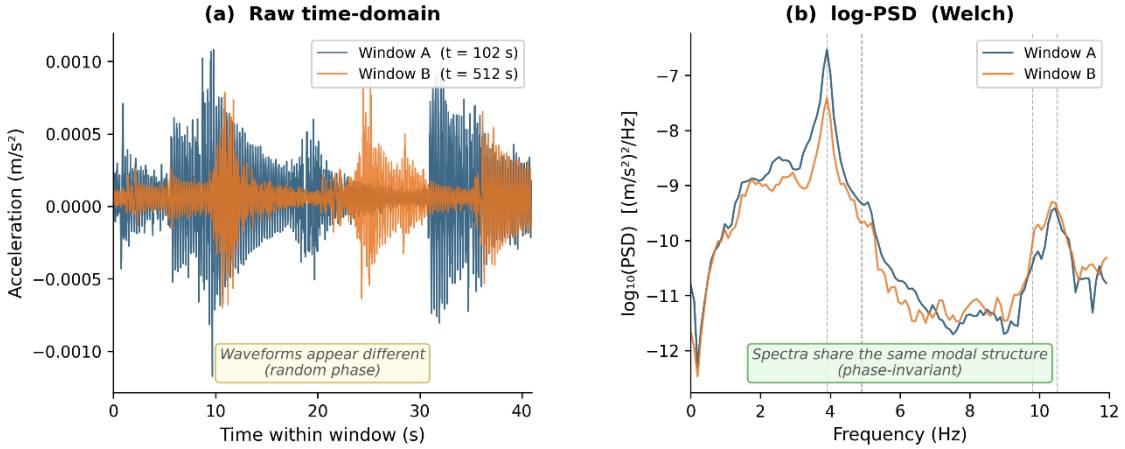


Figure 3.4: Illustration of the phase problem and its resolution

3.2.2 Windowing

A single recording is divided into shorter overlapping segments called windows, and each window is processed rather than the entire 10.9-minute recording. Doing this serves two purposes: first, it increases the number of training examples per recording, and second, it ensures that the PSD estimate within each window reflects an approximately

stationary structural response. A window length of 4,096 samples (40.96 seconds at 100 Hz) was selected. Since the lowest modal frequency of the Z24 bridge is at 3.9 Hz, as confirmed through the EDA (see Appendix), this window duration captures approximately 160 cycles of the lowest mode at 3.9 Hz, thereby providing a statistically stable PSD estimate for the modal content of interest (Brincker and Ventura 2015). A 50% overlap between consecutive windows was applied, giving a step size of 2,048 samples. Figure 3.5 below illustrates the windowing and overlapping visually. The number of windows extracted from a recording of n samples is given by equation (7):

$$N_w = \left\lfloor \frac{n - W}{S} \right\rfloor + 1 \qquad (7)$$

where $W = 4096$ is the window length and $S = 2048$ is the step size. For the nominal recording length of 65,530 samples, this yields approximately 30 windows per setup per scenario.

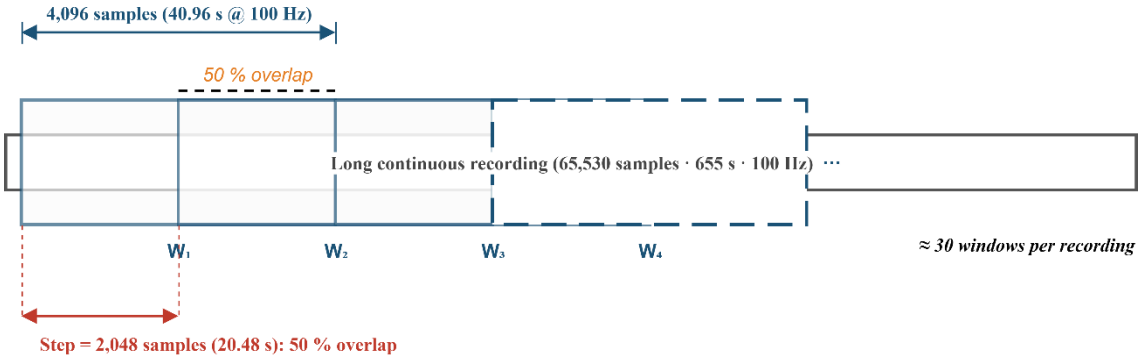


Figure 3.5: Illustration of sliding window segmentation with 50% overlap

3.2.3 Normalization using the StandardScaler

Before each log-PSD window is passed to the autoencoder, it is normalised using a StandardScaler. As shown below in equation (8), the scaler subtracts the mean and divides by the standard deviation of the training data on a per-channel basis:

$$\tilde{x}_{f,c} = \frac{x_{f,c} - \mu_c}{\sigma_c} \qquad (8)$$

where $x_{f,c}$ is the log-PSD value at frequency bin f for channel c, and μ_c and σ_c are the mean and standard deviation computed from the healthy training windows for that channel.

There was an issue during training that required a critical design decision about which data the scaler should be fitted to. Initially, a separate scaler was fitted for each of the nine AVT setups for the healthy scenario (S01). This produced a validation loss of approximately 2.2, which is far above the acceptable range. This is because different setup sessions have different ambient excitation amplitudes, and a scaler fitted on one session does not generalize to another. The solution was to fit a single scaler on all nine S01 setup sessions pooled together, giving each channel a stable mean and standard deviation that reflects the full range of healthy measurement conditions. After applying this pooled scaler, validation loss dropped to 0.10. Subsequently, all windows from any setup of any scenario are normalised using this single healthy reference. This follows the principles in the book of Goodfellow et al. (Goodfellow et al. 2016). The signal processing procedure is summarized in Algorithm 3.1.

Algorithm 3.1: Signal Preprocessing and Feature Representation

Input: MAT files (PDT scenarios, AVT mode), reference channel indices
fs=100 Hz, W=4,096, step=2,048, $nperseg$=1,024, $\varepsilon = 10^{-30}$
Output: X_{scaled}: normalised log-PSD tensor, shape (N × 308 × 5)
scaler: StandardScaler fitted on S01 windows only

for each scenario s, setup k ∈ {1,...,9}:
  data ← load_mat(s, k) // (n_samples × 33)
  ref ← extract_ref_channels(data) // (n_samples × 5)
  for each window w (W=4,096, step=2,048):
    for each reference channel c:
      f, pxx ← Welch(ref[w,:,c], fs=100, $nperseg$=1024)
      PSD[w,:,c] ← $\log_{10}(pxx + \varepsilon)$ // log-PSD transform

// Fit scaler on S01; 9 setups pooled → 270 healthy windows
X_{S01} ← concat(PSD[s=01,k] for k=1,...,9)
scaler ← StandardScaler.fit(X_{S01})
// Apply to all scenarios and setups
for each (s, k):
  X_{scaled}[s,k] ← scaler.transform(PSD[s,k])
return X_{scaled}, scaler

3.3 One-Class Learning and Autoencoder

This section presents and justifies the method, architecture, and training protocol of the Conv1D autoencoder used for anomaly detection in this study.

3.3.1 The Paradigm of One-Class Learning

Supervised classification is a very common machine learning technique, but it requires labelled examples of every damage class the model is expected to identify. In SHM, this requirement cannot be met in practice because anomalies or damaged states are rare, structurally diverse, and unanticipated. It is possible for a bridge to fail through a mechanism that was never observed in any prior structure, making it impossible to enumerate all damage classes at training time (Chandola et al. 2009). One-class learning can address this issue by training solely on the normal (healthy) class and classifying any significant deviation from learned normality as anomalous (Tax and Duin 2004). This is the appropriate design for SHM. The model is trained to learn what a healthy bridge looks like and raises an alarm when the observed data no longer matches that description, irrespective of what type of damage caused the deviation.

3.3.2 Autoencoder for Anomaly Detection

An autoencoder is a neural network trained to compress its input into a compact intermediate representation and then reconstruct the original input from that compact representation. When the autoencoder is trained solely on healthy data, it becomes specialised in reconstructing healthy structural fingerprints. When presented with a damaged bridge, the autoencoder maps its log-PSD input to an unusual latent representation, leading to reconstruction failure. The mean squared error (MSE) between the input and its reconstruction, therefore, serves as a scalar anomaly score and is given by equation (9) below:

$$s = \frac{1}{F \cdot C} \sum_{f=1}^{F} \sum_{c=1}^{C} \left(x_{f,c} - \hat{x}_{f,c} \right)^2 \qquad (9)$$

where $F = 308$ is the number of frequency bins, $C = 5$ is the number of reference channels, x is the scaled log-PSD input, and $\hat{x}$ is the reconstruction. A high score indicates that the autoencoder could not reconstruct the input faithfully, which is interpreted as evidence of a structural anomaly.

3.3.3 Architecture

Before selecting the Conv1D autoencoder architecture, several alternative architectures were considered. LSTM autoencoders, for instance, are designed for temporally sequential data and exploit the causal ordering of samples. However, since the input representation considered here is the log-PSD, which is a frequency-domain quantity whose adjacent bins correspond to nearby frequencies rather than consecutive time steps. The temporal inductive bias of LSTM is therefore not a match with the spectral structure of the input. Transformer autoencoders are also powerful, but to avoid overfitting, their self-attention mechanism requires considerably more training data. A transformer would be severely limited in data, given only 270 healthy training windows. Dense (MLP) autoencoders treat each frequency bin as an independent input and lack a mechanism to exploit the local correlation structure among adjacent frequency bins in the PSD input (Goodfellow et al. 2016; Okur et al. 2025). Conv1D layers were selected because their sliding filters operate locally along the frequency axis, learning the modal peak shapes, inter-modal gradients, and band-level energy patterns, which are structurally meaningful features of ambient vibration spectra (Brincker and Ventura 2015; Ni et al. 2025). This architectural alignment between the inductive bias of Conv1D and the physical structure of the PSD input is the primary justification for the chosen architecture.

As shown in Figure 3.6 below, the autoencoder takes as input a log-PSD window of shape (308, 5) representing 308 frequency bins across 5 reference channels. One-dimensional convolutional layers (Conv1D) are used in both the encoder and decoder because they treat the frequency axis as a sequential dimension and learn local spectral patterns, such as modal peaks, inter-modal gradients, and band-level energy distributions, using learnable filters that slide along the frequency axis.

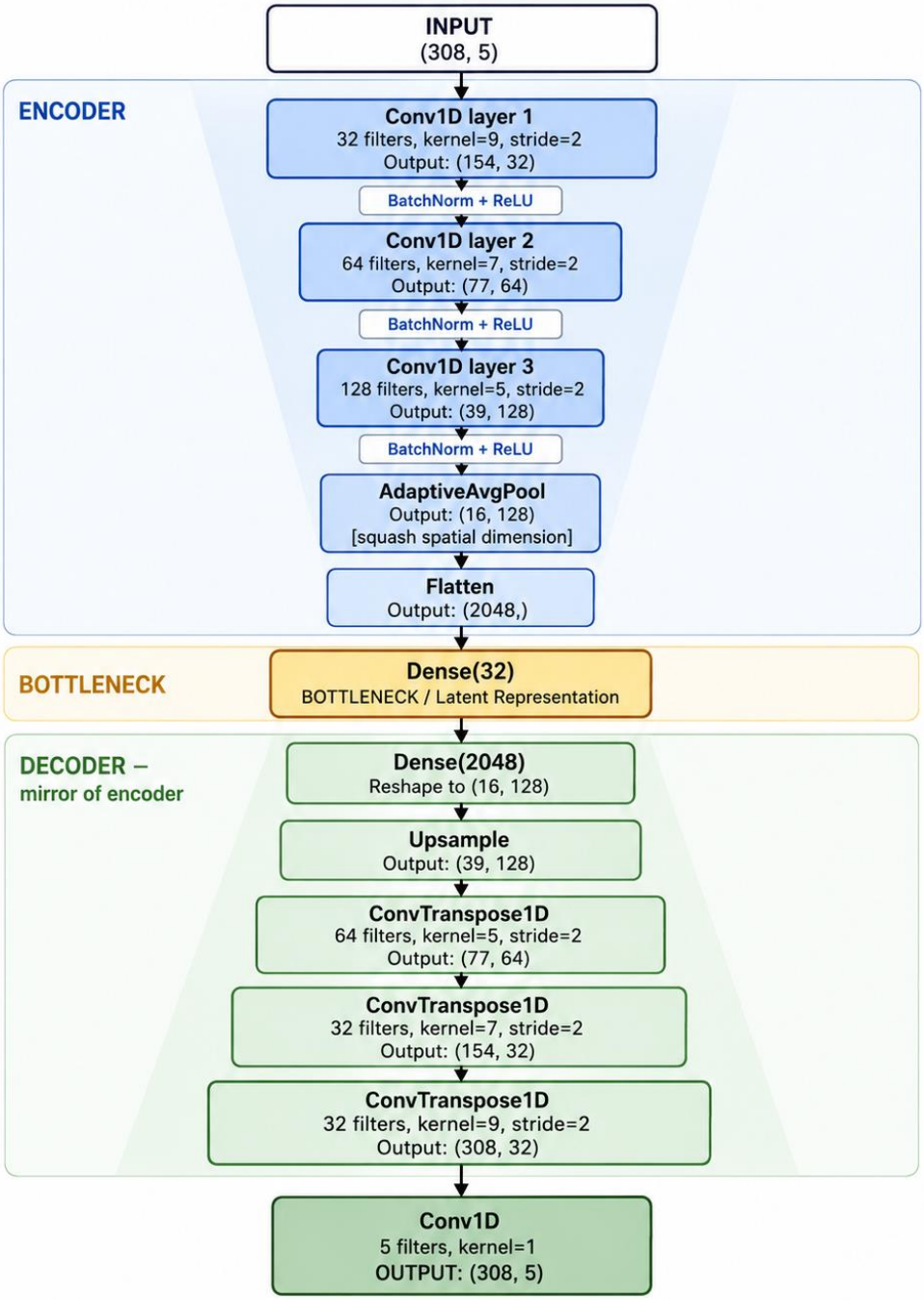


Figure 3.6: Autoencoder architecture diagram showing the full encoder-decoder

The encoder applies three Conv1D layers with 32, 64, and 128 filters, respectively, each using a stride of 2 to progressively compress the frequency dimension from 308 to 154 to 77 to 39 bins. Batch normalisation and ReLU activation follow each convolution. Adaptive average pooling layer reduces the spatial dimension from 39 to 16 before a fully connected layer compresses the representation to a 32-dimensional latent vector. The decoder mirrors the encoder using transposed convolutional layers with upsampling to

restore the original (308, 5) shape. The total number of trainable parameters is 1,187,461.

3.3.4 Model Training Protocol

The autoencoder was trained on all nine AVT setups from Scenario 01, pooled together, yielding 270 windows in total. Training on a single setup (approximately 30 windows) produced a model that memorised the training data rather than learning general structural patterns. Pooling across all reference channels in all setups ensures the model always receives measurements from the same physical sensors, regardless of which setup the window was drawn from. An 80/20 random split was applied to produce 216 training windows and 54 validation windows. A global random seed of 42 + setup was set prior to execution. The model was trained for up to 100 epochs with a batch size of 32 using the Adam optimiser with an initial learning rate of 0.001. MSE was used as the loss function. Also, early stopping with a patience of 15 epochs was applied to halt training when validation loss stopped improving, and a learning rate reduction schedule (factor 0.5, patience 5 epochs) enabled finer parameter adjustment in later training stages (Goodfellow et al. 2016). The best model weights, with the lowest validation loss, were saved and used for all evaluations.

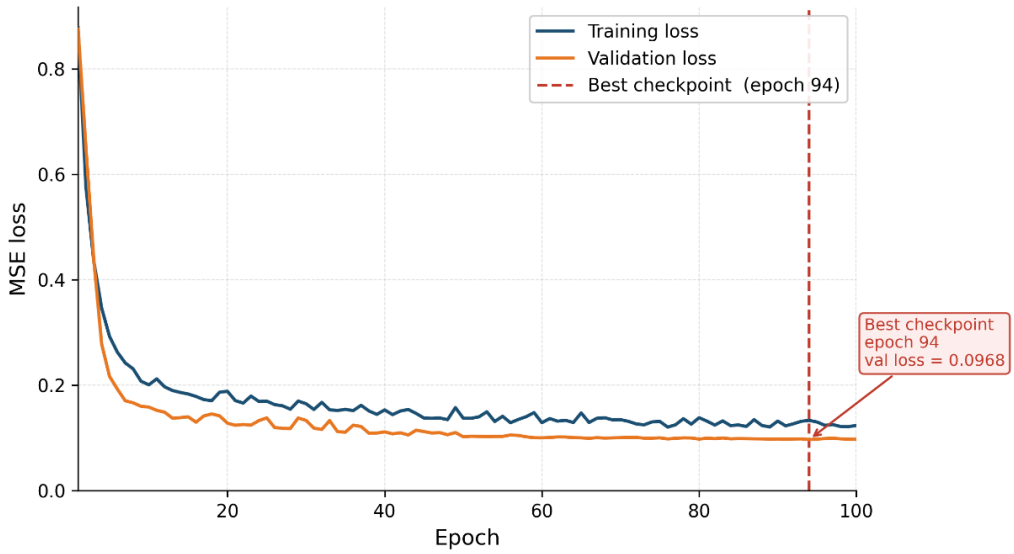


Figure 3.7: Training and validation MSE loss curves with best-checkpoint epoch

The complete training protocol for the Conv1D autoencoder is summarized in Algorithm 3.2 below.

Algorithm 3.2: Conv1D Autoencoder Training Protocol

Input: X_{train} (216 windows, shape 216×308×5)
X_{val} (54 windows), AE with encoder E and decoder D
Output: θ^*: model weights at best validation loss
// Hyperparameters
$lr \leftarrow 0.001$; $batch\_size \leftarrow 32$; $max\_epochs \leftarrow 100$
$patience\_stop \leftarrow 15$; $patience\_lr \leftarrow 5$; $lr\_factor \leftarrow 0.5$
$best\_val\_loss \leftarrow \infty$; $wait \leftarrow 0$
for epoch = 1 to max_epochs:
  // Training
  for each mini-batch $B \subseteq X_{train}$:
    $\hat{B} \leftarrow D(E(B))$ // forward pass
    $\mathcal{L} \leftarrow \text{MSE}(B, \hat{B})$ // MSE loss
    $\theta \leftarrow \text{Adam\_update}(\theta, \nabla_\theta \mathcal{L})$ // update

  // Validation
  $\mathcal{L}_{val} \leftarrow \text{MSE}(X_{val}, D(E(X_{val})))$

  // Learning rate schedule
  if $\mathcal{L}_{val}$ stagnant for $patience\_lr$ epochs:
    $lr \leftarrow lr \times lr\_factor$

  // Early stopping
  if $\mathcal{L}_{val} < best\_val\_loss$:
    $best\_val\_loss \leftarrow \mathcal{L}_{val}$; $\theta^* \leftarrow \theta$; $wait \leftarrow 0$
  else:
    $wait \leftarrow wait + 1$
    if $wait \geq patience\_stop$: break

return θ^*

3.4 Anomaly Scores and Thresholds

This section describes the detection threshold used by the model to make decisions based on the anomaly score. It also explains and justifies the evaluation metrics used and provides an overview of the scoring and threshold algorithm.

3.4.1 Detection Threshold

After training the autoencoder, an anomaly score is computed for each window by passing it through the trained model and calculating the MSE between the input and the reconstruction, as defined in Section 3.3.2. To convert this continuous score into a binary detection decision, a threshold is needed to classify windows with scores above it as anomalous and those below as healthy, as illustrated by Figure 3.8.

This threshold is derived entirely from healthy data, with no reference to any damaged scenario. Anomaly scores were computed for the held-out validation windows from Scenario 01 and the windows from the corresponding Setup of Scenario 08 (the restored condition). As established in Section 3.1.3, Scenario 08 is spectrally equivalent to Scenario 01 in four of five reference channels and is treated as a second healthy reference. Three thresholds were defined at the 95th, 97.5th, and 99th percentiles of this combined healthy score distribution, denoted p95, p97.5, and p99, respectively. These percentiles correspond to false alarm rates (FAR) of 5%, 2.5%, and 1%. At a p99 threshold, for instance, only 1% of genuinely healthy windows will be incorrectly flagged.

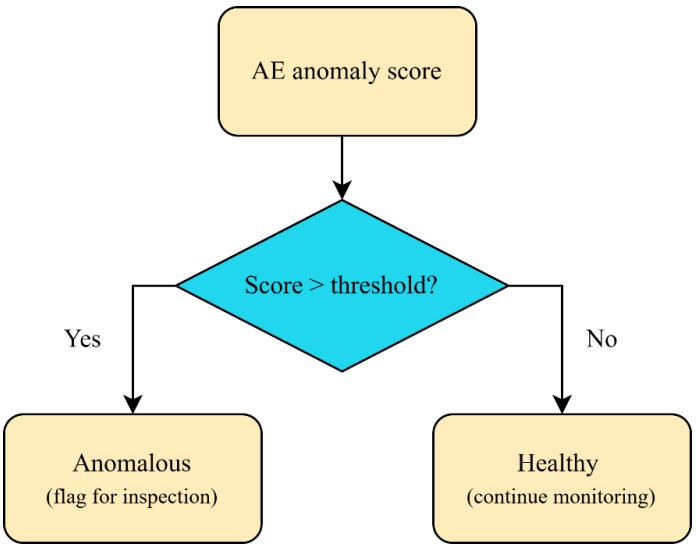


Figure 3.8: Binary detection decision from anomaly score based on threshold

These values are standard in classical statistics, with the 95th and 99th percentiles corresponding to the normal distribution one-tailed critical values at significance levels (α)

of 0.05 and 0.01, which are conventional in hypothesis testing. The 97.5th percentile also corresponds to the one-tailed critical value at α value of 0.025, which is the standard for two-tailed 95% confidence intervals. The three thresholds were included to allow for the recall-FAR trade-off to be evaluated over a range of sensitivity requirements.

3.4.2 Evaluation Metrics

The anomaly detection performance is evaluated using a combination of threshold-free and threshold-dependent metrics. The Area Under the Receiver Operating Characteristic Curve (AUROC) provides a threshold-free summary. It is the probability that a randomly selected damaged window receives a higher anomaly score than a randomly selected healthy window, with values ranging from 0.5 (random) to 1.0 (perfect separation) (Davis and Goadrich 2006). The Area Under the Precision-Recall Curve (AUPRC) complements AUROC by describing the performance under class imbalance conditions, which is more representative of real SHM deployment, where healthy windows far outnumber damaged ones (Saito and Rehmsmeier 2015).

For the threshold-dependent metrics, Recall, Precision, F1 score, and FAR are reported at each of the three percentile thresholds. Recall measures the proportion of damaged windows correctly identified, while Precision measures the proportion of flagged windows that are genuinely damaged. F1 is the harmonic mean of both Precision and Recall, and FAR quantifies the proportion of healthy windows incorrectly flagged as anomalous. In a safety-critical monitoring, Recall is the primary threshold-dependent concern because a missed detection carries greater consequences than a false alarm. Algorithm 3.3 below summarises the scoring and threshold procedure.

Algorithm 3.3: Anomaly Scoring and Threshold-Based Detection

Input: θ^*: trained autoencoder weights
$\mathcal{H}$: healthy windows (S01 + S08, all setups)
$\mathcal{T}$: scaled log-PSD windows from test scenario
Output: scores: per-window MSE reconstruction error
τ_α: thresholds at $\alpha \in \{0.95, 0.975, 0.99\}$
Metrics: AUROC, AUPRC, Recall, Precision, F1, FAR

// Anomaly score: mean squared reconstruction error

$$\text{score}(x) = \frac{1}{F \cdot C} \sum_{f=1}^{F} \sum_{c=1}^{C} \left(x_{f,c} - \hat{x}_{f,c}\right)^2$$

where $x_{rec} = D(E(x))$, $F = 308$ freq. bins, and $C = 5$ channels

// Set percentile thresholds from healthy score distribution

$\mathcal{S}_{\mathcal{H}} \leftarrow \{\text{score}(x) : x \in \mathcal{H}\}$
for each $\alpha \in \{0.95, 0.975, 0.99\}$:
  $\tau_\alpha \leftarrow \text{percentile}(\mathcal{S}_{\mathcal{H}}, \alpha \times 100)$

// Score test windows and classify

for each window $x_i \in \mathcal{T}$:
  $scores[i] \leftarrow \text{score}(x_i)$
  for each α:
    $\hat{y}_\alpha[i] \leftarrow 1$ if $scores[i] > \tau_\alpha$ else 0

// Compute evaluation metrics

AUROC, AUPRC $\leftarrow$ evaluate_ranked($scores$, $\hat{y}_{true}$)
for each α: Recall, Prec, F1, FAR $\leftarrow$ evaluate($\hat{y}_\alpha$, $\hat{y}_{true}$)

return $scores$, τ_α, AUROC, AUPRC, Recall, Prec, F1, FAR

The formulas for the different metrics are shown below in equations (10) to (14). AUPRC follows the same integral logic as AUROC, but over a precision-recall curve.

$$\text{AUROC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (12)$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

$$\text{FAR} = \frac{FP}{FP + TN} \quad (14)$$

where $TPR(t)$ and $FPR(t)$ are the True Positive Rate and the False Positive Rate, TP and FP are True Positive and False Positive, and TN is True Negative.

3.5 SHAP Explainability Pipeline

This section explains the rationale for the post-hoc analysis, establishes a feature bank to attribute the anomaly detection model's output, and provides an overview of the explainability model and its output.

3.5.1 Motivation for Post-hoc Explainability

As previously established in Section 1.2, the autoencoder anomaly score only indicates that the structural behaviour of the bridge has changed. It does not identify which sensor recorded the change, which frequency band was affected, or whether the change is consistent with a known structural failure mode. This information is essential for translating a detection output into an actionable maintenance decision. The post-hoc explainability methodology used addressed this gap by analysing a trained model in order to attribute its outputs to specific input features (Adadi and Berrada 2018). SHAP (SHapley Additive exPlanations) was selected for this purpose because, unlike the gradient-based alternatives or LIME (Ribeiro et al. 2016), it provides attributions with formal theoretical guarantees derived from cooperative game theory, that attributions are consistent, locally accurate, and sum to the difference between the model's prediction and its baseline expectation (Lundberg and Lee 2017).

3.5.2 Feature Bank

It is generally impractical to apply SHAP directly to the autoencoder due to interpretation difficulties. The input space of the autoencoder has $308 \times 5 = 1{,}540$ raw frequency bins and would produce attributions at a very high resolution that has no physical meaning to a structural engineer. Instead of this high number of features (1540), 92 physically named features are extracted from each window, and SHAP is applied to a surrogate model trained on these features, as described in Section 3.5.4. The 92 features are organised into four different families, which are described below:

3.5.2.1 Family A - Spectral statistics

For this family, the mean log-PSD, standard deviation of log-PSD, spectral entropy, and spectral centroid were considered. These features capture the overall energy level, spread, and distributional shape of the PSD across the full frequency range (Buckley et al. 2023). This gives a total of 20 features (5 channels × 4 features). The spectral centroid and entropy are computed as shown in equations (15) and (16) below:

$$\text{Spectral centroid,} \quad C = \frac{\sum_f f \cdot P(f)}{\sum_f P(f)} \tag{15}$$

where f is frequency and $P(f)$is the PSD value at that frequency. It gives a single number representing where the spectral energy is concentrated. A stiffness loss that shifts energy to lower frequencies will pull the centroid downward.

$$\text{Spectral entropy,} \quad H = -\sum_f p(f) \log_2 p(f) \tag{16}$$

where $p(f) = \frac{P(f)}{\sum_f P(f)}$ is the normalised PSD treated as a probability distribution. A healthy bridge with well-defined modal peaks would have a low entropy because the energy is concentrated at a few frequencies. While a damaged or noise-contaminated bridge signal with energy spread widely across frequencies would have higher entropy.

3.5.2.2 Family B - Band energies

Here, the mean log-PSD within eight frequency bands spanning 0–30 Hz was utilized to describe the band energies. The bands are 0–1, 1–2.5, 2.5–5, 5–7.5, 7.5–10, 10–12.5, 12.5–15, and 15–30 Hz. These bands target the modal content at known Z24 natural frequencies, especially the 2.5–5 Hz band, which contains the first two bending modes as confirmed in the EDA and across other Z24 literature. This set of band energies contributes a total of 40 features (5 channels × 8 features).

3.5.2.3 Family C - Modal proxies

For this family, the peak frequency and peak amplitude of the dominant spectral peak, and mean band energy in the first (3.4–4.4 Hz) and second (4.4–5.4 Hz) bending mode bands. These features directly quantify the modal frequency and amplitude characteristics most sensitive to stiffness changes. Adding extra 20 features (5 channels × 4 features).

3.5.2.4 Family D - Transmissibility ratios

For this family, R1V, R2L, R2T, and R3V are paired with the reference channel R2V. For each of these pairings, the ratio of their peak PSD amplitudes, the frequency of maximum transmissibility, and the ratio of band energies in the 2.5–10 Hz range were computed. These three features are referred to as the transmissibility features. Transmissibility describes how vibration energy transmits between structural locations, which changes when the damage alters the structural path between sensors (Abdrabo 2024). R2V is used as the denominator because it is the vertical mid-span reference channel, which is present in all setups and most widely adopted as the reference sensor in Z24 transmissibility studies (Abdrabo 2024; Maeck and De Roeck 2003; Peeters and De Roeck 2001; Reynders and Roeck 2008).

To further explain each transmissibility feature, the peak amplitude ratio captures how much vibration energy the numerator sensor receives relative to the mid-span sensor. This is because when pier settlement alters the distribution of structural stiffness, the energy balance between sensors changes. Peak frequency ratio, on the other hand, is

the ratio of the frequency at which each channel reaches its maximum PSD value. In a healthy bridge, both sensors are expected to resonate at the same dominant modal frequency, so this ratio should be close to 1. But a structural change that shifts the dominant mode differently at two different sensor locations will move this ratio away from 1. Finally, the band energy ratio is the ratio of mean log-PSD energy in the 2.5–10 Hz band (which covers the first two bending modes) between the numerator channel and R2V. This captures how modal energy transmits between sensor locations across the structurally most important frequency range. This adds 12 features to the list (4 sensor pairs × 3).

The selection of these four families is motivated by the types of change pier settlement is known to induce. When stiffness decreases, the modal frequencies shift (captured by Families B and C), the spectral energy distribution changes (Family A), and the vibration transmission between different sensors is modified (Family D). Table 3.2 below gives a summary of the feature bank.

Table 3.2: Summary of the 92 features used for SHAP explanation

Family	Features	Channels	Total	Physical motivation
A: Spectral statistics	mean, std, entropy, centroid	5	20	Global spectral shape change
B: Band energies	8 frequency bands	5	40	Modal band energy redistribution
C: Modal proxies	peak_freq, peak_amp, mode1, mode2	5	20	Direct modal parameter tracking
D: Transmissibility	peak_ratio, peak_freq, band_ratio	4 pairs	12	Inter-sensor path change

3.5.3 Per-Setup Normalization

Before training the surrogate model, each of the 92 features is normalised per setup using the mean and standard deviation computed from the healthy (S01) windows of that setup. Equation (17) below shows the formula for the normalization:

$$z_{i,k} = \frac{f_{i,k} - \mu_{i,k}^{\text{healthy}}}{\sigma_{i,k}^{\text{healthy}}} \quad (17)$$

where $f_{i,k}$is the value of feature i in window k, and $\mu_{i,k}^{\text{healthy}}$ and $\sigma_{i,k}^{\text{healthy}}$ are the mean and standard deviation of feature i across the healthy windows of the same setup. After normalization, the feature values are expressed in units of standard deviations above the healthy baseline for that setup. For instance, a value of $+4$ indicates the feature is 4 standard deviations above the typical value for a healthy bridge in that measurement zone. Carrying out this normalisation was essential because, without it, the surrogate R^2 collapsed to near zero ($R^2 \approx 0$) due to significant differences in the amplitudes of the raw features across setups. This is due to varying ambient excitation levels at each setup, making it impossible for the model to learn a consistent mapping from healthy to damaged.

3.5.4 Surrogate Model

An XGBoost regressor, as described in the work of Chen et al. (2016), was trained to predict the autoencoder's anomaly scores from the 92 normalised features. XGBoost was selected over other types of regressors because it is compatible with TreeSHAP (Lundberg and Lee 2017), which computes exact SHAP values for tree-based models in polynomial time. For non-tree models, SHAP values must be approximated, and this can introduce uncertainty into feature attributions.

The surrogate model was evaluated using 5-fold cross-validation, and SHAP attributions were considered reliable only if $R^2 \geq 0.70$. This value was selected because it is practically sufficient to capture the effect of EOVs while preserving damage sensitivity (Peeters and De Roeck 2001), and regression models yielding moderate performance in the 0.70s are commonly utilized and evaluated in modern SHM literature (Chen et al. 2024; Kim, Kim, and Hwang 2025). Below this threshold, the surrogate model is considered not to faithfully represent the autoencoder's prediction or scoring behaviour. The surrogate was trained on the pooled healthy and damaged windows for a given combination of setup and scenario. Figure 3.9 presents the flowchart of the surrogate strategy.

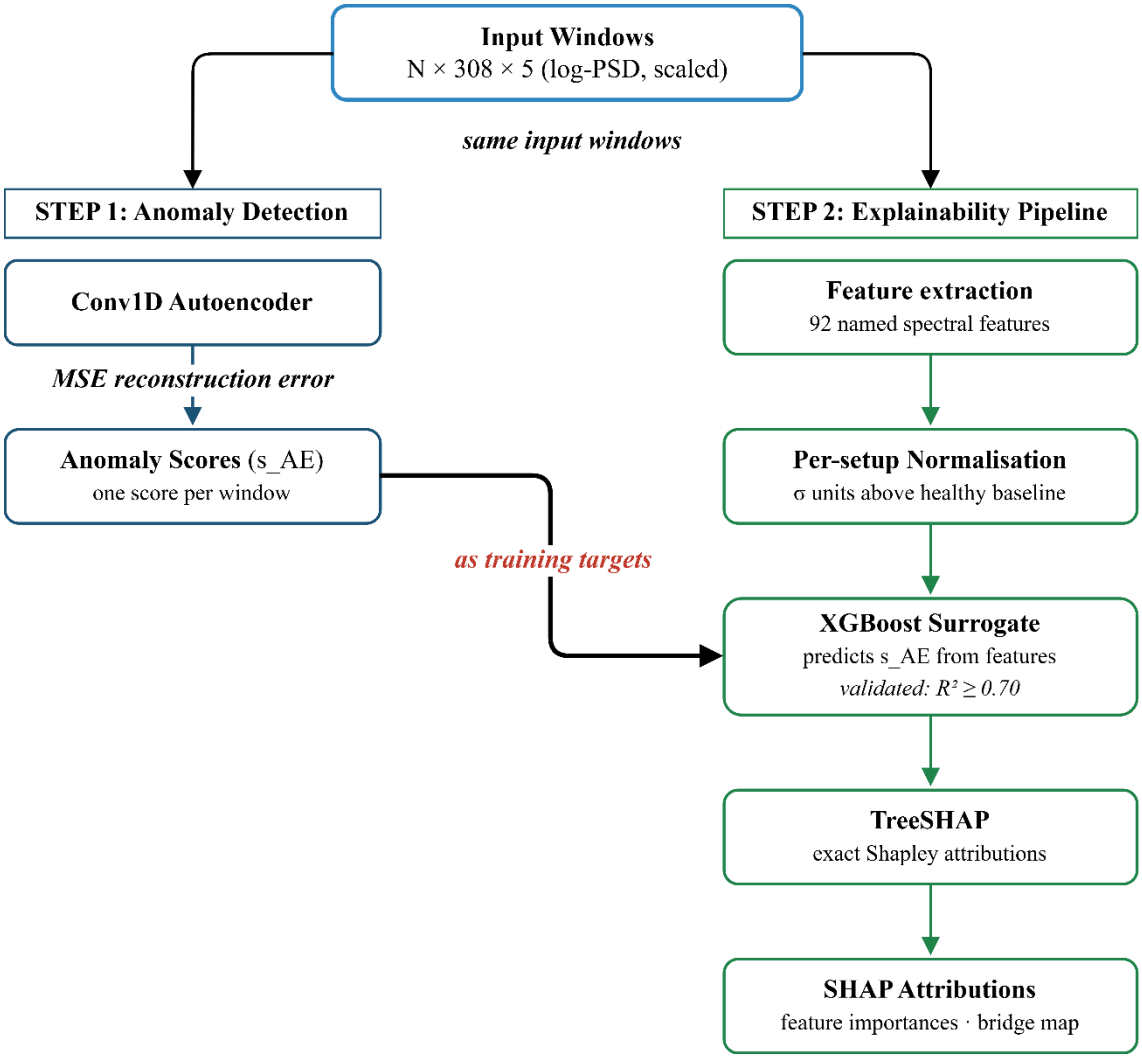


Figure 3.9: Two-step surrogate strategy diagram

3.5.5 SHAP Output Interpretation

After a surrogate is validated, TreeSHAP is used to compute SHAP attributions. For each window, every feature receives a SHAP value ϕ_i which represents its marginal contribution to the predicted anomaly score of that window relative to the expected score across the dataset. Positive values push the prediction toward anomalous, and negative values push it toward healthy.

The SHAP values satisfy the efficiency property, which means they sum exactly to the difference between the prediction and the baseline expectation, thereby guaranteeing that the full prediction is accounted for with no residual unexplained component

(Lundberg and Lee 2017). The summary of the explainability model is shown in Algorithm 3.4 below.

Algorithm 3.4: SHAP Explainability via XGBoost Surrogate

Input: θ*: trained autoencoder weights
X: windows for scenario $\boldsymbol{s}$ and setup $\boldsymbol{k}$
s_AE: AE anomaly scores for X (Algorithm 3.3)
Φ_{names}: 92 named spectral features (Sec. 3.5.2)
R^2_{min} = 0.70 (minimum surrogate fidelity)
Output: shap_vals: SHAP values, shape (N × 92)
φ_{global}: global importance (mean |Φ| per feature)

```
// Step 1: Extract 92 named spectral features
for each window x_i ∈ X:
    F_i ← extract_features(x_i, Φ_names)    // shape: (92,)

// Step 2: Per-setup normalisation (σ units above healthy baseline)
for each feature j ∈ {1,...,92}:
    μ_j, σ_j ← mean, std of F[:,j] over S01 windows of setup k
    Z_i[j] ← (F_i[j] − μ_j) / σ_j

// Step 3: Train XGBoost surrogate and validate fidelity
g ← XGBoost(n_estimators=300, max_depth=4)
R² ← cross_val_score(g, Z, s_AE, cv=5)
if mean(R²) < R²_min log low-fidelity warning
g.fit(Z, s_AE)

// Step 4: Exact Shapley values via TreeSHAP
for each normalised window z_i:
    shap_vals[i] ← TreeSHAP(g, z_i)  // exact – verify efficiency

// Step 5: Global feature importance
φ_global[j] ← (1/N) × Σ_i |shap_vals[i, j]|
return shap_vals, φ_global
```

Four visualisations are produced from the SHAP values and are used as the primary outputs of the explainability analysis in Chapter 4:

1. **Global importance bar chart:** This chart shows the mean absolute SHAP value per feature, computed separately over healthy and damaged windows, revealing which features the model relies on to distinguish the two classes.

2. **Beeswarm plot:** In this plot, each window is represented as a point, which is positioned by its SHAP value and coloured by its normalised feature value, revealing the direction and consistency of each feature's contribution.
3. **Waterfall plot:** A single window's full SHAP decomposition showing the stepwise contribution from the baseline expectation to the final predicted score, used for case-by-case anomaly explanation.
4. **Bridge schematic:** This is a schematic outline of the bridge on which the summed mean absolute SHAP values, aggregated by reference channel, are plotted spatially, mapping the importance of each sensor location to the detected anomaly.

4 Results

This chapter presents the experimental results of the Conv1D autoencoder and SHAP explainability pipeline evaluated on the Z24 bridge PDT dataset. The results are organized to address the three research questions stated in Section 1.3. The chapter begins by presenting Section 4.1 which provides qualitative inspection of the model's behaviour through reconstruction instances, a surrogate decision tree, and anomaly score distributions, thereby laying the foundation for what the models do before quantitative analysis. Section 4.2 reports the autoencoder training convergence. Section 4.3 quantifies detection performance across all nine measurement setups using metrics such as AUROC, AUPRC, and other previously mentioned threshold-dependent metrics, and compares it with baseline models. Sections 4.4 and 4.5 present the results of the SHAP feature attribution for the primary benchmark scenario (S05) and assess their consistency across the different setups. Finally, Section 4.6 examines how feature attributions change across the four pier settlement scenarios (S03, S04, S05 and S06), identifying a damage severity threshold for reliable structural explanation.

4.1 Qualitative Model Inspection

This section provides an overview of the autoencoder and surrogate models before the thesis results are presented.

4.1.1 Autoencoder Reconstruction

Figure 4.1 below shows the autoencoder's reconstruction of a median healthy window from S01 (top) and the most anomalous damaged window from S05 (bottom), the 80mm pier settlement, both selected for Setup 09 and R2L channel. The solid line shows the scaled log-PSD input, and the dashed line shows the autoencoder's reconstruction. The shaded region is the squared reconstruction error. The healthy window is reconstructed with high fidelity (MSE = 0.0718), showing that the model has learned the healthy spectral fingerprint. The damaged window, on the other hand, shows systematic

reconstruction error concentrated at the modal frequency bands (MSE = 0.3569, approximately 5× higher), which the model cannot reproduce because the spectral pattern deviates from that of the healthy training data. The scalar mean squared error across all frequency bins and channels serves as the anomaly score defined in Section 3.4.

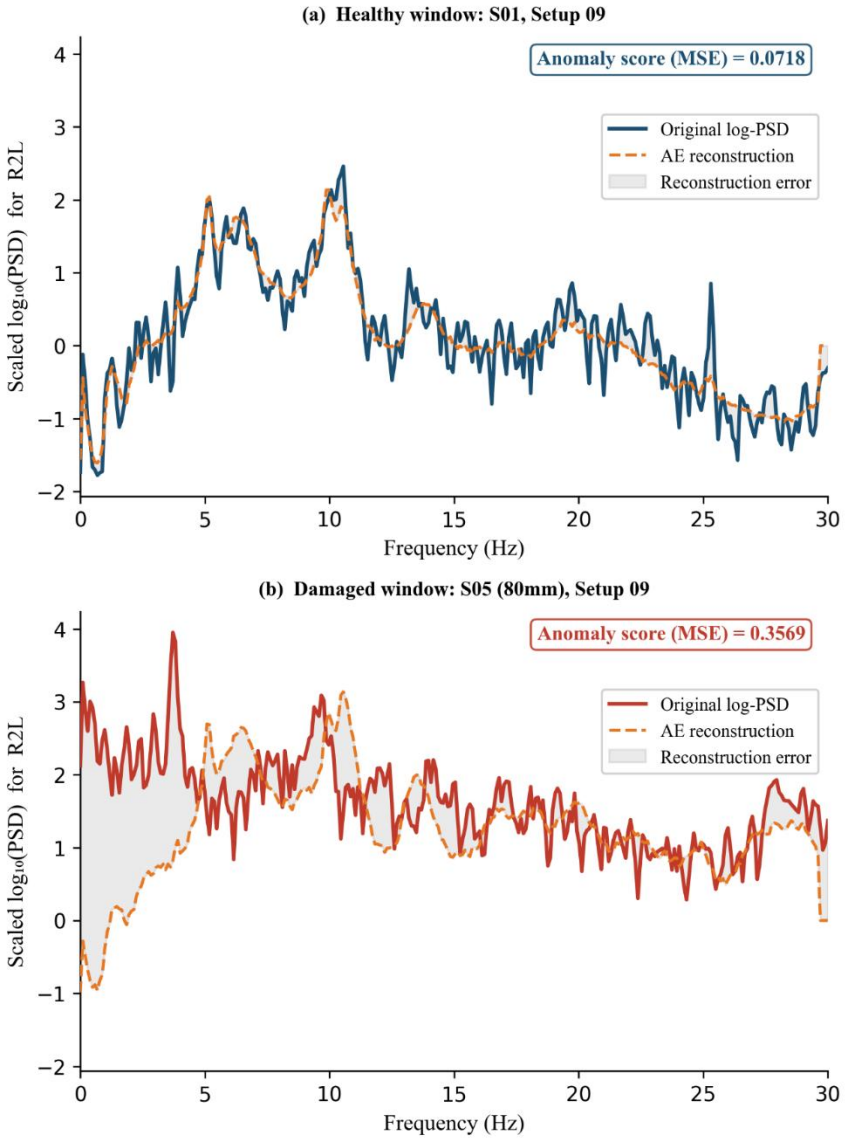


Figure 4.1: AE reconstruction of a healthy and damaged window

4.1.2 Surrogate Decision Tree

Figure 4.2 below presents a single tree (tree 0, with depth limited to 3 out of the full model depth of 4) extracted from the fitted XGBoost surrogate for Setup 09 against Scenario 05.

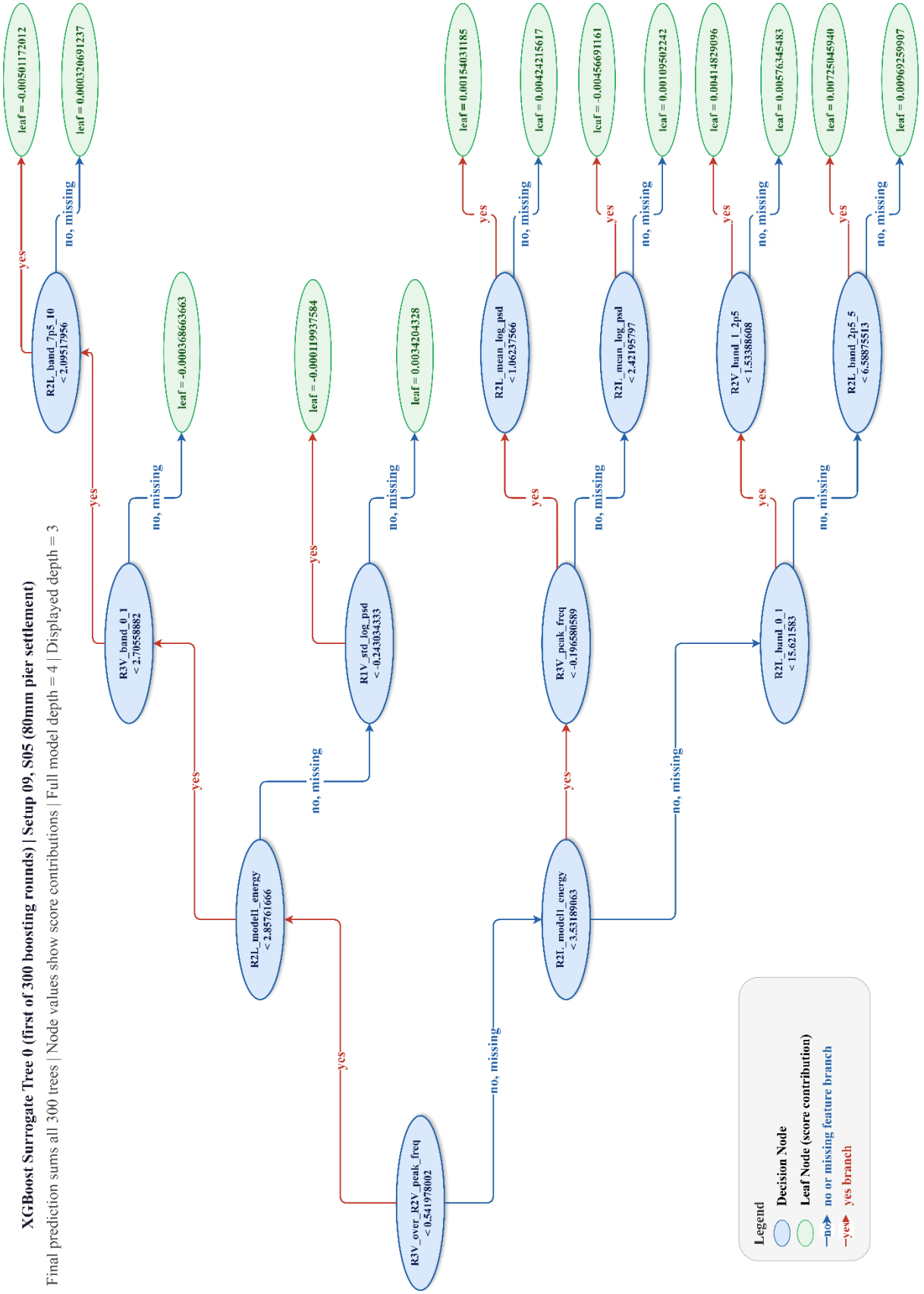


Figure 4.2: One tree from the XGBoost surrogate (Setup 09, S05)

The surrogate consists of 300 such trees whose predictions are summed to approximate the autoencoder's anomaly score. Each of the internal nodes applies a threshold to one of the 92 normalised spectral features. The leaf values represent the incremental score contribution from that branch for the given tree. The dominant splitting feature at the root node of the tree above ($R3V\_over\_R2V\_peak\_freq$) matches the global SHAP importance rankings, which are reported in Section 4.4. This confirms that the surrogate has learned a physically coherent mapping from features to anomaly scores.

4.1.3 Anomaly Score Distributions

Figure 4.3 below shows the anomaly score distribution corresponding to the distribution of the reconstruction error (MSE) of the healthy S01 and the damaged S05. The Figure clearly shows score separation between the healthy and damaged scenarios, and the yellow, orange, and blue dashed lines represent the thresholds at p95, p98 (p97.5), and p99 corresponding to false alarm rates of 5%, 2.5%, and 1%, respectively.

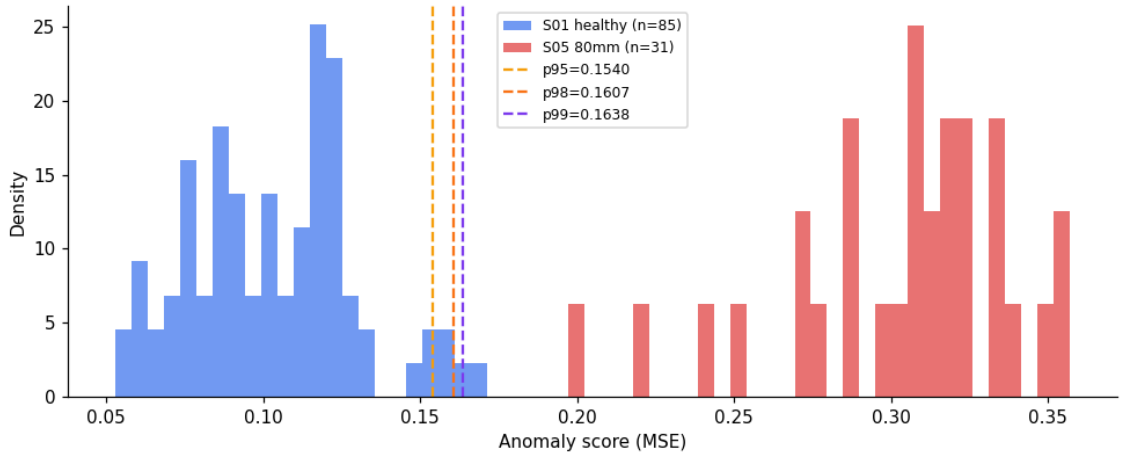


Figure 4.3: Anomaly score distribution for setup09 (S01 vs S05)

4.2 Autoencoder Training

For each of the nine AVT setups, the Conv1D autoencoder was trained separately using the same pooled 270 healthy windows from Scenario 1 and the same fixed random seed schedule (seed = 42 + setup number). All of the nine models for each setup converged

within 100 epochs without triggering early stopping. The training and validation curves for Setup 01 are shown in Figure 3.7 as a representative example. The validation loss decreased steadily from approximately 0.19 at epoch 1 to 0.097 at the marked best checkpoint. All nine models achieved final validation losses of 0.09-0.10, indicating consistent convergence across the different setups.

4.3 Detection Performance

This section reports the results of the proposed model across different evaluation metrics and compares them with those of two baseline methods: Principal Component Analysis (PCA) reconstruction error and Isolation Forest.

4.3.1 Proposed model results

Table 4.1 below presents the AUROC, AUPRC, and F1 scores at the p95, p98, and p99 threshold percentiles for all nine setups, evaluated on Scenario 05 (80 mm pier settlement) against the pooled healthy validation scenario (S01). The AUROC for all nine setups exceeds 0.95, with a mean AUROC of 0.987 across all measurements.

Table 4.1: Autoencoder detection performance across all nine setups for S05

Setup	Zone	AUROC	AUPRC	F1@p95	F1@p98	F1@p99
01	Utzenstorf	1.0000	1.0000	0.925	0.954	0.984
02	Utzenstorf	0.9977	0.9941	0.925	0.954	0.951
03	Utzenstorf	0.9529	0.7346	0.925	0.000	0.000
04	Utzenstorf	0.9769	0.8660	0.925	0.786	0.316
05	Mid-span	0.9818	0.9274	0.925	0.903	0.439
06	Koppigen	0.9875	0.9634	0.892	0.867	0.877
07	Koppigen	0.9981	0.9953	0.925	0.937	0.968
08	Koppigen	0.9875	0.9571	0.925	0.921	0.745
09	Koppigen	1.0000	1.0000	0.925	0.954	0.984
Mean		**0.987**	**0.938**			

Setups 01 and 09 both obtained a perfect AUROC = 1.000, implying that every damaged window for these setups received a higher anomaly score than every healthy window, while Setup 03 is the weakest performer. Out of the nine setups, eight achieved a recall of 1.000 at the p95 threshold, meaning all damaged windows are correctly flagged with 5.9% false alarm rate. Setup 09 achieves a recall of 1.000 and an F1 of 0.984 at the strictest p99 threshold (1.2% FAR), the strongest threshold-level result of the experiment. Both the Koppigen and Utzenstorf-zone setups achieve high performance, indicating that the pier settlement propagates globally throughout the bridge, making it detectable from any measurement location.

4.3.2 Baseline Comparison

Two baseline methods were evaluated for comparison with the proposed model. The first is PCA reconstruction error, tested with three component configurations (k = 10, 20, and 50) fitted on the pooled healthy training windows, and the second is Isolation Forest with 300 estimators and a contamination parameter of 0.05. Both models were applied to the same scaled log-PSD windows used by the proposed model and evaluated against the same test scenarios. Table 4.2 below presents the results of both baselines and the proposed model.

Table 4.2: Baseline comparison of the proposed model with PCA and Isolation Forest

Method	AUROC			AUPRC		
	S03	**S04**	**S05**	**S03**	**S04**	**S05**
Isolation Forest	0.680	0.585	0.961	0.149	0.127	0.559
PCA (k=10)	0.984	0.965	1.000	0.906	0.753	1.000
PCA (k=20)	0.998	0.995	1.000	0.980	0.959	1.000
PCA (k=50)	1.000	1.000	1.000	1.000	1.000	1.000
Conv1D AE	**0.953**	**0.977**	**1.000**	**0.735**	**0.866**	**1.000**

From the Table above, Isolation Forest is clearly inferior to PCA reconstruction error and the proposed Conv1D autoencoder model across all scenarios. Also, the Conv1D AE

underperforms PCA at k=20 across S03 and S04 but outperforms it at k=10 on S04. PCA with k=50 matches or exceeds the Conv1D AE on every metric across all scenarios.

Based on the comparison above, the key finding for the Z24 Bridge dataset is that its log-PSD spectral representation is so discriminative that even a linear method such as PCA can separate healthy from damaged windows. The fact that the PCA reconstruction error method outperforms the Conv1D autoencoder, however, is related to the properties of the dataset and the representation, not a failure of the experiment. The unique value of the Conv1D autoencoder, however, lies in enabling the SHAP pipeline (which PCA cannot provide) and not in outperforming PCA on AUROC and AUPRC.

4.4 SHAP Explainability Result: S05, Setup 09

For the SHAP analysis, Setup 09 (within the Koppigen zone), which achieved an AUROC of 1.000, was selected as the primary setup. The XGBoost surrogate trained on 92 normalised spectral features achieved a cross-validated R^2 of 0.792 ± 0.064, confirming adequate fidelity to the autoencoder's scoring behaviour. The healthy and damaged score distributions are well separated (healthy mean = 0.080, damaged mean = 0.275), confirming that the surrogate is trained on a meaningful signal. Table 4.3 lists the top ten features by mean absolute SHAP value.

Table 4.3: Top 10 SHAP features for Setup 09, S05 (80mm settlement)

Rank	Feature	Mean |SHAP|	Family
1	R3V_over_R2V_peak_freq	0.0366	D - Transmissibility
2	R2L_mode1_energy	0.0256	C - Modal proxy
3	R2L_band_0_1	0.0194	B - Band energy
4	R2V_peak_freq	0.0123	C - Modal proxy
5	R3V_peak_freq	0.0065	C - Modal proxy
6	R2L_band_15_30	0.0030	B - Band energy
7	R2L_mean_log_psd	0.0028	A - Spectral statistic
8	R2L_band_2p5_5	0.0025	B - Band energy
9	R2T_over_R2V_peak_ratio	0.0021	D - Transmissibility
10	R2L_mode2_energy	0.0021	C - Modal proxy

The dominant feature is $R3V\_over\_R2V\_peak\_freq$, which is the ratio of the dominant spectral peak frequency of the R3V channel (vertical, near Utzenstorf pier) to that of the mid-span reference R2V. Under healthy conditions, this ratio reflects the relative modal tuning between a sensor on the far side of the bridge and a mid-span sensor. The pier settlement in the Koppigen zone redistributes bending moments (the internal forces that cause a beam or bridge span to bend under load) across the span, shifting the relative peak frequencies differently at R3V and R2V and causing this ratio to deviate from its healthy baseline. The second feature is $R2L\_mode1\_energy$ (energy of the lateral mid-span sensor in the first bending mode band 3.4-4.4 Hz), which captures the amplitude consequence of the same stiffness redistribution due to the pier settlement.

The most important channel across all features is R2L (lateral mid-span), which accumulates the highest summed SHAP importance. This is shown spatially in Figure 4.8 (bridge schematic). Even though R1V is the reference sensor physically closest to the damaged Koppigen pier, it does not rank first on the list (see Figure A.5). This finding is discussed further in Section 5.1. The global feature importance, waterfall, beeswarm, and bridge schematic figures are shown below in Figure 4.4 to Figure 4.8.

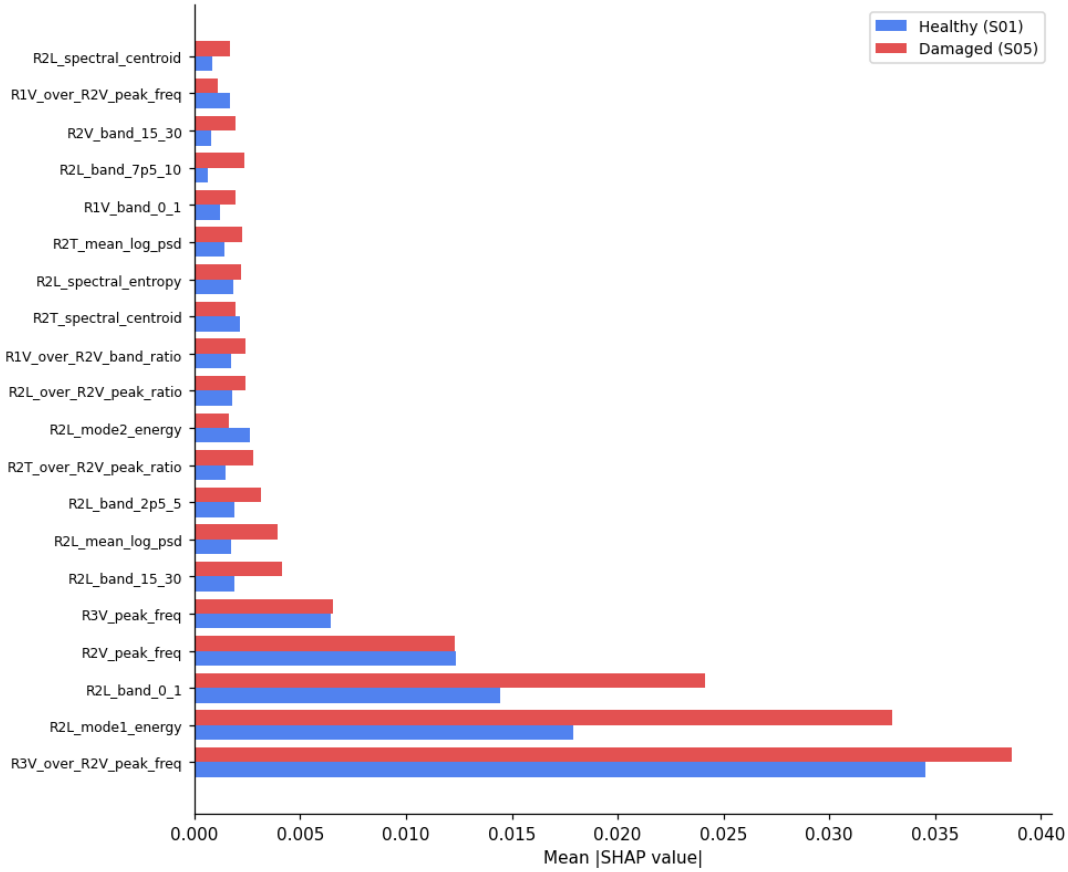


Figure 4.4: Global feature importance plot of the top features based on mean absolute SHAP values

The global feature importance plot in Figure 4.4 above shows the contribution of each feature, measured as the mean absolute SHAP value, to the model's overall prediction across all healthy and damaged windows. Similar to what was presented in Table 4.3, the chart shows that $R3V\_over\_R2V\_peak\_freq$ is dominant on average across all damaged windows.

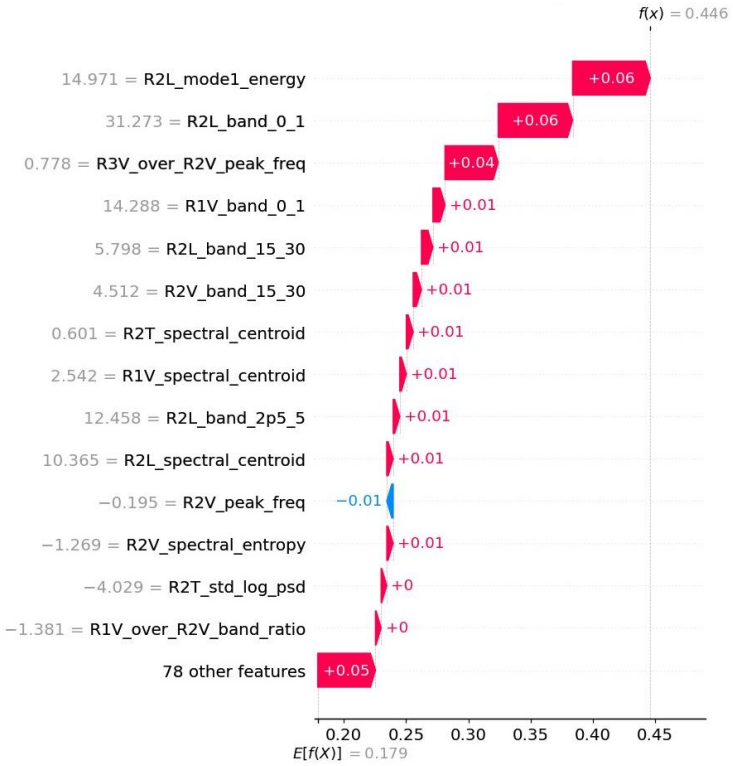


Figure 4.5: SHAP waterfall plot for the most-anomalous window of Setup09, S05

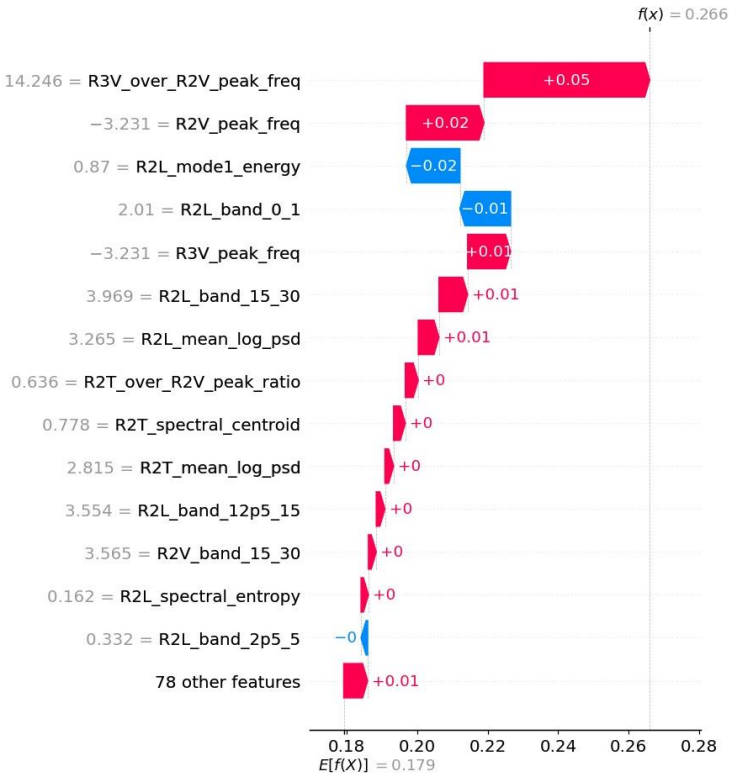


Figure 4.6: SHAP waterfall plot for a median anomalous window of Setup09, S05

From the two waterfall plots shown above in Figure 4.6 (w231, median anomalous, AE score = 0.268) shows $R3V\_over\_R2V\_peak\_freq$ as the dominant positive contributor (+0.05), which is consistent with the global importance ranking. On the other hand, Figure 4.5 (w163, most anomalous, AE score = 0.463) reveals a broader attribution pattern, where $R2L\_mode1\_energy$ and $R2L\_band\_0\_1$ each contributes +0.06, with $R3V\_over\_R2V\_peak\_freq$ contributing +0.04. This indicates that at extreme anomaly levels, the first bending mode energy at the lateral mid-span sensor becomes co-dominant with the transmissibility ratio, reflecting a more severe structural deviation in which multiple spectral features simultaneously exceed their healthy baseline.

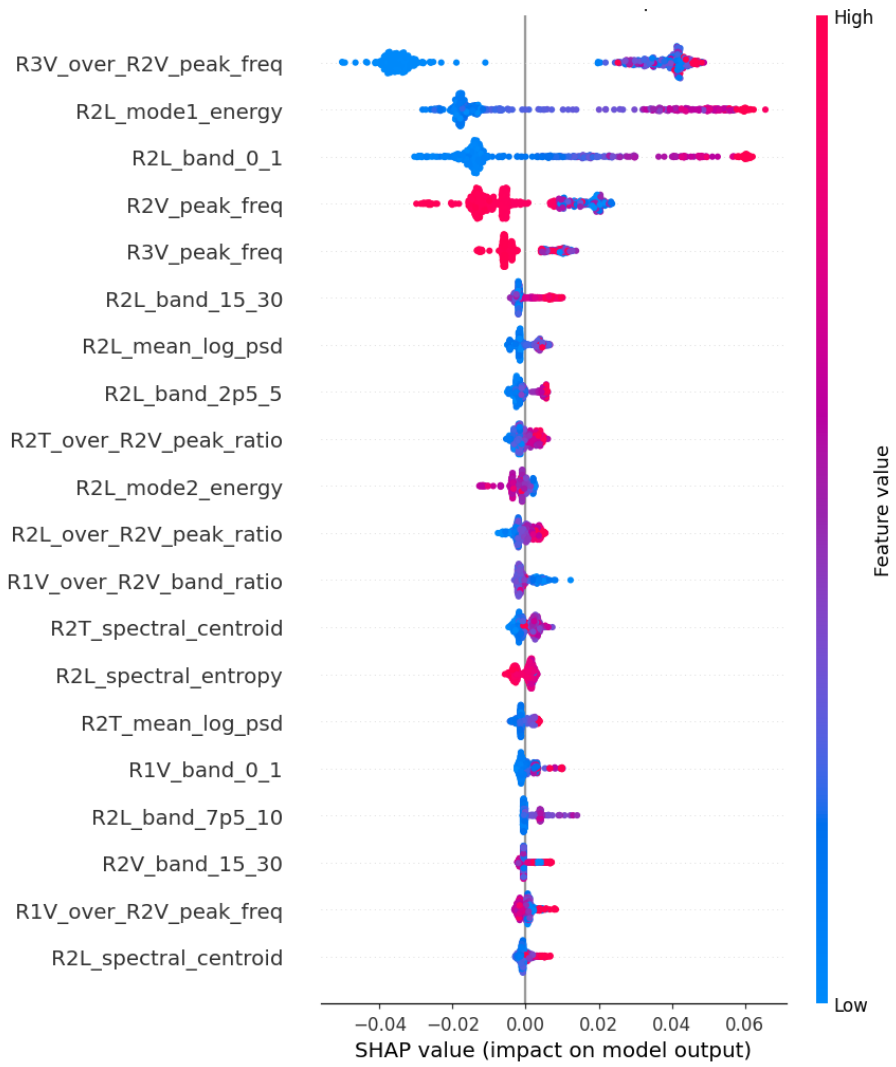


Figure 4.7: SHAP beeswarm plot for Setup09 Scenario 05

The beeswarm plot in Figure 4.7 above shows the SHAP feature importance of Setup 09, and just like the global importance plot, confirms that $R3V\_over\_R2V\_peak\_freq$, $R2L\_mode1\_energy$, and $R2L\_band\_0\_1$, are the top contributors and can push the anomaly score to +0.06 when their values are high. $R2V\_peak\_freq$, on the other hand, behaves inversely, with a high value giving a negative SHAP contribution.

Figure 4.8 below shows the spatial distribution of mean absolute SHAP importance values across the five reference channels. The R2L channel shows the highest SHAP importance, confirming that the lateral vibration response at the centre of the span is the most informative for detecting the Koppigen pier settlement, while R1V and R3V contribute secondary importance, and T2T and R2V have relatively lower attributions.

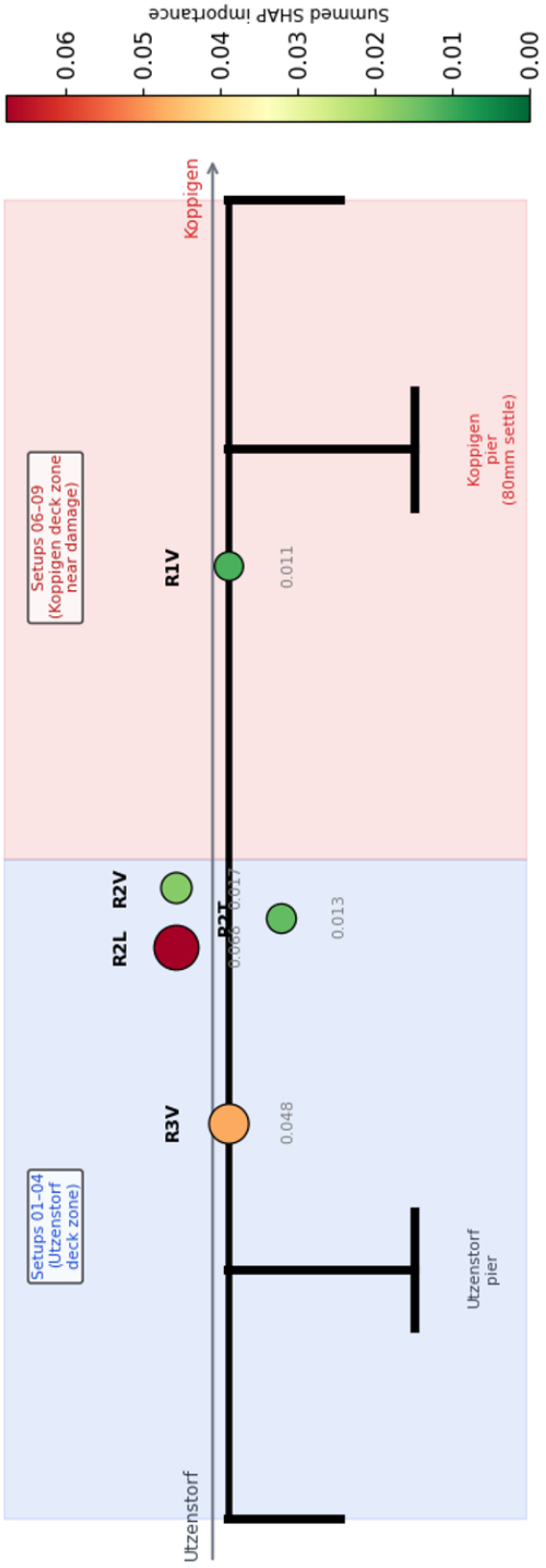


Figure 4.8: Bridge schematic showing SHAP channel importance for Setup09 Scenario 05

4.5 Multi-Setup SHAP Consistency

To assess whether the detected features reflect genuine structural phenomena and not merely artefacts of a single model, the SHAP pipeline was applied independently to four setups spanning both measurement zones. Table 4.4 summarises the results.

Table 4.4: Multi-setup SHAP comparison for Scenario 05 (80mm settlement)

Setup	Zone	Surrogate R^2	Top Feature	Top Channel
01	Utzenstorf	0.798 ± 0.084	R2L_mode1_energy	R2L
02	Utzenstorf	0.811 ± 0.041	R3V_over_R2V_peak_freq	R2L
07	Koppigen	0.810 ± 0.059	R3V_over_R2V_peak_freq	R2L
09	Koppigen	0.792 ± 0.064	R3V_over_R2V_peak_freq	R2L

Out of the four setups, three setups identify $R3V\_over\_R2V\_peak\_freq$ as the top feature. In Setup 01, $R2L\_mode1\_energy$ ranks first, and $R3V\_over\_R2V\_peak\_freq$ closely followed with a nearly identical importance (0.0287 and 0.0278, respectively), and the same set of features appears in every top five of the lists across all setups. Most notably, R2L is the top channel for all four setups. This consistency demonstrates that the detection mechanism operates through global modal changes rather than local sensor responses, and that the identified features reflect the bridge's genuine structural behaviour rather than setup-specific artefacts.

4.6 Settlement Progression

To understand and address the evolution of feature importance with increasing damage severity, the SHAP pipeline was applied to Setup 09 across the four pier settlement scenarios (S03–S06). Table 4.5 summarises the results.

Table 4.5: SHAP settlement progression for Setup 09

Scenario	Damaged score	Surrogate R^2	Top Feature	Top Channel
S03	0.208	0.620 ± 0.114	R3V_over_R2V_peak_freq	R2T
S04	0.146	0.512 ± 0.130	R3V_over_R2V_peak_freq	R3V
S05	0.275	0.792 ± 0.064	R3V_over_R2V_peak_freq	R2L
S06	0.317	0.750 ± 0.150	R3V_peak_freq	R3V

Three key findings can be derived from the Table above. First, surrogate fidelity is adequate ($R^2 \geq 0.70$) only from 80mm settlement onward. At 20mm and 40mm, the surrogate cannot faithfully reconstruct the autoencoder's scoring behaviour from the 92 structural features. This implies that the model detects some deviation at low settlement, but that deviation does not map cleanly onto interpretable structural features.

Second, the anomaly score progression is non-monotonic (variable) at low settlement. This is notable as S03 (20mm) scores higher than S04 (40mm) despite representing less structural damage. This is consistent with the ambient excitation variation across measurement days previously mentioned in Section 3.1.3, which affects absolute score magnitudes even after normalisation.

Third, the top feature changes qualitatively between S05 and S06. At 80mm, the transmissibility ratio $R3V\_over\_R2V\_peak\_freq$ dominates, reflecting how vibration propagates differently between sensor locations under moderate settlement. At 95mm, this gives way to $R3V\_peak\_freq$ and $R2V\_peak\_freq$, which are direct modal frequency features, suggesting that at maximum settlement, the pier displacement is large enough to produce a detectable shift in the resonant frequencies themselves, beyond the earlier noted transmissibility redistribution effect.

5 Discussion

This chapter explains the key findings from the results presented in Chapter 4 and compares the work in this thesis with other related work on the Z24 bridge benchmark. The chapter begins by providing responses to the research questions established in Section 1.3 to achieve the aim of this study and discuss the implications of each outcome. This is followed by a comparison of the experimental results of this study with prior work on the Z24 bridge dataset. The chapter concludes by acknowledging and stating the limitations of this study.

5.1 Research Question Responses

RQ1: Can a Conv1D autoencoder utilizing frequency-domain of the ambient vibration data reliably detect Koppigen pier settlement across all spatial measurement zones of the Z24 benchmark?

Yes. The proposed Conv1D autoencoder trained on log-PSD AVT data reliably detected the Koppigen pier settlement across all nine setups, achieving AUROC > 0.95 with a mean of 0.987, thereby answering RQ1 affirmatively. The model successfully separates the damaged windows of Scenario 05 from the healthy baseline, regardless of where the sensors are positioned on the bridge. At the p99 threshold, Setup 09 achieves a recall of 1.000 and an F1 of 0.984, demonstrating that the model can maintain near-perfect recall at a very low false alarm rate (1.2%) in the best-performing measurement setup.

The baseline comparison in Section 4.3.2 reveals that PCA reconstruction error, especially with high component configurations, achieves superior AUROC to the proposed Conv1D autoencoder on the log-PSD representation of the Z24 data across all three settlement scenarios. This finding is consistent with other SHM and Z24 literature suggesting that frequency-domain representations are fundamentally discriminative for structural damage detection, and that the detection task on this benchmark can be achieved by multiple methods, including simple linear models (Abdrabo 2024; Buckley et al. 2023; Peeters and De Roeck 2001). The limitation of PCA reconstruction as an anomaly detector is not its performance on the spectral representation of this specific dataset, but

interpretability. The PCA components are just linear mixtures of raw frequency bins that are not related to any physically meaningful structural quantities and cannot be integrated into a SHAP attribution pipeline (Adadi and Berrada 2018; Lundberg and Lee 2017). Therefore, the key contribution of the proposed approach is not achieving the highest AUROC on this well-studied benchmark, but rather providing a detection pipeline that simultaneously flags anomalies and explains which structural features changed, which sensor locations were most affected, and how the explanation evolves with increasing damage severity. These capabilities cannot be offered by a reconstruction-based linear method.

RQ2: Which vibration features most strongly explain the autoencoder's anomaly scores, and do they correspond to physically interpretable structural phenomena?

The dominant explanatory feature is $R3V\_over\_R2V\_peak\_freq$, which is a transmissibility ratio between the Utzenstorf-side vertical sensor (R3V) and the mid-span vertical reference (R2V). This feature is consistent across three of the four independently trained models spanning both sides of the bridge, and R2L (mid-span lateral) is the top channel in all four. This would have been naturally expected to be R1V (physically nearest the Koppigen pier). The physical interpretation is that the pier settlement at Koppigen redistributes bending moments along the span, thereby altering the distribution and transmission of modal energy between the sensors, and not local pier response. The transmissibility ratio captures this redistribution more sensitively than any single-sensor feature, because it encodes the relative change in modal tuning between the far sensor and the reference rather than just the absolute amplitude. RQ2 is therefore also answered affirmatively, that the features are both statistically dominant and physically interpretable.

RQ3: How do SHAP-attributed feature importances evolve as Koppigen pier settlement progresses from 20 mm to 95 mm, and does this progression reveal a quantifiable threshold for damage severity?

At 20mm and 40mm settlement, the surrogate fidelity falls below 0.70 (with surrogate R^2 of 0.620 and 0.512, respectively), indicating that the structural signal is too weak

relative to ambient excitation variability for the feature-based surrogate to reliably explain the autoencoder's outputs. However, from 80mm settlement onwards, the surrogate fidelity reaches 0.79, and the feature hierarchy stabilises around $R3V\_over\_R2V\_peak\_freq$. At 95mm settlement, a qualitative shift occurs toward direct modal frequency features ($R3V\_peak\_freq$, $R2V\_peak\_freq$), implying a change in the structural response mechanism at the maximum settlement. The empirical explainability threshold is therefore *80mm*, which is the minimum settlement at which SHAP attributions can be considered reliable for this dataset and pipeline.

5.2 Comparison with Prior Z24 Work

Unsupervised and one-class learning methods have been applied in recent studies on the Z24 bridge dataset, providing a basis for contextual comparison of their findings with those in this study.

Giglioni et al. (2023) proposed a detection framework based on an autoencoder trained on raw time-domain acceleration data, and used reconstruction error as the anomaly indicator. Their work demonstrated effectiveness in local damage detection on the Z24 benchmark. However, working directly in the time domain exposed the reconstruction loss to phase variability, which is inherent in ambient vibration signals. Moreover, the method provides no feature-level explanations for the anomalies detected.

Pakzad and Masoodi (2025) also evaluated six VAE-based hybrid frameworks on the Z24 dataset. They found that integrating a variational autoencoder with a one-class support vector machine (VAE-OCSVM) yielded the best overall performance across both operational and environmental variability. Their work addresses missing data, a concern absent from benchmark studies, and underscores the importance of combining classical anomaly detectors with deep latent representations. However, they didn’t address the explanation of the latent representation.

The study by Abdrabo (2024) applied an online one-class classification method on the same Z24 short-term monitoring data, evaluating elliptic envelope, incremental SVC, local outlier factor, half-space trees, and entropy-guided envelopes. The work demonstrated that XGBoost-based feature selection helped to identify a set of discriminative

features and introduced an entropy-based drift-management strategy to address the challenges posed by gradual changes in structural state during continuous monitoring. This study is the closest prior work that uses a one-class constraint on the Z24 PDT data, but no reconstruction-based scoring, scenario-specific evaluation, or post-hoc feature attribution was performed.

Ni et al. (2025) proposed a TCN-GAT autoencoder that combines temporal convolution with graph attention to explicitly model spatiotemporal coupling across the sensor network. Their method processes raw vibration signals end-to-end without frequency-domain preprocessing and achieves effective anomaly detection on the Z24 benchmark. However, the spatial topology is incorporated directly into the model architecture and not examined post-hoc through feature attribution.

Nesackon et al. (2026) combined PCA and autoencoder reconstruction in an ensemble framework applied to modal frequencies extracted via stochastic subspace identification. This led to improved stability compared to a baseline density estimation method. Their work confirms that ensemble reconstruction approaches can outperform single-method detectors under environmental variability, and that using modal frequencies rather than raw signals provides inherent phase invariance.

Across these studies, it is clear that detection capabilities have already been established for the Z24 benchmark under a one-class constraint using different architectures. One thing that remains missing in the literature is the mechanism for explaining the structural feature responsible for the anomaly scores and how the explanation changes with increasing damage severity. The work in this thesis directly addresses this gap.

5.3 Limitations of Thesis

Although this thesis has achieved its aim, the study still has limitations, some of which are outlined below:

1. **Single damage type:** All results are validated only for pier settlement. This study does not establish whether the same pipeline can generalize to other Z24 damage types (such as anchor head failures, tendon ruptures, concrete spalling), to bridges with different structural configurations, or to real instrumented bridges

under continuous long-term monitoring. The reported AUROC values only represent performance under idealized benchmark conditions with respect to the dataset, and the operational robustness of the pipeline on real bridges is unvalidated.

2. **Model complexity:** The Conv1D autoencoder contains 1,187,461 trainable parameters, trained on approximately 216 healthy windows per setup. Given the high parameter-to-sample ratio, there is concern about overfitting to the healthy training distribution. Even though the model was trained only on healthy data in a one -class setting, where overfitting implies a poor generalization to unseen healthy data rather than damaged ones, and early stopping with patience of 15 epochs was implemented, the possibility of the model memorising some characteristics of the training data cannot be ruled out.
3. **Explainability threshold limitation:** The surrogate-based SHAP approach only provides reliable attributions from 80mm settlement onwards. Below this threshold, the model detects anomalies it cannot explain through structural features, limiting interpretability for early-stage damage. The surrogate method creates a gap, as the XGBoost model only approximates but does not perfectly reproduce the autoencoder's scoring behaviour. Therefore, any interpretation of the SHAP attribution is an interpretation of the surrogate and not a direct explanation of the neural network.
4. **Offline operation mode:** The pipeline is trained and evaluated on pre-recorded data. Deploying the pipeline for real-time monitoring would require incremental scaler update and online threshold recalibration, as the healthy structural state won't be constant but will change with temperature and traffic loading patterns.
5. **Limited dataset scope:** The Z24 PDT campaign used for this study spans only 36 days. Considering long-term environmental variability, such as seasonal temperature cycles, and progressive creep would require the healthy training window to cover at least one full annual cycle, as noted by Peeters and De Roeck (2001). Also, the features and threshold identified in this study are specific to Z24 and cannot be assumed to transfer to other bridge types.

6 Conclusion and future work

This chapter combines the main elements of this thesis together and presents the conclusion, summarizing all key findings, contributions, limitations, and recommending directions for future work in line with this study.

6.1 Summary and Key Findings

This thesis developed and evaluated an explainable one-class anomaly detection pipeline for vibration-based SHM of bridge infrastructure. The developed pipeline employed a Conv1D autoencoder trained on log-PSD representation of the ambient acceleration data from five reference channels (R1V, R2L, R2T, R2V, and R3V) using the popular Z24bridge PDT benchmark dataset. This resulted in training nine independent models, which were evaluated against the 80mm Koppigen pier settlement (Scenario 05). A TreeSHAP pipeline was employed to provide post-hoc explainability using an XGBoost surrogate that was trained to approximate the autoencoder's anomaly scores from 92 normalised spectral features, to enable per-feature attribution without modifying the detection model.

All three research questions posed in Chapter 1 are answered affirmatively. For RQ1, the pipeline achieved AUROC above 0.95 across all nine measurement setups (mean of 0.987), confirming reliable detection of the pier settlement. Setup 09 achieved perfect recall with a strict FAR of 1.2%. For RQ2, the dominant explanatory feature across three of four independently trained models was $R3V\_over\_R2V\_peak\_freq$, a transmissibility ratio between the Utzenstorf-side vertical sensor and the mid-span vertical reference, with R2L identified as the most important channel. Both findings correspond to global modal redistribution under the pier settlement. Finally, for RQ3, the surrogate fidelity analysis established 80mm as the empirical explainability threshold, below which structural signals were insufficient for reliable SHAP attribution, while a shift in dominant features at 95mm settlement indicated the onset of detectable resonant frequency changes.

6.2 Summary of Contributions

This thesis contributed the following to SHM studies: (1) a phase-invariant one-class detection pipeline using log-PSD representations; (2) a SHAP explainability framework via XGBoost surrogate with per-setup normalisation; (3) an experimental characterisation of the Z24 PDT benchmark across all nine setups under a one-class constraint; and (4) a surrogate fidelity analysis establishing an empirical explainability threshold for progressive foundation damage. The developed pipeline can be integrated into a digital twin for monitoring infrastructure in smart cities.

6.3 Summary of Limitations

Five key limitations have been identified for this work. First, all results are validated only on a single damage type (Koppigen pier settlement), and this study does not establish generalization to other Z24 damage scenarios or to other bridges. Secondly, there is concern about potential overfitting due to the high parameter-to-sample ratio in the Conv1D autoencoder. Future studies should investigate architectures with fewer parameters. Third, reliable SHAP attribution was achieved only from 80mm settlement onwards. Below this threshold, the model detects anomalies it cannot explain through the 92-feature spectral representation. Fourthly, the pipeline operates entirely offline on pre-recorded data; real-time deployment would require incremental scaler updates and online threshold recalibration to account for operational and environmental variability. Finally, the Z24 PDT campaign spans only 36 days, and studying long-term seasonal and creep effects would require a healthy training window covering at least one full annual cycle.

6.4 Recommendations for Future Work

Following the contributions and limitations, the following directions are recommended for extending this work:

- **Generalization to other damage types:** Future work should evaluate the pipeline against the remaining Z24 PDT scenarios (anchor head failures, tendon ruptures,

etc.), to establish whether the same spectral features and detection thresholds generalize beyond the pier settlement.

- **Online and continuous monitoring:** Extending the framework to a streaming setting with both healthy baseline updating and adaptive threshold recalibration would move the pipeline closer to practical deployment on instrumented bridges.
- **Using directly interpretable architecture:** Future research should consider replacing the two-stage surrogate approach employed in this study with directly interpretable models, like sparse autoencoder or prototype-based network. This should eliminate the fidelity gap between the anomaly score and the surrogate prediction.
- **Validation on additional bridge datasets:** By applying the pipeline to other SHM benchmark datasets, future work can evaluate its robustness across different structural types, sensor configurations, and environmental conditions.

.

References

Abdrabo, Amro. 2024. “Application of Online Anomaly Detection Using One-Class Classification to the Z24 Bridge.” *Sensors* 24(23):7866. doi:10.3390/s24237866.

Adadi, Amina, and Mohammed Berrada. 2018. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).” *IEEE Access* PP:1–1. doi:10.1109/ACCESS.2018.2870052.

Aktan, Emin, Ivan Bartoli, Branko Glišić, and Carlo Rainieri. 2024. “Lessons from Bridge Structural Health Monitoring (SHM) and Their Implications for the Development of Cyber-Physical Systems.” *Infrastructures* 9(2):30. doi:10.3390/infrastructures9020030.

An, Jinwon, and Sungzoon Cho. 2015. “Variational Autoencoder Based Anomaly Detection Using Reconstruction Probability.”

Bayane, Imane, John Leander, and Raid Karoumi. 2024. “An Unsupervised Machine Learning Approach for Real-Time Damage Detection in Bridges.” *Engineering Structures* 308:117971. doi:10.1016/j.engstruct.2024.117971.

Benfenati, Luca, Daniele Jahier Pagliari, Luca Zanatta, Yhorman Alexander Bedoya Velez, Andrea Acquaviva, Massimo Poncino, Enrico Macii, Luca Benini, and Alessio Burrello. 2025. “Foundation Models for Structural Health Monitoring.” *IEEE Transactions on Sustainable Computing* 10(6):1103–15. doi:10.1109/TSUSC.2025.3592097.

Brincker, Rune(Author), and Carlos Ventura. 2015. *Introduction to Operational Modal Analysis*. John Wiley & Sons.

Buckley, Tadhg, Bidisha Ghosh, and Vikram Pakrashi. 2023. “A Feature Extraction & Selection Benchmark for Structural Health Monitoring.” *Structural Health Monitoring* 22(3):2082–2127. doi:10.1177/14759217221111141.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009. “Anomaly Detection: A Survey.” *ACM Comput. Surv.* 41. doi:10.1145/1541880.1541882.

Chen, Mingyang, Jingzhou Xin, Qizhi Tang, Tianyu Hu, Yin Zhou, and Jianting Zhou. 2024. “Explainable Machine Learning Model for Load-Deformation Correlation in Long-Span Suspension Bridges Using XGBoost-SHAP.” *Developments in the Built Environment* 20:100569. doi:10.1016/j.dibe.2024.100569.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” Pp. 785–94 in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Dabbous, Ali, Riccardo Berta, Matteo Fresta, Hadi Ballout, Luca Lazzaroni, and Francesco Bellotti. 2024. “Bringing Intelligence to the Edge for Structural Health Monitoring: The Case Study of the Z24 Bridge.” *IEEE Open Journal of the Industrial Electronics Society* 5:781–94. doi:10.1109/OJIES.2024.3434341.

Davis, Jesse, and Mark Goadrich. 2006. *The Relationship Between Precision-Recall and ROC Curves*. Vol. 06.

Eltouny, Kareem, Mohamed Gomaa, and Xiao Liang. 2023. “Unsupervised Learning Methods for Data-Driven Vibration-Based Structural Health Monitoring: A Review.” *Sensors* 23(6):3290. doi:10.3390/s23063290.

Fang, Youliang, Chanpeng Li, and Jiaxin Li. 2025. “Structural Damage Identification Based on Transfer Learning and Power Spectral Density.” *Structural Control and Health Monitoring* 2025(1):5224063. doi:10.1155/stc/5224063.

Faris, Nour, Tarek Zayed, and Ali Fares. 2025. “Review of Condition Rating and Deterioration Modeling Approaches for Concrete Bridges.” *Buildings* 15(2):219. doi:10.3390/buildings15020219.

Favarelli, Elia, and Andrea Giorgetti. 2021. “Machine Learning for Automatic Processing of Modal Analysis in Damage Detection of Bridges.” *IEEE Transactions on Instrumentation and Measurement* 70:1–13. doi:10.1109/TIM.2020.3038288.

Giglioni, Valentina, Ilaria Venanzi, Valentina Poggioni, Alfredo Milani, and Filippo Ubertini. 2023. “Autoencoders for Unsupervised Real-Time Bridge Health Assessment.” *Computer-Aided Civil and Infrastructure Engineering* 38(8):959–74. doi:10.1111/mice.12943.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Adaptive Computation and Machine Learning. Cambridge, Mass: The MIT press.

Hassan Daneshvar, Mohammad, and Hassan Sarmadi. 2022. “Unsupervised Learning-Based Damage Assessment of Full-Scale Civil Structures under Long-Term and Short-Term Monitoring.” *Engineering Structures* 256:114059. doi:10.1016/j.engstruct.2022.114059.

Hinton, G. E., and R. R. Salakhutdinov. 2006. “Reducing the Dimensionality of Data with Neural Networks.” *Science* 313(5786):504–7. doi:10.1126/science.1127647.

Javadinasab Hormozabad, Sajad, Mariantonieta Gutierrez Soto, and Hojjat Adeli. 2021. “Integrating Structural Control, Health Monitoring, and Energy Harvesting for Smart Cities.” *Expert Systems* 38(8):e12845. doi:10.1111/exsy.12845.

Jiang, Yali, Gang Yang, Haijiang Li, and Tian Zhang. 2023. “Knowledge Driven Approach for Smart Bridge Maintenance Using Big Data Mining.” *Automation in Construction* 146:104673. doi:10.1016/j.autcon.2022.104673.

Kaewunruen, Sakdirat, Jessada Sresakoolchai, Wentao Ma, and Olisa Phil-Ebosie. 2021. “Digital Twin Aided Vulnerability Assessment and Risk-Based Maintenance Planning of Bridge Infrastructures Exposed to Extreme Conditions.” *Sustainability* 13(4):2051. doi:10.3390/su13042051.

Kim, Sunjoong, Hyemin Kim, and Youchan Hwang. 2025. “Data-Driven Dynamic Response Forecasting and Anomaly Detection in Long-Span Bridges.” *Journal of Civil Structural Health Monitoring* 15(7):3045–62. doi:10.1007/s13349-025-00952-8.

Krämer, C., CAM De Smet, and Guido De Roeck. 1999. “Z24 Bridge Damage Detection Tests.” Pp. 1023–29 in *IMAC 17, the International modal analysis conference*. Vol. 3727. Society of Photo-optical Instrumentation Engineers.

Lundberg, Scott, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.”

Maeck, J., and G. De Roeck. 2003. “DESCRIPTION OF Z24 BENCHMARK.” *Mechanical Systems and Signal Processing* 17(1):127–31. doi:10.1006/mssp.2002.1548.

Malekloo, Arman, Ekin Ozer, Mohammad AlHamaydeh, and Mark Girolami. 2022. “Machine Learning and Structural Health Monitoring Overview with Emerging Technology and High-Dimensional Data Source Highlights.” *Structural Health Monitoring* 21(4):1906–55. doi:10.1177/14759217211036880.

Matos, José C., Vanni Nicoletti, Jakub Kralovanec, Hélder S. Sousa, Fabrizio Gara, Martin Moravcik, and Maria J. Morais. 2023. “Comparison of Condition Rating Systems for Bridges in Three European Countries.” *Applied Sciences* 13(22):12343. doi:10.3390/app132212343.

Mehrjoo, Azin, Kyle L. Hom, Homayoon Beigi, and Raimondo Betti. 2025. “Zero-Shot Bridge Health Monitoring Using Cepstral Features and Streaming LSTM Networks.” *Infrastructures* 10(11):292. doi:10.3390/infrastructures10110292.

Nesackon Abraham, Jabez, Minh Q. Tran, Jerusha Samuel Jayaraj, Jose C. Matos, Maria Rosa Valluzzi, and Son N. Dang. 2026. “Unsupervised Learning-Based Anomaly Detection for Bridge Structural Health Monitoring: Identifying Deviations from Normal Structural Behaviour.” *Sensors* 26(2):561. doi:10.3390/s26020561.

Ni, Yanchun, Qiyuan Jin, and Rui Hu. 2025. “A Novel Unsupervised Structural Damage Detection Method Based on TCN-GAT Autoencoder.” *Sensors* 25(21):6724. doi:10.3390/s25216724.

Noel, Adam B., Abderrazak Abdaoui, Tarek Elfouly, Mohamed Hossam Ahmed, Ahmed Badawy, and Mohamed S. Shehata. 2017. “Structural Health Monitoring Using Wireless Sensor Networks: A Comprehensive Survey.” *IEEE Communications Surveys & Tutorials* 19(3):1403–23. doi:10.1109/COMST.2017.2691551.

Okur, Fatih, Ahmet Altunişik, and Ebru Kalkan Okur. 2025. “A Novel Approach for Anomaly Detection in Vibration-Based Structural Health Monitoring Using Autoencoders in Deep Learning.” *Structural Control and Health Monitoring* 2025. doi:10.1155/stc/5602604.

Pakzad, Seyed Soroush, and Amir R. Masoodi. 2025. “Early Damage Detection in Bridges Using a Variational Autoencoder–Based Hybrid Unsupervised Learning Framework.” *Scientific Reports* 16(1):1389. doi:10.1038/s41598-025-31115-w.

Peeters, Bart, and Guido De Roeck. 2001. “One-Year Monitoring of the Z24-Bridge: Environmental Effects versus Damage Events.” *Earthquake Engineering & Structural Dynamics* 30(2):149–71. doi:10.1002/1096-9845(200102)30:2<149::AID-EQE1>3.0.CO;2-Z.

Plevris, Vagelis, and George Papazafeiropoulos. 2024. “AI in Structural Health Monitoring for Infrastructure Maintenance and Safety.” *Infrastructures* 9(12):225. doi:10.3390/infrastructures9120225.

Rebenciuc, Mihai, Simona Mihaela Bibic, and Antonela Toma. 2021. “Assessment of Structural Monitoring by Analyzing Some Modal Parameters: An Extended Inventory of Methods and Developments.” *Archives of Computational Methods in Engineering* 28(3):1575–90. doi:10.1007/s11831-020-09433-1.

Reynders, Edwin, and Guido De Roeck. 2008. “Continuous Vibration Monitoring and Progressive Damage Testing on the Z 24 Bridge.” edited by C. Boller, F. Chang, and Y. Fujino. Wiley.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.”

Rytter, Anders. 1993. “Vibrational Based Inspection of Civil Engineering Structures.”

Saito, Takaya, and Marc Rehmsmeier. 2015. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets” edited by G. Brock. *PLOS ONE* 10(3):e0118432. doi:10.1371/journal.pone.0118432.

Salave, Harshwardhan. 2025. “IoT-Based Bridge Monitoring System.” *International Journal for Research in Applied Science and Engineering Technology* 13(4):4442–45. doi:10.22214/ijraset.2025.69099.

Santaniello, Pasquale, and Paolo Russo. 2023. “Bridge Damage Identification Using Deep Neural Networks on Time–Frequency Signals Representation.” *Sensors* 23(13):6152. doi:10.3390/s23136152.

Santos-Vila, Ivan, Ricardo Soto, Emanuel Vega, Alvaro Peña Fritz, and Broderick Crawford. 2024. “Anomaly Detection on Bridges Using Deep Learning With Partial Training.” *IEEE Access* 12:116530–45. doi:10.1109/ACCESS.2024.3447571.

Sivasuriyan, Arvindan, Dhanasingh Sivalinga Vijayan, Anna Piętocha, Wojciech Górski, Łukasz Wodzyński, and Eugeniusz Koda. 2026. “Advanced Sensing and Digital Monitoring Technologies for Structural Health Assessment of Civil Infrastructure.” *Buildings* 16(3):656. doi:10.3390/buildings16030656.

Sohn, Hoon. 2006. “Effects of Environmental and Operational Variability on Structural Health Monitoring.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. doi:10.1098/rsta.2006.1935.

Tang, Yunchao, Shuai Wan, Qingying Yang, Zheng Chen, and Yang Xu. 2025. “A Review: Research Progress in Bridge Structural Health Monitoring From the Perspective of AI Development.” *Structural Control and Health Monitoring* 2025(1):8870840. doi:10.1155/stc/8870840.

Tax, David, and Robert Duin. 2004. “Support Vector Data Description.” *Machine Learning* 54:45–66. doi:10.1023/B:MACH.0000008084.60811.49.

Welch, P. 1967. “The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short, Modified Periodograms.” *IEEE Transactions on Audio and Electroacoustics* 15(2):70–73. doi:10.1109/TAU.1967.1161901.

Yang, Linfei, and Enzo Huang. 2024. “Structural Health Monitoring Data Analysis Using Deep Learning Techniques.” Pp. 913–27 in *Proceedings of the 2024 8th International Conference on Electronic Information Technology and Computer Engineering*. Haikou Guangdong China: ACM.

Yu, Xiao, Yuguang Fu, Jian Li, Jianxiao Mao, Tu Hoang, and Hao Wang. 2024. “Recent Advances in Wireless Sensor Networks for Structural Health Monitoring of Civil Infrastructure.” *Journal of Infrastructure Intelligence and Resilience* 3(1):100066. doi:10.1016/j.iintel.2023.100066.

Z24 Bridge benchmark. n.d. Retrieved May 24, 2026. https://bwk.kuleuven.be/bwm/z24.

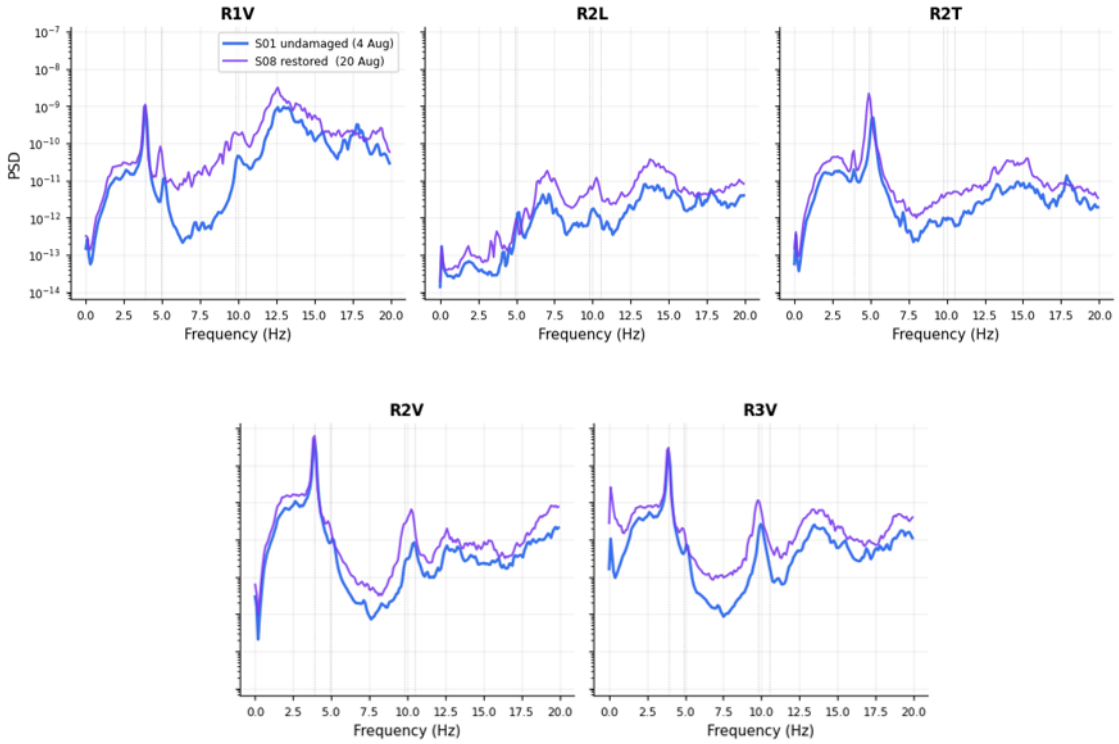
Zeng, Fan, Chuan Pang, and Huajun Tang. 2024. “Sensors on Internet of Things Systems for the Sustainable Development of Smart Cities: A Systematic Literature Review.” *Sensors* 24(7):2074. doi:10.3390/s24072074.

Zhang, Xiulin, and Wensong Zhou. 2023. “Structural Vibration Data Anomaly Detection Based on Multiple Feature Information Using CNN-LSTM Model.” *Structural Control and Health Monitoring* 2023(1):3906180. doi:10.1155/2023/3906180.

Zinno, Raffaele, Sina Shaffiee Haghshenas, Giuseppe Guido, and Alessandro VItale. 2022. "Artificial Intelligence and Structural Health Monitoring of Bridges: A Review of the State-of-the-Art." *IEEE Access* 10:88058–78. doi:10.1109/ACCESS.2022.3199443.

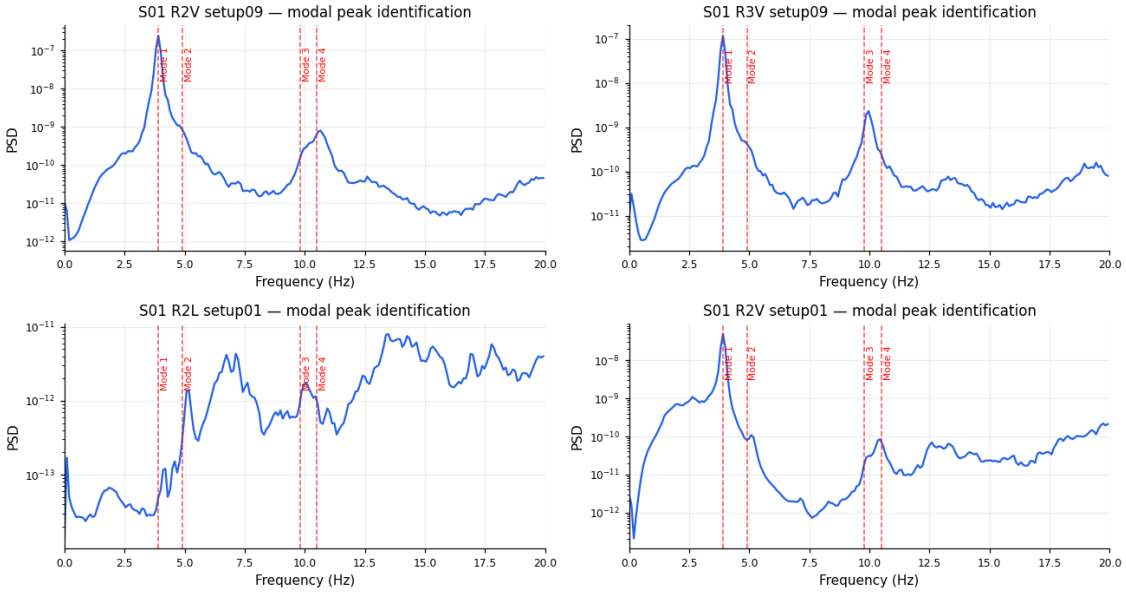
Appendices

Appendix 1. S01 vs S08 modal frequency comparison



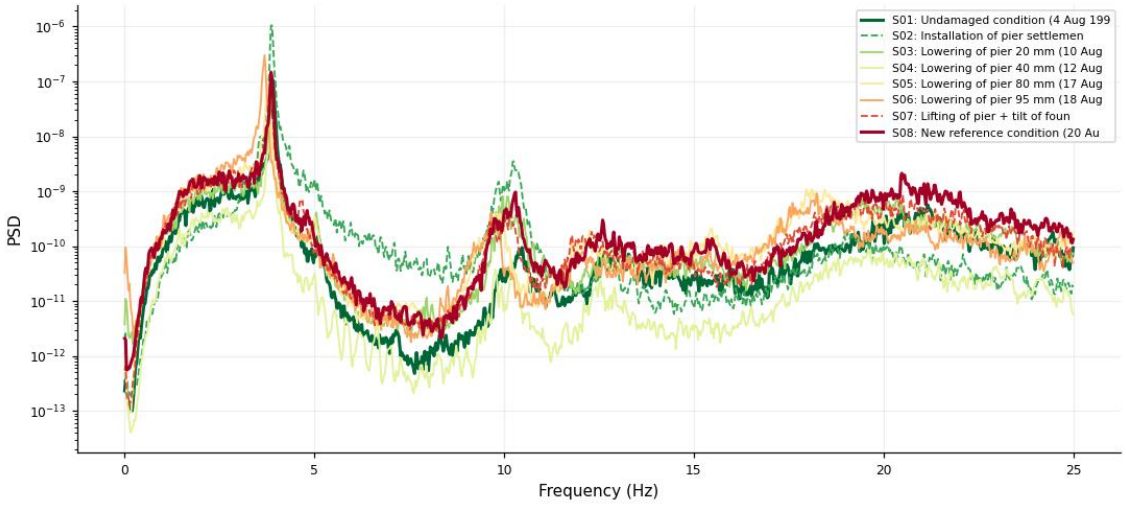
Mean log-PSD of all five reference channels for Scenario 01 (4 August, healthy baseline) and Scenario 08 (20 August, restored condition), plotted on the same axes. Shows modal peak alignment for R1V, R2T, R2V, R3V, and the ~0.5 Hz offset in R2L. This is the spectral equivalence justification for using S08 in threshold calibration.

Appendix 2. Reference channel PSDs for S01 showing Z24 natural frequencies



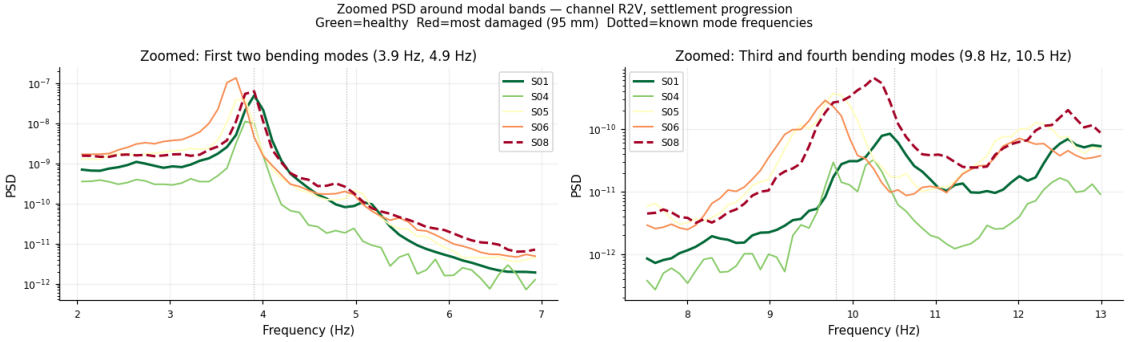
Mean log-PSD for each of the five reference channels (R1V, R2L, R2T, R2V, R3V) from S01, with the four known Z24 natural frequencies (3.9, 4.9, 9.8, 10.5 Hz) marked as vertical dashed lines. Confirms that the chosen frequency range and $nperseg = 1024$ resolution resolves all four structural modes.

Appendix 3. Settlement progression PSD (S03–S06 vs S01)



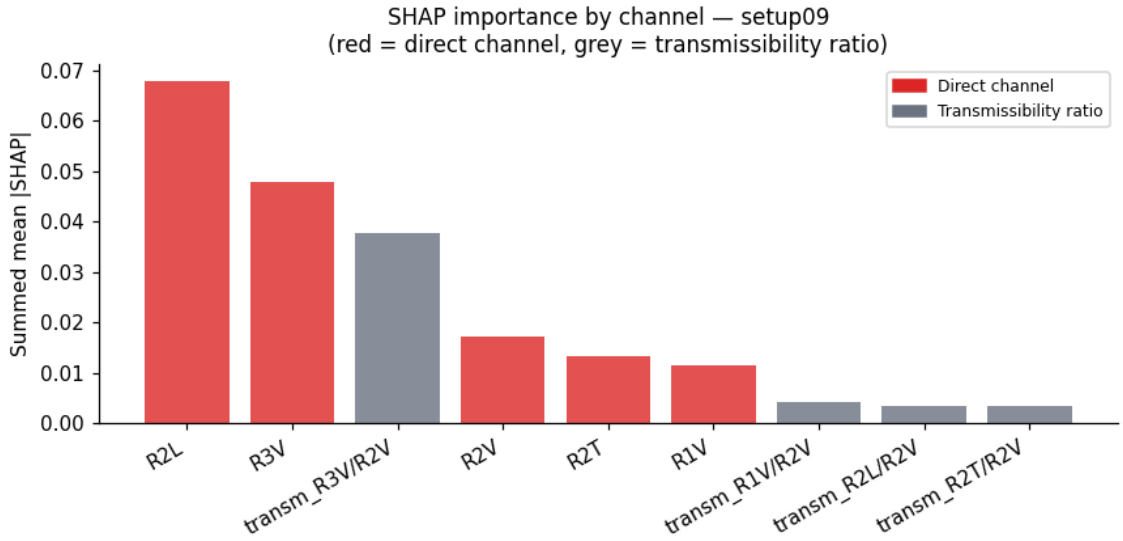
Mean log-PSD of R2V for the four pier settlement scenarios (S03, S04, S05, S06) overlaid against S01, showing non-monotonic amplitude ordering and the progressive spectral deviation from 40mm onwards.

Appendix 4. Zoomed PSD: modal bands under settlement



Zooming into the 2–7 Hz range (first two bending modes) and 7–13 Hz range (third and fourth modes) for the settlement progression makes the frequency shifts directly visible. The broad PSD view in Figure A.3 above compresses these into a single indistinguishable bump.

Appendix 2. SHAP channel importance, Setup 09, S05



Quantitative breakdown of summed mean absolute SHAP values per reference channel, confirming R2L as the dominant channel. Referenced from Section 4.4 with parentheses.