processes

MDPI

Article

Self-Supervised CNN–Transformer Anomaly Detection for Bearing Health Monitoring

Syed Haseeb Haider Zaidi [1,*], Alex Shenfield [1,*], Hongwei Zhang [1] and Augustine Ikpehai [2]

[1] Advanced Food Innovation Centre (AFIC), Sheffield Hallam University, Sheffield S9 2AA, UK; h.zhang@shu.ac.uk
[2] School of Engineering and Built Environment, Sheffield Hallam University, Sheffield S1 1WB, UK; a.ikpehai@shu.ac.uk
* Correspondence: h.zaidi@shu.ac.uk (S.H.H.Z.); a.shenfield@shu.ac.uk (A.S.)

Abstract

Reliable bearing fault detection is essential for predictive maintenance in industrial systems; however, obtaining labelled fault data is often expensive, time-consuming, and impractical in real-world deployments. To address this challenge, this study proposes a healthy-only self-supervised anomaly detection framework for bearing health monitoring using vibration measurements. The proposed approach combines convolutional neural networks and Transformer-based temporal modelling to learn informative representations from healthy vibration signals without requiring fault labels during representation learning. Three self-supervised learning strategies—reconstruction-based, contrastive, and a unified contrastive–reconstruction objective—are investigated to evaluate the effectiveness of different representation learning approaches. The learned latent representations are subsequently analysed using Isolation Forest and Mahalanobis-distance anomaly scoring methods. To provide a realistic assessment of generalisation, a strict grouped cross-validation protocol is employed, where data are partitioned at the sample level to prevent information leakage between training and testing sets. Furthermore, prevalence-aware experiments are conducted under 5% and 10% fault prevalence scenarios to assess deployment robustness. Experimental results on the Paderborn bearing dataset demonstrate that the proposed CNN + Transformer model trained with combined contrastive and reconstruction objectives and evaluated using Isolation Forest achieves the best overall performance, obtaining a ROC-AUC of 0.878 ± 0.015, a PR-AUC of 0.958 ± 0.005, and an F1-score of 0.590 ± 0.040. The results consistently outperform classical feature-based approaches, One-Class SVM, and autoencoder baselines. Ablation analysis further shows that combining contrastive and reconstruction objectives produces more informative representations than either objective alone. The findings demonstrate that the proposed healthy-only self-supervised framework provides an effective and label-efficient approach for rolling bearing anomaly detection and shows promise for predictive maintenance applications where labelled fault data are limited or unavailable.

Keywords: bearing anomaly detection; self-supervised learning; contrastive learning; healthy-only learning; one-class learning; Transformer; convolutional neural network; predictive maintenance; vibration monitoring; Isolation Forest

Check for updates

Academic Editor: José de Jesús Rangel Magdaleno

Received: 22 June 2026
Revised: 17 July 2026
Accepted: 24 July 2026
Published: 30 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Breakdown represents a functional deterioration of a machine or equipment that progressively reduces performance and may ultimately lead to operational failure [1]. In

industrial environments, unexpected equipment failures cause production interruptions, increased maintenance costs, and potential safety risks [2]. Modern manufacturing systems therefore increasingly emphasize maintenance strategies that move beyond reactive repair toward early fault detection and condition-based intervention. In this context, failure is not limited to total system stoppage but also includes the inability to meet required quality or productivity targets [3]. Consequently, predictive maintenance has become a central objective in intelligent and data-driven industrial operations. Machine learning has emerged as a transformative technology in predictive maintenance, enabling automated health assessment, fault classification, and remaining useful life estimation, thereby improving equipment reliability and reducing unplanned downtime [4,5].

Among mechanical components, rolling bearings are particularly critical elements in machinery. They support both static and dynamic loads while enabling smooth motion, and they account for a substantial proportion of machine breakdowns [6]. Under prolonged high-speed operation and harsh working conditions, bearings are especially susceptible to progressive degradation and sudden failure, often resulting in significant economic losses and operational disruption [7]. Therefore, accurate and timely fault diagnosis of rolling bearings remains a problem of considerable practical and theoretical importance [8].

Vibration analysis has become one of the most widely adopted techniques for bearing condition monitoring. Even during normal operation, rotating machinery generates vibration signals due to shaft rotation, gear meshing, fluid flow, and electromagnetic interactions that produce periodic magnetic forces and torque ripple in electric machines. These signals contain characteristic patterns that reflect the internal mechanical state of the system. Vibration-based monitoring is particularly advantageous because it responds immediately to structural changes, supports continuous or intermittent measurement, and frequently enables localization of faulty components [9]. As a result, vibration signals provide a rich source of information for early-stage fault detection.

Traditional vibration-based diagnosis methods rely on handcrafted statistical and signal-processing features extracted from measured vibration signals [10,11]. Both time-domain descriptors (e.g., RMS, kurtosis, skewness, crest factor) and frequency-domain features derived from spectral and envelope analysis are widely used to characterize impulsive behaviors and characteristic fault frequencies associated with bearing defects [12]. These features are typically combined with classical machine learning algorithms such as SVM or kNN for classification or anomaly detection [11]. However, such approaches depend heavily on domain expertise and manual feature engineering, and their performance may degrade under varying operating conditions or distribution shifts [10,13]. To overcome these limitations, deep learning techniques have increasingly been adopted for automated feature extraction and fault classification [14,15]. Convolutional Neural Networks (CNNs) automatically learn hierarchical features from raw vibration signals through convolution and pooling operations, eliminating the need for manual feature design [2]. Recurrent Neural Networks (RNNs), including LSTM and GRU variants, have been widely applied to sequential vibration analysis; however, their recurrent structure limits parallelization and may hinder the modeling of very long-range dependencies. Transformer architectures, based on self-attention mechanisms, alleviate these limitations by enabling direct global context modeling without relying on recurrence [16]. Despite their strong performance, most deep learning approaches for bearing fault diagnosis rely on fully supervised learning paradigms and require large quantities of labeled fault data.

In real industrial environments, however, fault data are inherently scarce. Failures occur infrequently, and collecting labeled fault samples can be costly, time-consuming, and sometimes unsafe. Operational datasets are typically dominated by healthy signals, with fault instances representing only a small fraction of total data. These limitations motivate

the investigation of learning paradigms that can operate effectively with little or no labeled fault data.

Anomaly detection offers an alternative framework that requires only normal data during training and identifies abnormal behavior as deviations from learned patterns [17]. Anomalies may appear as point anomalies—individual abnormal observations—or as collective anomalies, where a sequence of data becomes abnormal as a whole [18]. Recent deep learning-based methods have demonstrated superior performance in complex real-world scenarios [19]. Nevertheless, effective anomaly detection for industrial monitoring systems requires models capable of detecting subtle early-stage deviations while maintaining very low false-positive rates in predominantly healthy operational streams.

Self-supervised learning (SSL) has recently emerged as a promising solution for representation learning without manual annotations. Recent studies indicate that self-supervised approaches can achieve competitive performance in time-series anomaly detection while significantly reducing reliance on labeled datasets [20]. However, the integration of self-supervised representation learning with strict one-class anomaly detection for vibration-based rolling bearing monitoring remains relatively underexplored. Moreover, many existing studies adopt window-level random data splitting that risks information leakage, or evaluate performance under artificially balanced class distributions that do not reflect industrial reality.

To address these limitations, this work formulates rolling bearing fault detection as a strict one-class anomaly detection problem using a unified self-supervised learning framework. Unlike many existing studies that employ either contrastive learning or reconstruction objectives independently, the proposed framework integrates both strategies within a hybrid CNN–Transformer encoder trained exclusively on healthy vibration signals. The learned latent representations are subsequently modelled using an Isolation Forest trained solely on healthy embeddings, while anomaly thresholds are calibrated using held-out healthy validation windows within each training fold to prevent test data leakage. This combination enables anomaly detection without requiring labelled fault samples while providing a realistic evaluation under a strict healthy-only training protocol.

To ensure realistic evaluation, sample-level grouped cross-validation is employed to eliminate window-level information leakage. In addition to full-test evaluation, anomaly prevalence scenarios of 5% and 10% are simulated to reflect low fault prevalence industrial environments. Operational metrics, including false-positive rate and false alarms per 1000 healthy windows, are reported to assess performance under deployment-relevant conditions. The main contributions of this work are summarized as follows:

- A unified healthy-only self-supervised anomaly detection framework for rolling bearing health monitoring is proposed. The framework integrates contrastive and reconstruction-based self-supervised learning within a hybrid CNN–Transformer encoder and combines the learned representations with Isolation Forest for one-class anomaly detection. By learning exclusively from healthy vibration signals, the proposed approach eliminates the need for labelled fault samples during representation learning.
- A strict one-class evaluation protocol is introduced in which representation learning, anomaly detector training, and threshold calibration are performed exclusively using healthy data. Isolation Forest is trained only on healthy embeddings, while anomaly thresholds are determined using held-out healthy validation samples, preventing information leakage from faulty data.
- The robustness of the proposed framework is evaluated under realistic deployment conditions using prevalence-aware analyses at 5% and 10% fault prevalence levels. In addition to conventional ROC-AUC and PR-AUC metrics, operational performance is

assessed through specificity and false-positive behaviour, providing practical insights for predictive maintenance applications.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed methodology. Section 4 presents results and discussion. Section 5 concludes the paper.

2. Related Work

Vibration-based bearing fault diagnosis forms the foundation of condition monitoring in rotating machinery, as vibration signatures directly reflect the progression of incipient defects [10,11]. Localized defects generate characteristic frequency components governed by bearing geometry and rotational speed, making vibration monitoring highly sensitive to early-stage damage [11].

Traditional diagnostic frameworks rely on handcrafted feature extraction from time-domain, frequency-domain, and time–frequency representations. Common features include statistical descriptors such as RMS, skewness, and kurtosis, along with spectral indicators derived from FFT and PSD analysis [12]. Techniques such as envelope analysis and wavelet transforms are often employed to enhance impulsive fault signatures [21]. These features are typically combined with shallow classifiers such as SVM, kNN, Isolation Forest, or OCSVM [11]. While effective under controlled laboratory conditions, such approaches depend heavily on expert feature engineering and often exhibit limited robustness under non-stationary operating conditions or distribution shifts [10,13].

To address these limitations, deep learning (DL) methods have been widely adopted for end-to-end feature learning from raw vibration signals [14]. Convolutional neural networks (CNNs) automatically learn hierarchical local representations and have become dominant in vibration-based fault diagnosis [15,22]. However, convolutional architectures are inherently constrained by limited receptive fields, which restrict their ability to model long-range temporal dependencies [23]. Recurrent models such as LSTM and GRU have been introduced to capture sequential dependencies, but they suffer from scalability limitations and training instability for long sequences [24,25].

More recently, Transformer architectures have gained attention due to their self-attention mechanisms, which enable direct modeling of long-range dependencies without recursive computation [23,24]. Hybrid CNN–Transformer models leverage convolutional inductive bias for local feature extraction while employing attention mechanisms for global temporal modeling [25,26]. Empirical studies report improved performance compared to standalone CNN or RNN models in sequential learning tasks [25,26]. Alongside CNN- and Transformer-based architectures, generative models, particularly Generative Adversarial Networks (GANs), have emerged as an alternative paradigm for anomaly detection in industrial time-series data. Rather than directly performing supervised fault classification, GAN-based methods learn the distribution of healthy operating data and identify anomalies through deviations from the learned normal data manifold. For example, Jiang et al. proposed an encoder–decoder–encoder GAN framework trained exclusively on healthy bearing signals, where reconstruction and latent-space consistency losses were jointly used to detect anomalous samples without requiring labelled fault data. Such approaches have demonstrated promising performance for bearing fault diagnosis under highly imbalanced conditions. However, GAN-based methods rely on adversarial optimisation, which can be computationally demanding and susceptible to training instability, motivating the exploration of more stable self-supervised representation learning approaches [27]. Nevertheless, most deep learning-based bearing diagnosis approaches remain formulated as supervised multi-class classification problems, relying on large labeled datasets such as CWRU [28]. These datasets are often collected under seeded fault conditions and controlled environ-

ments, limiting generalization to real industrial settings [14]. Furthermore, supervised methods assume consistent training and testing distributions, an assumption that is frequently violated in practical industrial environments due to variations in rotational speed, mechanical load, sensor placement, measurement noise, and changing environmental conditions [29,30]. Such factors introduce domain shifts that can substantially degrade the generalisation performance of models trained under controlled laboratory conditions [30]. In addition, industrial monitoring systems typically generate large volumes of unlabelled healthy operating data, whereas fault events occur infrequently and are often difficult, expensive, or unsafe to label accurately. Consequently, obtaining representative labelled fault datasets covering diverse operating conditions remains a major challenge, making fully supervised learning paradigms difficult to deploy reliably in predictive maintenance applications [25,30,31].

These constraints motivate the use of anomaly detection and one-class learning frameworks, where models are trained primarily on normal data. Classical one-class methods and subspace techniques such as PCA and OCSVM have been applied in industrial settings [32–34]. Deep reconstruction-based approaches, particularly autoencoders, have further expanded industrial anomaly detection [32,35]. In parallel, deep one-class approaches such as Deep SVDD learn a compact hypersphere representation of healthy data and have emerged as strong baselines for unsupervised anomaly detection [36]. However, reconstruction objectives may reduce separability when model capacity allows accurate reconstruction of anomalous patterns [37]. In addition, many anomaly detection studies rely on offline validation, simulation-based datasets, or window-level random splits that risk information leakage and overestimated performance [38,39].

Contrastive self-supervised learning (SSL) has recently emerged as a promising alternative for representation learning without labeled fault data [40]. Instead of minimizing reconstruction error, contrastive methods maximize agreement between augmented views of the same sample while separating representations of different samples in embedding space [40]. This learning paradigm has been successfully applied to anomaly detection tasks, where contrastive objectives enhance separability between normal and abnormal patterns without requiring labeled datasets [41]. Hybrid deep–classical frameworks further combine neural encoders with lightweight one-class detectors, such as OC-SVM or related boundary-based methods, enabling nonlinear representation learning followed by anomaly boundary estimation [42]

Despite these advances, several gaps remain in vibration-based industrial monitoring. First, most approaches remain supervised or reconstruction-driven. Second, the integration of contrastive SSL with strict one-class modeling for rolling bearing vibration data is still limited. Third, evaluation protocols frequently ignore realistic fault prevalence and risk window-level data leakage. Finally, deployment-oriented metrics such as operational false alarm burden are rarely emphasized. These limitations highlight the need for a unified framework that combines contrastive self-supervised representation learning, leakage-free sample-level validation, healthy-only threshold calibration, and prevalence-aware evaluation under realistic industrial constraints.

3. Methodology

Figure 1 summarizes the complete workflow of the proposed approach, including preprocessing, self-supervised representation learning, one-class modeling, threshold calibration, and prevalence-aware evaluation.

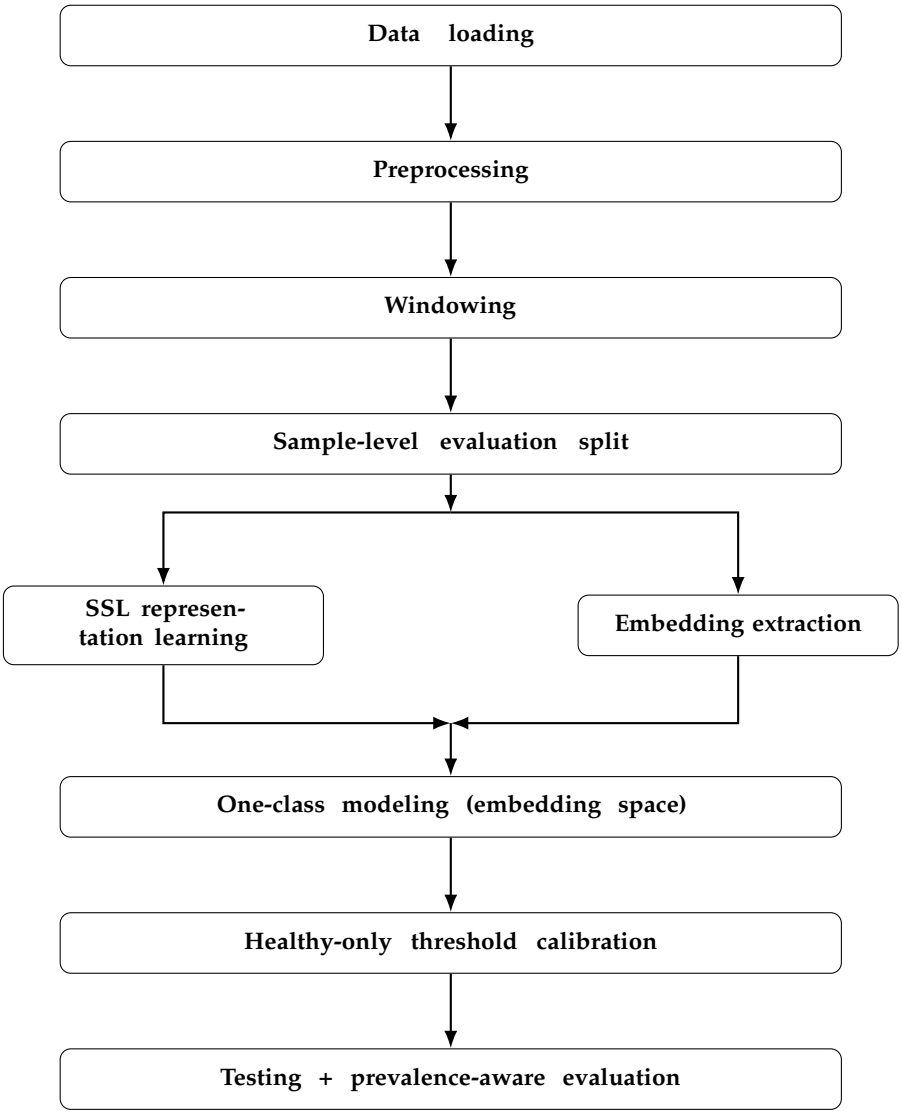

Figure 1. Workflow of the proposed self-supervised CNN–Transformer anomaly detection framework with sample-level grouped evaluation and healthy-only threshold calibration.

The workflow begins by loading raw bearing vibration data, applying preprocessing, and segmenting each signal into overlapping windows. Each window retains the identity of its source sample, and the dataset is then split using a sample-level grouped strategy with three-fold stratified group cross-validation protocol, so that all windows from the same sample are assigned entirely to either the training or test fold, preventing data leakage while maintaining approximate healthy-versus-anomalous class balance. Healthy training windows are used for self-supervised representation learning, embeddings are then extracted, and a one-class anomaly detector is trained in the learned embedding space. The anomaly threshold is calibrated using held-out healthy validation windows only, and the final model is evaluated on unseen test samples under both full-test and prevalence-aware conditions.

3.1. Dataset and Problem Formulation

A limited subset of vibration measurement recordings was intentionally selected from the Paderborn bearing dataset [43] to enable strict sample-level grouped evaluation while maintaining manageable computational requirements for extensive cross-validation and ablation experiments. The selected subset comprised 580 vibration measurement recordings, including 120 healthy recordings and 460 faulty recordings. The faulty recordings consisted

of 220 measurements with artificially induced defects (100 inner-race and 120 outer-race faults) and 240 measurements with naturally developed bearing defects (120 inner-race and 120 outer-race faults). All recordings were acquired under the fixed operating condition `N15_M07_F10`, corresponding to a rotational speed of 1500 rpm, a load torque of 0.7 N m, and a radial force of 1000 N [43]. Although the complete subset was not balanced between healthy and faulty recordings, it provided representation of both artificial and real fault categories while preserving the class imbalance typically encountered in industrial anomaly detection. The vibration signals were recorded at 64 kHz and consist of healthy, artificially induced, and real fault conditions. For representation learning, each signal was segmented into overlapping windows of 1024 points (≈16 ms) with a step size of 512 points (≈8 ms), producing fixed-length windows for training and evaluation. The task is formulated as anomaly detection, where healthy data defines the normal operating distribution and fault conditions are treated as deviations from this distribution during evaluation. Raw signals are normalized on a per-sample basis to reduce amplitude variations across measurement sessions. The proposed framework combines CNN-based local feature extraction with Transformer-based temporal modeling within a self-supervised learning setting, allowing representations to be learned using healthy data only, while anomaly detection is performed through unsupervised scoring. Performance is assessed using ranking metrics together with operational metrics relevant to industrial deployment, including false alarms per 1000 healthy windows, and additional experiments simulate realistic low fault prevalence scenarios of 5% and 10%. This design emphasizes a deployment-oriented anomaly detection pipeline that integrates hybrid temporal representation learning with rigorous evaluation under realistic operating conditions.

3.2. CNN–Transformer Encoder Architecture

The proposed encoder combines convolutional feature extraction with Transformer-based temporal modeling to capture both local vibration characteristics and longer-range temporal dependencies within bearing vibration signals.

Each input consists of a vibration window of 1024 points represented as a one-dimensional sequence. The signal is first processed by a convolutional backbone comprising three one-dimensional convolutional layers with kernel sizes of 7, 5, and 3, respectively. Each convolutional layer is followed by batch normalization and a rectified linear unit (ReLU) activation function. All convolutional layers employ a stride of 2, progressively reducing the temporal resolution while increasing the representational capacity of the feature maps.

The final convolutional layer produces a latent representation with an embedding dimension of 64. The resulting feature sequence is then passed to a Transformer encoder consisting of two TransformerEncoderLayer blocks. Each Transformer layer employs four self-attention heads and a feedforward network with dimensionality equal to twice the embedding dimension. A dropout rate of 0.1 is applied within the Transformer layers to improve generalization.

Following Transformer processing, global average pooling is applied across the temporal dimension to obtain a fixed-length embedding vector. The embedding is subsequently ℓ_2-normalized and serves as the latent representation used for anomaly detection (Figure 2).

During self-supervised pretraining, a projection head composed of two fully connected layers with batch normalization and ReLU activation is attached to the encoder. The projection head is used only during contrastive representation learning and is discarded after training. During inference, anomaly detection is performed directly on the pooled encoder embeddings.

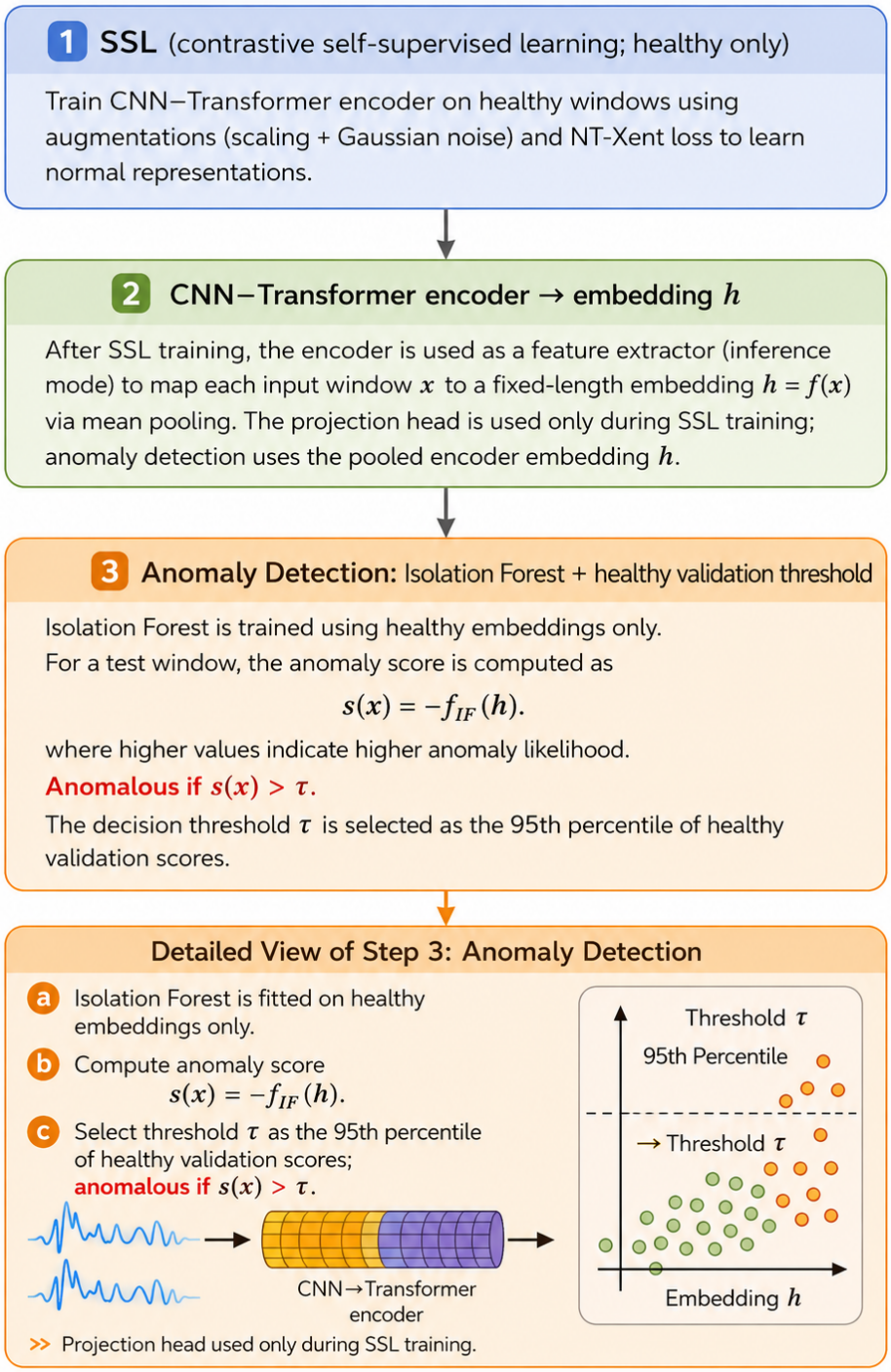

Figure 2. Workflow of the proposed self-supervised CNN–Transformer anomaly detection framework. Healthy vibration windows are used for self-supervised representation learning, and the resulting embeddings are employed by an Isolation Forest for anomaly detection.

3.3. Self-Supervised Representation Learning

A self-supervised learning (SSL) framework is employed to learn discriminative representations from healthy vibration data without requiring fault annotations. Within each training fold, only healthy vibration windows are used for representation learning. Three SSL strategies are investigated: contrastive learning, reconstruction-based learning, and a combined contrastive–reconstruction approach (Figure 3).

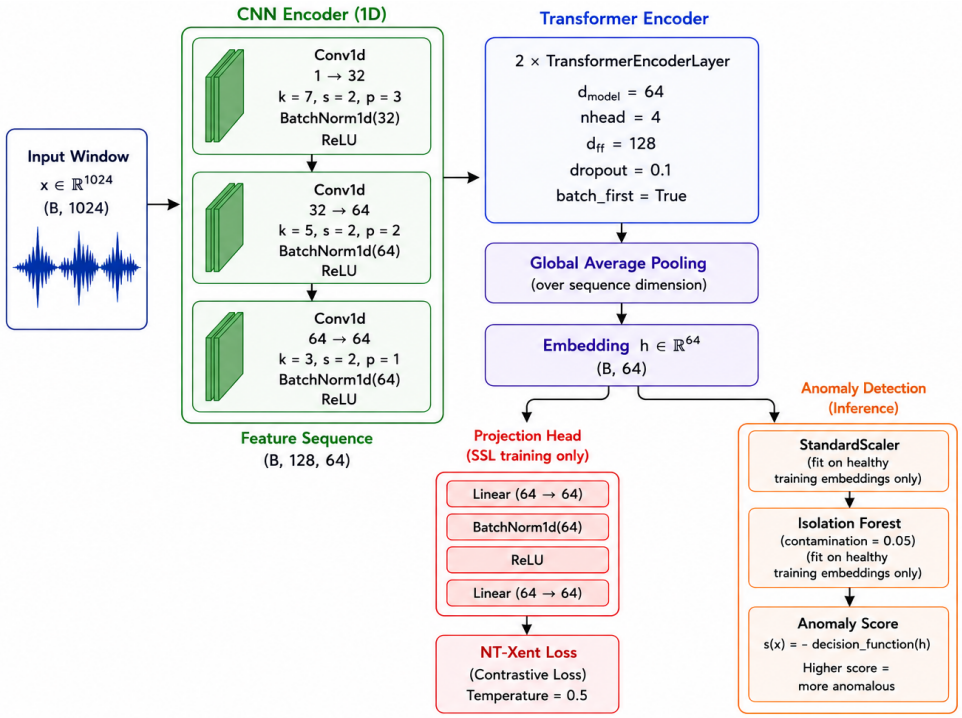

Figure 3. Implementation-level architecture of the proposed self-supervised CNN–Transformer anomaly detection framework. Vibration windows are encoded using a CNN–Transformer backbone, and the resulting embeddings are optimized through contrastive self-supervised learning. During inference, the learned embeddings are analyzed using an Isolation Forest for anomaly detection.

For contrastive learning, positive pairs are constructed by selecting two healthy vibration windows originating from the same measurement file. Each window is subsequently subjected to weak physically plausible augmentations to improve robustness while preserving the underlying signal characteristics. The augmentation strategy consists of additive Gaussian noise, amplitude scaling, and small temporal shifts. Gaussian noise is added according to $\epsilon \sim \mathcal{N}(0, 0.01^2)$, while amplitude scaling factors are sampled around unity. Small temporal shifts are introduced through circular displacement of the vibration sequence, accounting for minor phase variations that may occur during signal acquisition. These augmentations preserve the physical characteristics of the vibration signal while encouraging the encoder to learn invariant and robust representations. The augmented views are processed by the encoder and optimized using the normalized temperature-scaled cross-entropy (NT-Xent) contrastive loss. Embeddings are normalized prior to similarity computation, and the temperature parameter is fixed at 0.5.

For the combined contrastive–reconstruction self-supervised learning model, the overall training objective is defined as

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{NT-Xent}} + \lambda \mathcal{L}_{\text{rec}}, \tag{1}$$

where $\mathcal{L}_{\text{NT-Xent}}$ denotes the normalized temperature-scaled cross-entropy contrastive loss, $\mathcal{L}_{\text{rec}}$ represents the mean squared error (MSE) between the input vibration window and its reconstructed output, and λ is the reconstruction weighting coefficient. In this study, λ was fixed at 1.0, assigning equal importance to the contrastive and reconstruction objectives during training.

In addition to contrastive learning, a reconstruction-based SSL strategy is also evaluated. In this setting, the encoder is coupled with a lightweight decoder comprising fully connected layers that reconstruct the input vibration window from the latent embedding. The model is trained exclusively on healthy vibration windows, and the reconstruction objective en-

courages the latent representation to capture the underlying structure and normal operating characteristics of the vibration signals. Reconstruction quality is optimized by minimizing the mean squared error between the original and reconstructed vibration windows.

Finally, a combined SSL strategy is investigated to exploit the complementary strengths of contrastive and reconstruction learning. In this approach, the model simultaneously optimizes both the NT-Xent contrastive objective and the reconstruction objective during training. Positive pairs are generated and augmented in the same manner as the contrastive framework, while both augmented views are reconstructed by the decoder. The total training objective is computed as a weighted combination of contrastive and reconstruction losses, with the reconstruction weight fixed at 1.0. This joint optimization encourages the encoder to learn representations that are both invariant to local signal perturbations and informative of the underlying signal structure.

Training is performed for 20 epochs using the Adam optimizer with a learning rate of 5×10^{-4} and a batch size of 128. A cosine annealing learning-rate scheduler is employed throughout training to improve optimization stability and convergence. During SSL training, a two-layer projection head comprising fully connected layers, batch normalization, and ReLU activation is attached to the encoder output. The contrastive objective is applied in the projection space, whereas the encoder embedding is retained as the learned representation. After training, the projection head and decoder, when present, are discarded, and the encoder embeddings are used for downstream anomaly detection.

3.4. Anomaly Detection in the Embedding Space

Following self-supervised representation learning, the encoder is frozen and used to extract embeddings for all samples in the training, validation, and test sets. To ensure consistent scaling across latent dimensions, embeddings are standardized using a StandardScaler fitted exclusively on healthy training embeddings. The same transformation is subsequently applied to the validation and test embeddings.

Anomaly detection is performed in the learned embedding space using an Isolation Forest (IF). The IF model is trained exclusively on healthy embeddings from the training fold, preserving the one-class anomaly detection setting and preventing the use of fault information during model fitting.

For a test sample x, the anomaly score is computed as the negative output of the Isolation Forest decision function:

$$s(x) = -f_{\mathrm{IF}}(z(x)), \tag{2}$$

where $z(x)$ denotes the standardized encoder embedding. Higher scores indicate a greater likelihood of anomalous behaviour.

The Isolation Forest is implemented using scikit-learn with a contamination parameter of 0.05 and a fold-dependent random seed (42+ fold index). All remaining hyperparameters are retained at their default values. The resulting anomaly scores are subsequently used for threshold calibration and performance evaluation.

For comparative analysis, a Mahalanobis-based anomaly scoring approach was also evaluated. Healthy training embeddings were first standardized using statistics computed exclusively from the healthy training set. Subsequently, the mean and variance of each latent dimension were estimated from the standardized healthy embeddings. Anomaly scores for validation and test samples were then computed as the variance-normalized distance from the healthy embedding mean, where larger distances indicate a greater deviation from normal operating behaviour. This distance-based approach was included as a statistical baseline to assess whether the learned embedding space could be effectively characterized using deviations from the healthy embedding distribution.

3.5. Healthy-Only Validation and Threshold Selection

To avoid bias from test data and simulate realistic deployment conditions, threshold selection is performed using an internal validation split derived solely from healthy training windows.

Within each cross-validation fold:

1. Healthy windows are divided into an SSL training subset and a healthy validation subset.
2. The SSL encoder is trained using the SSL training healthy subset only.
3. Healthy embeddings are extracted, standardized using statistics computed from the healthy training set, and used to train the Isolation Forest.
4. The decision threshold is defined as the 95th percentile of anomaly scores computed on the healthy validation subset.

During testing, a window is classified as anomalous if its anomaly score exceeds this 95% threshold.

3.6. Cross-Validation Protocol

Model evaluation is conducted using a three-fold stratified group K-Fold cross-validation procedure. Stratification is performed with respect to the binary label (healthy versus anomalous), while grouping is enforced using measurement sample identifiers.

Within each fold, all windows originating from a given measurement sample are assigned exclusively to either the training or testing partition. The grouped evaluation protocol therefore assesses the ability of the model to generalize to previously unseen measurement sample rather than unseen windows from previously observed sample.

Because the experimental subset contains a limited number of measurement samples, strict sample-level grouping reduces the risk of overly optimistic performance estimates while providing a more realistic assessment of deployment performance under unseen operating conditions.

3.7. Evaluation Metrics and Prevalence Simulation

Performance is assessed using both ranking-based and threshold-dependent metrics. Ranking performance is evaluated using the area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision–recall curve (AUC-PR).

Threshold-dependent evaluation includes precision, recall, F1-score, false-positive rate (FPR), and specificity. In addition, false alarms per 1000 healthy windows (FA/1000) are computed as:

$$\mathrm{FA}/1000 = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \times 1000 \tag{3}$$

where FP denotes the number of false positives and TN denotes the number of true negatives, so that $\mathrm{TN} + \mathrm{FP}$ represents the total number of healthy windows in the evaluated set. This metric quantifies the number of false alarms generated per 1000 healthy windows and provides an operationally meaningful measure of reliability for industrial monitoring systems.

Because the dataset is approximately balanced at the window level, realistic deployment conditions are simulated by enforcing anomaly prevalence levels of 5% and 10% in the test set. This is achieved by retaining all healthy windows and randomly subsampling anomalous windows to obtain the desired prevalence. The subsampling procedure is repeated five times using different random seeds, and performance metrics are averaged across repetitions to obtain stable estimates under low-prevalence operating conditions.

All experiments were implemented in Python using PyTorch for self-supervised representation learning and scikit-learn for anomaly detection, threshold calibration, and performance evaluation.

4. Results and Discussion

4.1. Training Behaviour

The training loss curves are shown in Figure 4. The autoencoder reconstruction loss decreased steadily across epochs, indicating stable optimisation. Contrastive SSL losses for both CNN-only and CNN + Transformer encoders also decreased gradually, showing that the encoders learned increasingly consistent representations for augmented healthy vibration windows. The full SSL model showed stable optimisation of the combined contrastive and reconstruction loss.

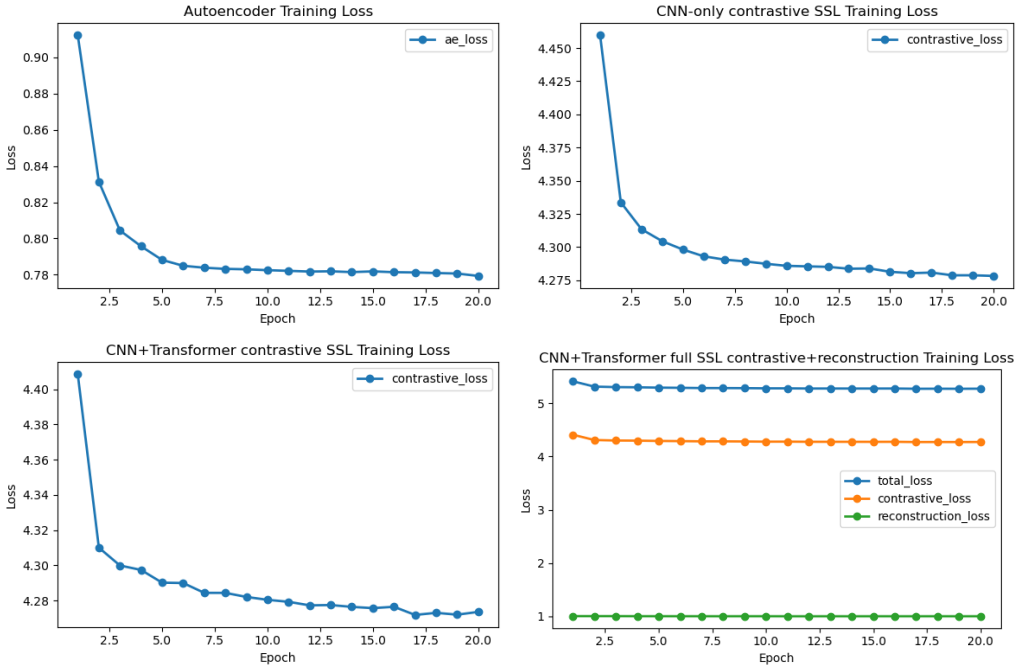

Figure 4. Training loss curves for the autoencoder, CNN-only contrastive SSL, CNN + Transformer contrastive SSL, and CNN + Transformer full SSL models.

4.2. Experimental Performance Comparison

Table 1 presents the mean performance obtained over three-fold stratified group cross-validation. The classical machine-learning baselines based on handcrafted vibration features achieved moderate performance. Isolation Forest using classical statistical features obtained a mean ROC-AUC of 0.625 ± 0.003 and a PR-AUC of 0.854 ± 0.004, while One-Class SVM achieved a higher ROC-AUC of 0.734 ± 0.004 and PR-AUC of 0.901 ± 0.003. However, both methods showed limited recall and F1-score, indicating that handcrafted statistical features were insufficient to capture the full complexity of bearing fault behaviour.

The autoencoder baseline produced the weakest performance, with a ROC-AUC of 0.429 ± 0.011 and an F1-score of 0.067 ± 0.006. This suggests that reconstruction error alone was not effective for distinguishing healthy and faulty bearing signals in this experimental setting. The autoencoder may reconstruct both normal and abnormal vibration windows with similar error, thereby reducing its discriminative capability.

To provide a comparison with a recent deep one-class anomaly detection approach, Deep SVDD was implemented and evaluated using the same healthy-only training protocol, grouped three-fold cross-validation, and healthy-validation threshold calibration as the proposed framework. As shown in Table 1, Deep SVDD achieved a ROC-AUC of 0.654 ± 0.014, a PR-AUC of 0.891 ± 0.005, and an F1-score of 0.422 ± 0.021. Compared with the Autoencoder baseline, Deep SVDD substantially improved anomaly detection performance and also achieved a higher F1-score than the classical OCSVM baseline. Nev-

ertheless, all proposed self-supervised learning models achieved higher ROC-AUC and PR-AUC values, indicating that the learned self-supervised representations provide more discriminative features for bearing fault anomaly detection than the hypersphere-based representation learned by Deep SVDD.

Table 1. Main performance comparison across three-fold cross-validation. The bold values indicate the best performance achieved for each evaluation metric among all compared methods.

Method	ROC-AUC	PR-AUC	Precision	Recall	F1-Score
IF (classical features)	0.625 ± 0.003	0.854 ± 0.004	0.888 ± 0.011	0.103 ± 0.016	0.184 ± 0.026
OCSVM (classical features)	0.734 ± 0.004	0.901 ± 0.003	0.940 ± 0.001	0.202 ± 0.012	0.332 ± 0.016
Autoencoder	0.429 ± 0.011	0.758 ± 0.008	0.730 ± 0.019	0.035 ± 0.003	0.067 ± 0.006
Deep SVDD	0.654 ± 0.014	0.891 ± 0.005	0.951 ± 0.006	0.271 ± 0.017	0.422 ± 0.021
CNN-only contrastive SSL + IF	0.790 ± 0.016	0.937 ± 0.007	$\mathbf{0.970 \pm 0.005}$	$\mathbf{0.439 \pm 0.051}$	$\mathbf{0.603 \pm 0.050}$
CNN + Transformer contrastive SSL + IF	0.816 ± 0.037	0.934 ± 0.012	0.957 ± 0.009	0.302 ± 0.091	0.455 ± 0.102
CNN + Transformer full SSL + IF	$\mathbf{0.878 \pm 0.015}$	$\mathbf{0.958 \pm 0.005}$	0.970 ± 0.003	0.425 ± 0.040	0.590 ± 0.040

In contrast, the proposed self-supervised learning approaches substantially improved anomaly detection performance. The CNN-only contrastive SSL model combined with Isolation Forest achieved a ROC-AUC of 0.790 ± 0.016 and a PR-AUC of 0.937 ± 0.007. The CNN + Transformer contrastive SSL model further improved ROC-AUC to 0.816 ± 0.037, showing the benefit of temporal modelling using transformer layers.

Although the CNN + Transformer contrastive model achieved a higher mean ROC-AUC than the CNN-only model, its F1-score and recall were lower and exhibited greater variation across folds. This is because ROC-AUC evaluates ranking performance across all decision thresholds, whereas the F1-score depends on a single calibrated threshold. The higher variance may also reflect the greater capacity of the Transformer and the limited number of grouped recordings available in each fold.

The best overall detection performance was obtained by the CNN + Transformer full SSL model, which combined contrastive and reconstruction objectives and used Isolation Forest for anomaly scoring. This model achieved the highest ROC-AUC of 0.878 ± 0.015 and the highest PR-AUC of 0.958 ± 0.005. These results demonstrate that the proposed SSL framework learns informative latent representations from raw vibration windows and significantly outperforms classical feature-based and reconstruction-only baselines.

Although the proposed framework achieved high ROC-AUC and PR-AUC values, the recall remained relatively low, indicating that some faulty windows were classified as healthy. The anomaly detection threshold was calibrated using the 95th percentile of healthy validation scores to avoid the use of fault labels during threshold selection. This choice was intended to provide a practical balance between missed detections and false alarms while remaining consistent with the healthy-only learning setting adopted in this work. In practical deployments, the operating threshold can be adjusted according to application requirements, depending on whether reducing false alarms or minimising missed detections is the primary objective.

Additional analyses investigating the contribution of different self-supervised learning objectives and the influence of alternative anomaly scoring strategies are provided in Appendix A.

4.3. Linear Probe Evaluation

To assess the quality of the learned self-supervised representations, a linear probe classifier was trained on frozen embeddings generated by the CNN + Transformer contrastive SSL encoder.

Across the three cross-validation folds, the linear probe achieved ROC-AUC values of 0.974, 0.979, and 0.980, corresponding to a mean ROC-AUC of 0.978. The consistently high performance indicates that fault-related information is strongly encoded within the learned latent space despite the encoder being trained exclusively on healthy vibration signals.

These findings suggest that the self-supervised representations capture discriminative characteristics relevant to bearing health monitoring and provide additional evidence supporting the effectiveness of the proposed representation learning framework.

4.4. ROC and Precision–Recall Analysis

The ROC and precision–recall curves of the best-performing model are shown in Figures 5 and 6. The ROC curve demonstrates strong separation between healthy and faulty windows, while the precision–recall curve confirms strong performance under class imbalance. The high PR-AUC is particularly important because the dataset contains substantially more faulty windows than healthy windows after windowing.

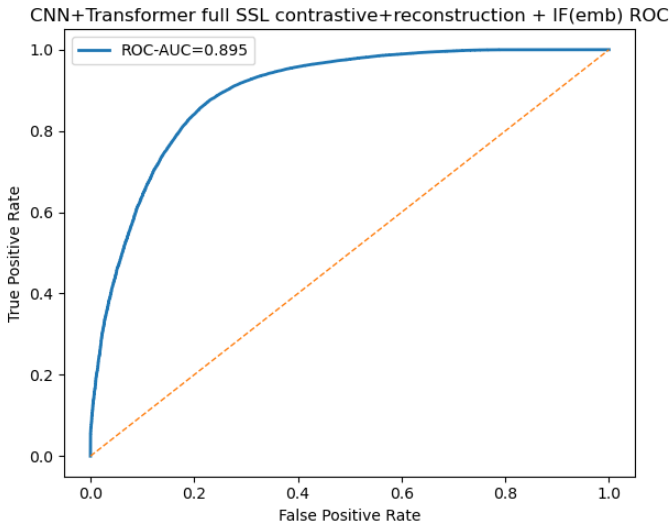

Figure 5. ROC curve of the best-performing CNN + Transformer full SSL model with Isolation Forest anomaly scoring.

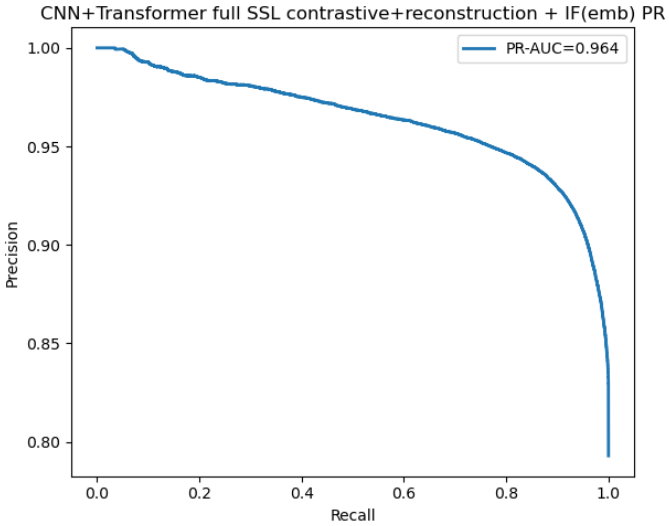

Figure 6. Precision–recall curve of the best-performing CNN + Transformer full SSL model with Isolation Forest anomaly scoring.

4.5. Anomaly Score Distribution and Confusion Matrix

Figure 7 shows the anomaly score distribution for healthy and faulty windows. The separation between the two distributions confirms that the learned representation and Isolation Forest scorer assign higher anomaly scores to faulty samples. The vertical threshold line represents the decision boundary derived from the healthy validation set.

Figure 8 presents the confusion matrix for the best-performing model. The model achieved high precision and specificity, indicating that most detected anomalies correspond to true faults and that the false-positive rate remained low. However, recall was lower than precision, suggesting that some faulty windows remained close to the healthy distribution and were therefore not detected at the selected threshold.

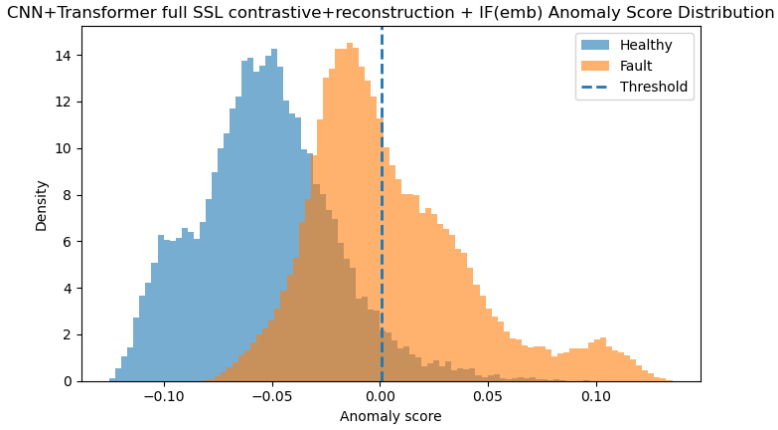

Figure 7. Anomaly score distribution of healthy and faulty windows for the best-performing model.

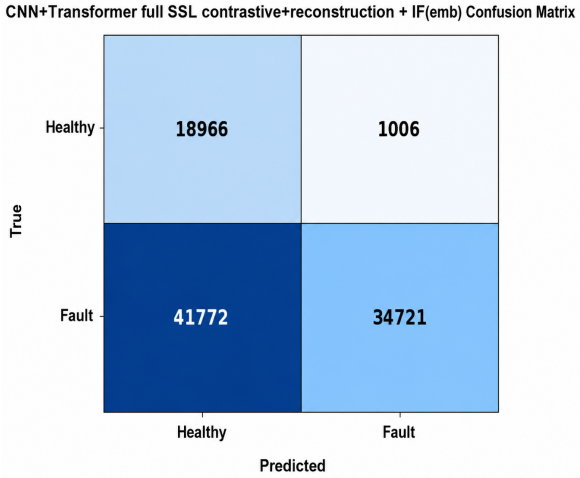

Figure 8. Confusion matrix of the best-performing CNN + Transformer full SSL model with Isolation Forest anomaly scoring.

4.6. Latent Space Visualisation

Figures 9 and 10 show two-dimensional projections of the learned embeddings using PCA and t-SNE, respectively. The projections indicate that healthy and faulty windows become partially separated in the learned latent space. This supports the quantitative results and confirms that the SSL encoder captures fault-sensitive structure despite being trained using healthy windows only.

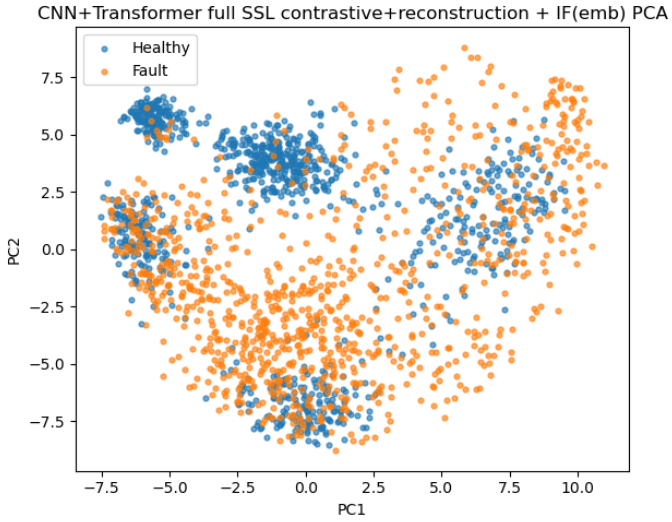

Figure 9. PCA projection of balanced healthy and faulty embeddings from the best-performing SSL model.

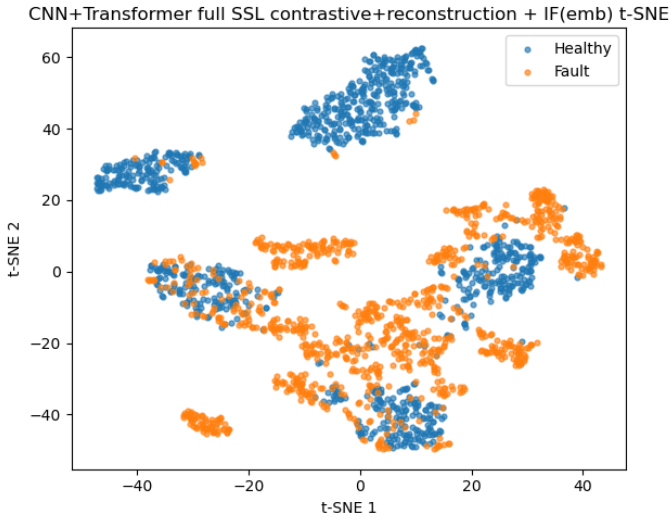

Figure 10. t-SNE projection of balanced healthy and faulty embeddings from the best-performing SSL model.

4.7. Robustness Under Low Fault Prevalence

To evaluate practical deployment behaviour, additional prevalence analysis was performed at 5% and 10% fault prevalence. Table 2 summarises the observed ranges across folds. The model maintained specificity above 0.94 in both settings, indicating stable false-positive behaviour. As expected, PR-AUC decreased under lower prevalence because precision–recall metrics are sensitive to class imbalance. Nevertheless, ROC-AUC remained between 0.770 and 0.842 at 5% prevalence and between 0.774 and 0.840 at 10% prevalence.

Table 2. Prevalence analysis of the CNN + Transformer contrastive SSL model with Isolation Forest.

Fault Prevalence	ROC-AUC Range	PR-AUC Range	Specificity Range
5%	0.770–0.842	0.195–0.292	0.947–0.952
10%	0.774–0.840	0.321–0.425	0.947–0.952

4.8. Discussion

The experimental results demonstrate that self-supervised representation learning provides a strong alternative to classical feature engineering for bearing anomaly detection.

The proposed CNN + Transformer full SSL model achieved the highest ROC-AUC and PR-AUC among all evaluated methods, showing that combining convolutional feature extraction, transformer-based temporal modelling, and self-supervised objectives produces highly informative embeddings for anomaly detection.

The results also show that the full SSL objective outperformed both reconstruction-only and contrastive-only alternatives. Reconstruction-only learning was insufficient because reconstruction error does not necessarily correspond to fault severity. In contrast, contrastive learning improved separability by encouraging consistency between augmented healthy windows. The combined objective achieved the best performance, suggesting that contrastive discrimination and reconstruction-based regularisation are complementary.

Another important finding is the importance of the anomaly scoring mechanism. Isolation Forest consistently outperformed Mahalanobis distance, indicating that the learned latent space does not follow a simple Gaussian structure. This suggests that non-parametric anomaly scoring is more appropriate for SSL embeddings in bearing fault detection.

Overall, the proposed CNN + Transformer SSL framework provides an effective, label-efficient, and scalable approach for bearing health monitoring. The leakage-free sample-level evaluation protocol further increases confidence that the reported performance reflects generalisation to unseen measurement sessions rather than memorisation of window-level signal characteristics. The method achieves strong anomaly detection performance while relying primarily on healthy data for representation learning, making it suitable for predictive maintenance applications in real industrial environments. These findings are particularly relevant for industrial predictive maintenance, where healthy operating data are abundant but labelled fault data are scarce, making healthy-only anomaly detection frameworks attractive for practical deployment. Although the proposed framework demonstrated strong performance on the selected subset of the Paderborn bearing dataset, several limitations remain. The experiments were conducted using a single benchmark dataset and a limited number of measurement sample. Additional validation across multiple bearing datasets and varying operating conditions is required to assess generalization robustness. Furthermore, the study focused on anomaly detection rather than fault-type classification. Future work will investigate cross-domain adaptation, online deployment, and transfer learning under changing industrial operating conditions.

5. Conclusions

This study presented a self-supervised learning framework for bearing anomaly detection that combines convolutional neural networks and Transformer-based temporal modelling with unsupervised anomaly scoring. The proposed approach was designed to operate in a healthy-only training setting, where fault labels are not required during representation learning. By integrating contrastive and reconstruction-based self-supervised objectives with Isolation Forest anomaly detection, the framework learns informative latent representations directly from raw vibration signals while preserving the practical advantages of one-class learning.

Despite these promising results, several limitations should be acknowledged. The proposed framework was evaluated using a subset of the Paderborn bearing dataset under a single operating condition and therefore does not explicitly assess robustness to measurement noise, varying rotational speeds, changing load conditions, or cross-domain distribution shifts commonly encountered in industrial environments. Furthermore, the proposed framework focuses on anomaly detection rather than fault classification and therefore identifies abnormal operating behaviour without distinguishing specific fault types. These limitations should be considered when interpreting the results and motivate further investigation under more diverse industrial scenarios.

Experimental evaluation on the Paderborn bearing dataset using strict three-fold sample-level stratified group cross-validation demonstrated that self-supervised representation learning substantially outperforms classical feature-engineering approaches and reconstruction-only baselines. Among all evaluated methods, the CNN + Transformer full self-supervised learning model combined with Isolation Forest achieved the best overall performance, obtaining a ROC-AUC of 0.878 ± 0.015, a PR-AUC of 0.958 ± 0.005, and an F1-score of 0.590 ± 0.040. These results indicate that the proposed framework effectively captures fault-sensitive information from vibration signals despite being trained exclusively on healthy data.

The ablation study demonstrated that contrastive learning plays the dominant role in generating discriminative representations, while reconstruction learning provides complementary regularisation that further improves performance. In addition, Isolation Forest consistently outperformed Mahalanobis distance across all learned representations, suggesting that the latent space learned by self-supervised models exhibits complex non-Gaussian structure that benefits from non-parametric anomaly scoring.

Additional prevalence-aware evaluation under 5% and 10% fault prevalence conditions showed that the proposed framework remains robust in highly imbalanced deployment scenarios. The model maintained high specificity and stable ranking performance while preserving practical anomaly detection capability under realistic industrial operating conditions.

The results of this study indicate that combining convolutional feature extraction, Transformer-based temporal modelling, self-supervised representation learning, and unsupervised anomaly detection can provide an effective and label-efficient approach for bearing anomaly detection. Although the proposed framework achieved strong ranking performance (ROC-AUC and PR-AUC), the moderate F1-score indicates that further improvements in threshold selection and decision calibration are required to enhance detection performance. Moreover, the conclusions are based on experiments conducted using a selected subset of the Paderborn bearing dataset and should therefore be interpreted with appropriate caution until validated across additional datasets and operating conditions.

Future work will focus on validating the proposed framework across multiple bearing datasets and diverse operating conditions, including varying rotational speeds, load conditions, and measurement noise. Additional research will investigate adaptive threshold selection, domain adaptation, online learning, and lightweight deployment on resource-constrained edge devices. Extending the framework to fault classification and fault localisation while preserving the advantages of healthy-only self-supervised representation learning also represents an important direction for future research.

Author Contributions: Conceptualization, S.H.H.Z. and A.S.; methodology, S.H.H.Z. and A.S.; investigation, S.H.H.Z.; writing—original draft, S.H.H.Z.; writing—review and editing, S.H.H.Z. and A.S.; reviewing, A.S., H.Z. and A.I.; supervision, A.S., H.Z. and A.I. All authors have read and agreed to the published version of the manuscript.

Funding: This study received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s). Code and data analysis will be made available on GitHub following publication.

Acknowledgments: The authors gratefully acknowledge support from the Advanced Food Innovation Centre (AFIC) and Sheffield Hallam University. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version of this paper arising from this submission.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Supplementary Experimental Analysis

Appendix A.1. Ablation Study of SSL Objectives

Table A1 presents the ablation study of the self-supervised learning objectives. The results show that contrastive learning substantially improves representation quality compared with reconstruction-only learning, while combining contrastive and reconstruction objectives achieves the best overall performance.

Table A1. Ablation study of self-supervised learning objectives using Isolation Forest anomaly scoring. The bold values indicate the best performance achieved for each evaluation metric among all compared methods.

SSL Objective	ROC-AUC	PR-AUC
CNN + Transformer reconstruction-only SSL + IF	0.615 ± 0.022	0.854 ± 0.011
CNN + Transformer contrastive SSL + IF	0.816 ± 0.037	0.934 ± 0.012
CNN + Transformer full SSL + IF	$\mathbf{0.878 \pm 0.015}$	$\mathbf{0.958 \pm 0.005}$

Appendix A.2. Effect of Anomaly Scoring Strategy

Table A2 compares the anomaly scoring strategies used with the learned SSL representations. Isolation Forest consistently outperformed Mahalanobis distance across all representations, indicating its greater suitability for modelling the complex structure of the learned embedding space.

Table A2. Comparison of anomaly scoring strategies for learned SSL representations. The bold values indicate the best performance achieved for each evaluation metric among all compared methods.

Representation and Scorer	ROC-AUC	PR-AUC	F1-Score
CNN-only contrastive SSL + Mahalanobis	0.661 ± 0.042	0.897 ± 0.022	0.485 ± 0.089
CNN-only contrastive SSL + IF	0.790 ± 0.016	0.937 ± 0.007	0.603 ± 0.050
CNN + Transformer contrastive SSL + Mahalanobis	0.497 ± 0.067	0.823 ± 0.028	0.246 ± 0.046
CNN + Transformer contrastive SSL + IF	0.816 ± 0.037	0.934 ± 0.012	0.455 ± 0.102
CNN + Transformer full SSL + Mahalanobis	0.564 ± 0.005	0.845 ± 0.012	0.272 ± 0.074
CNN + Transformer full SSL + IF	$\mathbf{0.878 \pm 0.015}$	$\mathbf{0.958 \pm 0.005}$	0.590 ± 0.040

References

1. Mohan, T.R.; Roselyn, J.P.; Uthra, R.A.; Devaraj, D.; Umachandran, K. Intelligent machine learning based total productive maintenance approach for achieving zero downtime in industrial machinery. *Comput. Ind. Eng.* **2021**, *157*, 107267. [CrossRef]
2. Zhang, W.; Li, C.; Peng, G.; Chen, Y.; Zhang, Z. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mech. Syst. Signal Process.* **2018**, *100*, 439–453. [CrossRef]
3. Thomas, D.; Weiss, B. Maintenance costs and advanced maintenance techniques in manufacturing machinery: Survey and analysis. *Int. J. Progn. Health Manag.* **2021**, *12*, 10-36001. [CrossRef]
4. Masani, K.I.; Oza, P.; Agrawal, S. Predictive maintenance and monitoring of industrial machine using machine learning. *Scalable Comput. Pract. Exp.* **2019**, *20*, 663–668. [CrossRef]
5. Patil, C.R.; Jadhav, S.K.; Bardiya, A.L.; Davande, A.P.; Raverkar, M.P. Machine learning-based predictive maintenance of industrial machines. *Int. J. Comput. Trends Technol.* **2023**, *71*, 50–56. [CrossRef]
6. Burda, E.A.; Zusman, G.V.; Kudryavtseva, I.S.; Naumenko, A.P. An overview of vibration analysis techniques for the fault diagnostics of rolling bearings in machinery. *Shock Vib.* **2022**, *2022*, 6136231. [CrossRef]
7. Chen, Y.; Chen, J.; Qiang, Y.; Yuan, Z.; Yang, J. Refined composite moving average fluctuation dispersion entropy and its application on rolling bearing fault diagnosis. *Rev. Sci. Instrum.* **2023**, *94*, 105110. [CrossRef] [PubMed]
8. Chen, X.; Chen, Y.; Ju, Z.; Qiang, Y.Q.; Yang, J. Rolling bearing fault diagnosis under limited data using Conditional Mask-Enhanced VQGAN and Convolutional-Embedded vision Mamba. *Meas. Sci. Technol.* **2026**, *37*, 046206. [CrossRef]
9. Randall, R.B. *Vibration-Based Condition Monitoring: Industrial, Automotive and Aerospace Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
10. Kumar, S.; Lokesha, M.; Kumar, K.; Srinivas, K.R. Vibration based fault diagnosis techniques for rotating mechanical components: Review paper. *IOP Conf. Ser. Mater. Sci. Eng.* **2018**, *376*, 12109. [CrossRef]

11. Hu, W.; Xin, G.; Wu, J.; An, G.; Li, Y.; Feng, K.; Antoni, J. Vibration-based bearing fault diagnosis of high-speed trains: A literature review. *High-Speed Railw.* **2023**, *1*, 219–223. [CrossRef]
12. Altaf, M.; Akram, T.; Khan, M.A.; Iqbal, M.; Ch, M.M.I.; Hsu, C.H. A new statistical features based approach for bearing fault diagnosis using vibration signals. *Sensors* **2022**, *22*, 2012. [CrossRef] [PubMed]
13. Song, X.; Liu, J.; Xia, M. Advanced Vibration-Based Fault Diagnosis and Vibration Control Methods. *Sensors* **2023**, *23*, 7704. [CrossRef] [PubMed]
14. Shenfield, A.; Howarth, M. A novel deep learning model for the detection and identification of rolling element-bearing faults. *Sensors* **2020**, *20*, 5112. [CrossRef] [PubMed]
15. Chen, Z.; Mauricio, A.; Li, W.; Gryllias, K. A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks. *Mech. Syst. Signal Process.* **2020**, *140*, 106683. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
17. Dong, H.; Zhang, Y.; Gu, H.; Konz, N.; Zhang, Y.; Mazurowski, M.A. SWSSL: Sliding window-based self-supervised learning for anomaly detection in high-resolution images. *IEEE Trans. Med. Imaging* **2023**, *42*, 3860–3870. [CrossRef] [PubMed]
18. Hasan, M.; Islam, M.M.; Zarif, M.I.I.; Hashem, M. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. *Internet Things* **2019**, *7*, 100059. [CrossRef]
19. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep learning for anomaly detection: A review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–38. [CrossRef]
20. Tran, D.H.; Nguyen, V.L.; Nguyen, H.; Jang, Y.M. Self-supervised learning for time-series anomaly detection in Industrial Internet of Things. *Electronics* **2022**, *11*, 2146. [CrossRef]
21. Hamadache, M.; Lee, D.; Mucchi, E.; Dalpiaz, G. Vibration-based bearing fault detection and diagnosis via image recognition technique under constant and variable speed conditions. *Appl. Sci.* **2018**, *8*, 1392. [CrossRef]
22. Wang, J.; Mo, Z.; Zhang, H.; Miao, Q. A deep learning method for bearing fault diagnosis based on time-frequency image. *IEEE Access* **2019**, *7*, 42373–42383. [CrossRef]
23. Yuan, J.; Zhou, F.; Guo, Z.; Li, X.; Yu, H. HCformer: Hybrid CNN-transformer for LDCT image denoising. *J. Digit. Imaging* **2023**, *36*, 2290–2305. [CrossRef] [PubMed]
24. Bashir, T.; Wang, H.; Tahir, M.; Zhang, Y. Wind and solar power forecasting based on hybrid CNN-ABiLSTM, CNN-transformer-MLP models. *Renew. Energy* **2025**, *239*, 122055. [CrossRef]
25. Ullah, W.; Hussain, T.; Ullah, F.U.M.; Lee, M.Y.; Baik, S.W. TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106173. [CrossRef]
26. Chen, Y.; Chen, X.; Xu, A.; Sun, Q.; Peng, X. A hybrid CNN-Transformer model for ozone concentration prediction. *Air Qual. Atmos. Health* **2022**, *15*, 1533–1546. [CrossRef]
27. Jiang, W.; Hong, Y.; Zhou, B.; He, X.; Cheng, C. A GAN-based anomaly detection approach for imbalanced industrial time series. *IEEE Access* **2019**, *7*, 143608–143619. [CrossRef]
28. Neupane, D.; Seok, J. Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review. *IEEE Access* **2020**, *8*, 93155–93178. [CrossRef]
29. Li, X.; Zhang, W.; Ding, Q. Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal Process.* **2019**, *161*, 136–154. [CrossRef]
30. Zheng, H.; Wang, R.; Yang, Y.; Yin, J.; Li, Y.; Li, Y.; Xu, M. Cross-domain fault diagnosis using knowledge transfer strategy: A review. *IEEE Access* **2019**, *7*, 129260–129290. [CrossRef]
31. Xu, K.; Kong, X.; Wang, Q.; Yang, S.; Huang, N.; Wang, J. A bearing fault diagnosis method without fault data in new working condition combined dynamic model with deep learning. *Adv. Eng. Inform.* **2022**, *54*, 101795. [CrossRef]
32. Aslam, M.M.; Tufail, A.; De Silva, L.C.; Haji Mohd Apong, R.A.A.; Namoun, A. An improved autoencoder-based approach for anomaly detection in industrial control systems. *Syst. Sci. Control Eng.* **2024**, *12*, 2334303. [CrossRef]
33. Martí, L.; Sanchez-Pi, N.; Molina, J.M.; Garcia, A.C.B. Anomaly detection based on sensor data in petroleum industry applications. *Sensors* **2015**, *15*, 2774–2797. [CrossRef] [PubMed]
34. Zayas-Gato, F.; Jove, E.; Casteleiro-Roca, J.L.; Quintián, H.; Piñón-Pazos, A.; Simić, D.; Calvo-Rolle, J.L. A hybrid one-class approach for detecting anomalies in industrial systems. *Expert Syst.* **2022**, *39*, e12990. [CrossRef]
35. Ahmed, I.; Ahmad, M.; Chehri, A.; Jeon, G. A smart-anomaly-detection system for industrial machines based on feature autoencoder and deep learning. *Micromachines* **2023**, *14*, 154. [CrossRef] [PubMed]
36. Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S.A.; Binder, A.; Müller, E.; Kloft, M. Deep one-class classification. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 4393–4402.
37. Zhou, X.; Hu, Y.; Liang, W.; Ma, J.; Jin, Q. Variational LSTM enhanced anomaly detection for industrial big data. *IEEE Trans. Ind. Inform.* **2020**, *17*, 3469–3477. [CrossRef]

38. Ahmed, C.M.; MR, G.R.; Mathur, A.P. Challenges in machine learning based approaches for real-time anomaly detection in industrial control systems. In Proceedings of the 6th ACM on Cyber-Physical System Security Workshop, Online, 6 October 2020; pp. 23–29.
39. Paolini, D.; Dini, P.; Soldaini, E.; Saponara, S. One-class anomaly detection for industrial applications: A comparative survey and experimental study. *Computers* **2025**, *14*, 281. [CrossRef]
40. Kumar, P.; Rawat, P.; Chauhan, S. Contrastive self-supervised learning: Review, progress, challenges and future research directions. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 461–488. [CrossRef]
41. Sattar, S.; Khan, S.; Khan, M.I.; Akhmediyarova, A.; Mamyrbayev, O.; Kassymova, D.; Oralbekova, D.; Alimkulova, J. Anomaly detection in encrypted network traffic using self-supervised learning. *Sci. Rep.* **2025**, *15*, 26585. [CrossRef] [PubMed]
42. Chalapathy, R.; Menon, A.K.; Chawla, S. Anomaly detection using one-class neural networks. *arXiv* **2018**, arXiv:1802.06360.
43. Lessmeier, C.; Kimotho, J.K.; Zimmer, D.; Sextro, W. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In Proceedings of the PHM Society European Conference, Bilbao, Spain, 5–8 July 2016; Volume 3.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.