A Sequentially Optimized Stacked LSTM Framework for Residual-Based Bearing Fault Detection

ZAIDI, Haseeb <http://orcid.org/0009-0003-7357-5200>, SHENFIELD, Alex <http://orcid.org/0000-0002-2931-8077>, ZHANG, Hongwei <http://orcid.org/0000-0002-7718-021X> and IKPEHAI, Augustine <http://orcid.org/0000-0002-5254-8188>

electronics

MDPI

Article

A Sequentially Optimized Stacked LSTM Framework for Residual-Based Bearing Fault Detection

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

This study presents a healthy-only bearing fault detection framework in which a stacked long short-term memory (LSTM) predictor learns normal vibration dynamics through multi-step forecasting, and deviations between predicted and observed vibration sequences are used for residual-based anomaly detection. The methodological contribution lies in the controlled stage-wise development of the complete prediction-to-decision pipeline, including temporal configuration, predictor selection, residual scoring, and decision-threshold calibration, together with strict bearing-level separation between framework development and final evaluation. The framework was evaluated on the Paderborn bearing dataset using independent healthy bearings for training, validation, calibration, and testing, artificially damaged bearings for detector development, and previously unseen real-damage bearings for final fault evaluation. The finalized detector achieved an ROC–AUC of 0.8385 and a PR–AUC of 0.9200, with a healthy false-positive rate of 0.13% and precision of 0.9971. However, recall at the selected operating threshold was limited to 22.57%, showing that strong anomaly-score discrimination did not translate into equally high fault sensitivity under the conservative global threshold. Performance remained consistent across independent model initializations, while evaluation on a fully reserved healthy bearing produced a 17.95% false-positive rate, highlighting threshold calibration and healthy-domain generalization as the principal limitations of the current framework.

Keywords: bearing anomaly detection; LSTM; stacked LSTM; vibration signal prediction; residual-based anomaly detection; condition monitoring; predictive maintenance; time-series forecasting; healthy-only learning; vibration monitoring

Academic Editor: Ahmed Abu-Siada

Received: 18 August 2026
Revised: 3 September 2026
Accepted: 4 September 2026
Published: 5 September 2026

1. Introduction

Rolling-element bearings are critical components of rotating machinery, and vibration-based monitoring is widely used to identify bearing degradation because faults produce measurable changes in the acquired signals [1–3]. Vibration-based anomaly detection can support intelligent industrial condition monitoring and predictive maintenance by deriving machine-health information from routinely acquired sensor measurements [4,5]. While conventional diagnostic approaches rely on the time, frequency, or time–frequency-domain representations, deep-learning methods can learn diagnostic features directly from vibration data and have shown strong performance in bearing fault diagnosis [6–10]. However, much of this work remains based on supervised classification of predefined fault categories [1,3,9], requiring labelled fault examples that may be limited in practical monitoring

applications [11] and potentially reducing robustness when operating conditions or data distributions change [1,10].

Healthy-only normal-behaviour modelling provides an alternative by learning expected system dynamics from healthy observations and treating deviations as potential anomalies [12]. LSTM models are well suited to this setting because they can capture temporal dependencies and identify abnormal sequences through prediction or reconstruction errors [12–14]. Residual-based approaches similarly use deviations between observed and expected behaviour as evidence of abnormality [15,16]. A complete prediction-based detector therefore requires not only temporal prediction, but also residual-based anomaly scoring and decision-threshold calibration [11,14,16].

Reliable assessment also requires strict separation between framework development and final evaluation. Highly overlapping sliding windows drawn from the same underlying time series can introduce information leakage when distributed across development and test sets [17], while window overlap itself can influence reported performance [18]. Bearing-level separation is particularly important because vibration characteristics can vary with operating conditions [19,20]. Despite progress in prediction-based anomaly detection, the complete detector still depends on interacting choices involving temporal context, forecasting horizon, prediction architecture, residual scoring, and threshold calibration [13,16,21]. These components are often considered individually, even though decisions made at one stage influence the anomaly evidence available to subsequent stages. This creates a practical methodological problem: a detector with an effective forecasting model may still perform poorly if temporal context, residual scoring, or threshold calibration are selected independently. The present work therefore focuses not on introducing a new recurrent architecture, but on developing and evaluating the complete healthy-only prediction-to-decision pipeline through a controlled sequential procedure while preserving independent bearings for final generalization assessment.

Motivated by this gap, this study develops a sequential framework for constructing and evaluating a healthy-only residual-based bearing fault detector using LSTM forecasting as the temporal prediction component. The individual forecasting, residual-scoring, and thresholding components are established techniques; the methodological contribution lies in their controlled stage-wise development, strict separation of bearing roles, and evaluation of the finalized detector on previously unseen real-damage bearings. Normal vibration dynamics are learned through direct multi-step forecasting, and abnormalities are identified from residuals between predicted and observed vibration sequences. Input sequence length, forecasting horizon, prediction architecture, residual-based anomaly scoring, and decision-threshold calibration are investigated sequentially, with each selected component fixed before the subsequent stage is evaluated. The experimental protocol separates framework development from final assessment at the bearing level: independent healthy bearings are used for predictor training, validation, detector calibration, and healthy evaluation; artificially damaged bearings support detector development; and real-damage bearings are withheld for final fault evaluation. The finalized framework is additionally assessed across multiple operating conditions, independent model initializations, and a healthy bearing reserved throughout framework development.

Accordingly, the principal contributions of this study are as follows:

- A healthy-only residual-based bearing monitoring framework is developed in which normal vibration dynamics are learned through direct multi-step forecasting and abnormalities are detected from future-prediction residuals without requiring real-fault labels during predictor training.
- A structured stage-wise development strategy is introduced for the complete prediction-to-decision pipeline, in which temporal input length, forecasting hori-

zon, prediction architecture, residual-based anomaly scoring, and decision-threshold calibration are selected sequentially, with each decision conditioned on components fixed at preceding stages.

- A bearing-level experimental protocol separates predictor training, validation, detector calibration, artificial-fault development, healthy evaluation, and real-fault evaluation, reducing the potential for information leakage associated with highly overlapping vibration windows.
- The finalized detector is evaluated using previously unseen real-damage bearings, multiple operating conditions, independent model initializations, and an additional healthy bearing withheld throughout framework development. It is further compared with PatchTST and iTransformer under an aligned healthy-only prediction-residual protocol, enabling fault-detection performance, reproducibility, healthy-domain generalization, and competitiveness with contemporary time-series architectures to be examined separately.

The remainder of this paper is organized as follows. Section 2 reviews related work in deep-learning-based bearing condition monitoring and anomaly detection. Section 3 presents the proposed methodology and experimental protocol. Section 4 presents and discusses the experimental results. Finally, Section 5 summarizes the main findings, limitations, and directions for future research.

2. Related Work

Research on rolling-bearing condition monitoring has evolved from conventional signal-processing and feature-engineering pipelines toward increasingly data-driven diagnostic frameworks. Traditional vibration-based approaches commonly extract descriptive information from the time, frequency, or time–frequency domains before applying a statistical or machine-learning classifier. Such methods remain useful because bearing defects introduce characteristic changes in measured vibration; however, their effectiveness depends on the selection of suitable preprocessing, signal decomposition, and fault-sensitive features [1–3]. This dependence has motivated the adoption of deep learning, in which useful representations can be learned with less reliance on manually specified diagnostic features. Reviews of modern bearing diagnosis consequently identify convolutional, recurrent, autoencoding, and hybrid architectures as major directions in data-driven condition monitoring [1,3,9].

CNN-based methods have been particularly influential because local vibration patterns can be learned directly from one-dimensional signals or from transformed two-dimensional representations. A compact adaptive 1D CNN has been demonstrated to operate directly on raw bearing vibration measurements, reducing the need for predetermined signal transformations and handcrafted feature selection [6]. In contrast, vibration measurements have been transformed into time–frequency images before deep classification, illustrating an alternative strategy in which signal processing and learned representations are combined [7]. Subsequent work has incorporated multiscale feature extraction and recurrent processing. Multiscale CNN processing has been combined with LSTM layers to exploit both local characteristics and temporal information, while MSCNN–LSTM architectures have also been investigated for rolling-bearing fault diagnosis [8,10]. Collectively, these studies demonstrate the effectiveness of deep architectures for learning fault-sensitive representations, while their diagnostic formulations primarily focus on discrimination among predefined bearing-health or fault classes.

The classification paradigm has produced high benchmark performance, yet several studies and reviews caution that such results should not automatically be interpreted as evidence of deployment robustness. Bearing datasets frequently contain controlled or

artificially introduced faults, and experimental results obtained under fixed operating conditions can deteriorate when speed, load, noise, or measurement distributions change [1,3]. Limited and imbalanced fault observations present an additional challenge for supervised bearing diagnosis, motivating the use of strategies such as data augmentation, cost-sensitive learning, and transfer learning to improve diagnostic performance when representative fault data are scarce [3,9]. Although these approaches can improve supervised diagnosis, they do not remove the fundamental requirement that representative fault information be available during model development.

Recent data-driven fault-diagnosis research has expanded beyond conventional deep-learning architectures through hybrid optimization and Transformer-based modelling. A recent KPCA-ISSA-KELM framework combines kernel principal component analysis (KPCA) for nonlinear feature extraction with an improved Sparrow Search Algorithm (ISSA) to optimize the kernel parameters and regularization coefficient of a kernel extreme learning machine (KELM) [22]. Although developed for fuel-cell fault diagnosis, this approach illustrates the broader integration of nonlinear feature extraction, data-driven optimization, and machine-learning-based fault classification. In parallel, modern time-series architectures provide alternative mechanisms for representing temporal structure. PatchTST uses subsequence-level patches as Transformer input tokens [23], Crossformer models cross-dimensional dependencies in multivariate time series [24], FEDformer combines series decomposition with frequency-domain representations [25], and iTransformer reorganizes the representation of multivariate time series to capture dependencies between variables [26].

Recent advances in time-series foundation models have extended large-scale pre-training to general-purpose temporal representation learning. MOMENT, for example, is pre-trained across heterogeneous time-series datasets and supports downstream tasks including forecasting, classification, imputation, and anomaly detection [27]. Similarly, Timer employs large-scale generative pre-training and formulates forecasting, imputation, and anomaly detection within a unified generative framework [28]. These developments are relevant to condition monitoring because pre-trained temporal representations may reduce reliance on task-specific model training when labelled fault data are limited. However, recent evaluations also indicate that reconstruction- or forecasting-error-based anomaly detection with foundation models does not necessarily outperform simpler anomaly-detection baselines, suggesting that the choice of anomaly evidence and detection strategy remains important [29].

Recent studies have increasingly investigated Transformer-based representations for bearing and rotating-machinery fault diagnosis. PatchTST has been evaluated for supervised bearing fault classification and subsequently extended through MultiPatchTST, where multiple PatchTST models with different patch lengths are combined to reduce dependence on a single predefined patch scale [30]. The resulting framework was evaluated on the Paderborn, CWRU, and MFPT bearing datasets, including experiments under several forms of added industrial noise. More recently, iTransformer has also been evaluated as a comparative Transformer architecture in Paderborn bearing diagnosis, while ACSE-RNformer introduced amplitude-calibrated sequence embedding and response normalization to preserve fault-sensitive vibration responses and stabilize global dependency modelling under variations in signal amplitude [31]. These studies demonstrate the growing application of contemporary Transformer architectures to vibration-based diagnosis and motivate their consideration alongside recurrent approaches. However, these bearing studies primarily formulate diagnosis as supervised classification, whereas the present study considers healthy-only anomaly detection in which abnormality is inferred from future-prediction residuals.

This limitation has encouraged a shift from explicit fault classification toward anomaly detection based on normal-behaviour modelling. Deep time-series anomaly detection encompasses supervised, unsupervised, semi-supervised, and self-supervised learning formulations, with the appropriate formulation largely determined by the availability of labelled normal and anomalous observations [11]. In a healthy-only setting, the model is provided with normal operating data and abnormality is subsequently identified from deviations from the learned normal distribution or temporal behaviour. The scarcity and diversity of anomalous observations also create challenges for generalization, particularly when a detector encounters previously unseen anomaly patterns or changes in the target domain [32]. Recent bearing-fault detection research has also investigated few-shot one-class learning, in which incipient faults are detected without relying on representative labelled examples of the fault classes encountered during operation [33].

Within normal-behaviour modelling, recurrent neural networks provide a natural mechanism for exploiting temporal dependence. Rather than assigning a health label directly to a vibration segment, a recurrent model can learn how a normal sequence evolves and use disagreement between expected and observed behaviour as evidence of abnormality. LSTM networks have been applied to time-series anomaly detection by modelling normal temporal behaviour and evaluating prediction errors [13]. LSTM encoder–decoder formulations have subsequently been used to learn normal multivariate sequences and detect anomalies through reconstruction errors, showing how sequence models can be trained without requiring anomalous examples during model learning [12]. These two formulations represent closely related but distinct strategies: forecasting-based methods infer future observations from preceding context, whereas reconstruction-based approaches attempt to reproduce the observed sequence through a learned latent representation.

Building on these normal-behaviour modelling approaches, stacked LSTM architectures have been investigated for temporal anomaly detection, with prediction errors providing the basis for identifying abnormal sequences [14]. The resulting detection performance also depends on the selected anomaly threshold and its associated trade-off between false-positive and false-negative decisions, emphasizing that prediction performance alone does not define the final detector. Related LSTM-based studies have investigated forecasting and autoencoding formulations for temporal anomaly detection [34], while stacked LSTM encoder–decoder models have also been applied to industrial tool-condition monitoring using normal operating sequences [35].

Collectively, these studies indicate that architecture, temporal context, and detector configuration can materially influence the resulting anomaly decisions. Earlier run-to-failure bearing research also demonstrated that unsupervised representations learned from initially healthy vibration data can be used to construct health indicators that track subsequent bearing degradation [36]. More recently, label-free machinery fault detection has been formulated as a sequential learning problem in which health dynamics are learned from observational state transitions without requiring fault labels. Evaluation across multiple run-to-failure benchmarks further demonstrates the value of preserving temporal state transitions when detecting gradual machinery degradation [37].

Prediction and reconstruction models, however, provide only the first stage of an anomaly detector. Their errors must still be converted into a representation from which abnormality can be judged. Residual-based modelling has therefore been explored across several engineering domains. Autoencoders trained on normal machine sound have used reconstruction residuals as indicators of anomalous machine behaviour [15]. Two-step residual-error frameworks based on variational autoencoders have further demonstrated that residual information can be used not only to identify abnormal system behaviour but also to analyse the variables contributing to detected anomalies [16]. These studies reinforce

an important distinction between learning normal behaviour and making the final anomaly decision: a prediction or reconstruction model generates residual information, whereas the subsequent scoring and decision stages determine how that information is interpreted.

Threshold selection is consequently an important part of the detection pipeline. A continuous residual or anomaly score does not itself specify whether an observation should be regarded as normal or abnormal. Different approaches have been used to convert prediction or reconstruction errors into anomaly decisions, including statistical error modelling, empirically selected thresholds, and calibrated decision rules [11,14,16]. This issue becomes particularly important when the operational objective is not simply to maximize aggregate discrimination but to maintain a low false-alarm rate on predominantly healthy data. Recent work on calibrated industrial anomaly detection has therefore coupled forecasting with held-out calibration data and explicit false-alarm control. For example, causal forecasting has been combined with residual/nonconformity scoring and conformal calibration within an integrated anomaly-detection framework, illustrating that forecasting, score construction, calibration, and decision making can be treated as distinct components of the complete detector [21]. Adaptive and condition-aware thresholding provides a complementary strategy for improving anomaly decisions under changing operating conditions. Recent rotating-machinery anomaly-detection approaches have considered condition-aware thresholds in which the decision boundary is adjusted according to the operating condition or the evolving distribution of normal data [38].

Such approaches differ from the sequential optimization strategy adopted in the present study. Adaptive thresholding modifies the anomaly decision boundary in response to changes in operating state or data distribution, whereas the proposed sequential strategy is an offline development procedure that selects the forecasting configuration, architecture, residual score, and thresholding rule in a predefined order while retaining previously selected components. The two strategies are therefore complementary: condition-aware adaptation could subsequently be incorporated into a residual-based detector when deployment involves substantial operating-condition shifts.

This distinction is especially relevant to prediction-based monitoring because the quality and behaviour of the residual depend on choices made before thresholding. The amount of historical context supplied to the predictor determines the temporal information available for modelling normal dynamics, while the forecasting horizon defines the extent of the future sequence to be predicted. Experimental studies of recurrent multi-step forecasting have shown that prediction behaviour depends on both the available input history and the multi-step forecasting configuration [39,40]. Sliding-window construction is itself a central design element in deep time-series anomaly detection, with historical windows commonly used to provide sequential context [11]. Predictor architecture then determines how these temporal dependencies are represented, after which the multidimensional prediction error must be reduced to an anomaly score and calibrated against healthy behaviour. Consequently, input length, forecasting horizon, predictor architecture, residual aggregation, and decision threshold should not be regarded as unrelated implementation choices; each can influence the characteristics of the anomaly evidence presented to the next stage.

Experimental design introduces an additional concern. Overlapping windows extracted from long vibration recordings are highly related, and evaluation protocols must therefore prevent information from the same physical source from unintentionally appearing on both sides of a development–evaluation boundary. Recent predictive-maintenance research has demonstrated that random partitioning of overlapping sliding windows can introduce information leakage by placing highly similar samples across training and test sets, leading to substantially optimistic estimates of diagnostic performance [17]. This concern is consistent with the broader fault-diagnosis literature, which identifies changing

operating and environmental conditions, domain shifts, data scarcity, and robustness to unseen conditions as persistent challenges for data-driven diagnostic systems [1,3,41]. Thus, incorporating independent healthy data into the evaluation can provide an additional means of assessing whether an anomaly detector is sensitive to variation within the normal operating domain.

The Paderborn University bearing benchmark provides a useful setting for examining these issues because it contains measurements acquired under multiple operating conditions and includes healthy bearings together with artificially induced and real bearing damage [19]. Despite its earlier introduction, the benchmark continues to be used in recent bearing-diagnosis research, enabling contemporary methods to be evaluated against a well-established experimental reference [30,31,42]. It has consequently been used in bearing-diagnosis studies to examine performance under different combinations of rotational speed, load torque, and radial force. The distinction between artificial and real damage is particularly valuable for anomaly-detection research: artificially damaged bearings can support controlled detector development, whereas naturally developed damage provides a more demanding basis for evaluating whether the resulting detector responds to fault behaviour that was not used during its construction.

Taken together, the literature establishes the effectiveness of deep representation learning for bearing diagnosis, recurrent models for learning normal temporal behaviour, and prediction or reconstruction residuals for anomaly detection [6,8,12–16]. At the same time, these strands expose a broader methodological issue. A healthy-only prediction detector is not defined solely by its forecasting network: temporal context, forecasting horizon, predictor architecture, residual representation, score construction, threshold calibration, and the separation of development from final evaluation collectively determine its behaviour. Prior studies establish the importance of many of these elements individually, but their interaction motivates a more systematic treatment of the complete prediction-to-decision pipeline.

3. Methodology

3.1. Overall Framework of the Proposed System

The proposed framework formulates bearing anomaly detection as a healthy-behaviour prediction problem. A stacked LSTM is trained using vibration sequences from healthy bearings to learn normal temporal dynamics. During detection, the trained predictor estimates future vibration samples from the preceding observation window, and the discrepancy between the predicted and measured sequences is used as evidence of abnormal behaviour.

As illustrated in Figure 1a, the framework is developed sequentially by investigating input sequence length, forecasting horizon, prediction architecture, anomaly scoring method, and threshold calibration in a predefined order, with each selected component fixed before the subsequent stage. The resulting fixed prediction-to-decision pipeline, shown in Figure 1b, converts vibration sequences into multi-step forecasts, prediction residuals, anomaly scores, and ultimately threshold-based healthy or faulty decisions.

As summarized in Figure 1c, bearing-level data separation is maintained throughout framework development. Healthy development bearings and artificially damaged bearings are used for framework configuration, whereas held-out healthy and real-damage bearings are reserved for independent evaluation. This separation prevents the final evaluation bearings from influencing predictor training, detector configuration, or threshold calibration.

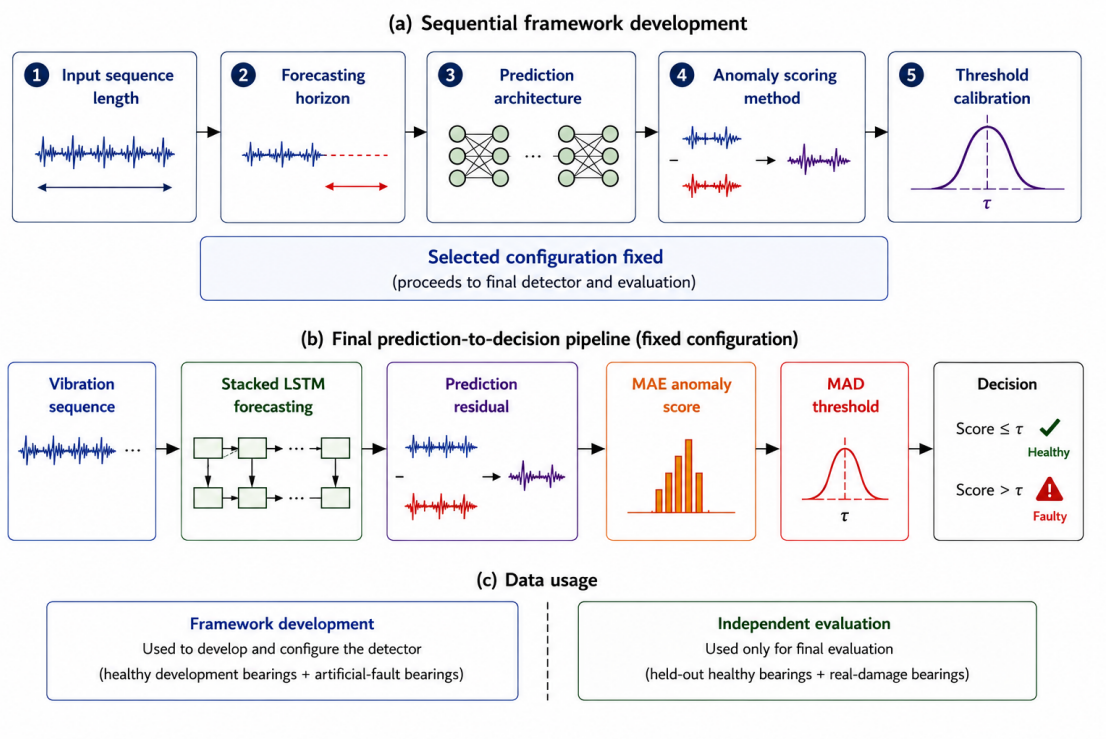

Figure 1. Overview of the proposed healthy-only bearing fault detection framework. (**a**) Sequential development of the prediction-to-decision pipeline, with each selected component fixed before the subsequent stage; (**b**) finalized residual-based detection pipeline; and (**c**) separation of framework-development data from the held-out bearings used for independent evaluation.

3.2. Bearing Dataset and Data Preparation

The experiments were conducted using vibration measurements from the Paderborn University bearing dataset. Ten bearings were included in the study: six healthy bearings (K001–K006), two bearings with artificially induced faults (KI01 and KA01), and two bearings with real damage (KI04 and KA04). In total, 800 vibration recordings were used, with 80 recordings available for each bearing. Each bearing was represented under four operating conditions, with 20 recordings per condition. The corresponding rotational speed, load torque, and radial force settings are summarized in Table 1. The recordings contained approximately 256,000 samples each; across the 800 recordings, approximately 2.05×10^8 raw vibration samples were processed.

Table 1. Operating conditions used in the experiments.

Condition	Dataset Code	Speed (rpm)	Torque (Nm)	Radial Force (N)
C1	N09_M07_F10	900	0.7	1000
C2	N15_M01_F10	1500	0.1	1000
C3	N15_M07_F04	1500	0.7	400
C4	N15_M07_F10	1500	0.7	1000

A bearing-level allocation strategy was established before model development. Rather than randomly dividing windows from the same bearing across different datasets, each bearing was assigned a distinct experimental role, as summarized in Table 2. Healthy bearings K001 and K002 were used to learn the normal vibration dynamics, while K003 was retained for predictor validation. K004 provided independent healthy data for detector calibration, K005 was reserved for healthy test evaluation, and K006 was kept completely outside the development procedure for the later assessment of unseen healthy-domain

behaviour. Artificially damaged bearings KI01 and KA01 were used during detector development, whereas the bearings with real damage, KI04 and KA04, were reserved for final fault evaluation.

Table 2. Role-based bearing allocation used throughout the experimental framework.

Experimental Role	Bearing ID(s)	Purpose
Healthy training	K001, K002	Predictor training
Healthy validation	K003	Predictor validation
Healthy calibration	K004	Detector calibration
Healthy testing	K005	Final healthy evaluation
Reserved healthy	K006	Healthy-domain robustness evaluation
Artificial inner-race fault	KI01	Detector development
Artificial outer-race fault	KA01	Detector development
Real inner-race fault	KI04	Final fault evaluation
Real outer-race fault	KA04	Final fault evaluation

Before sequence generation, each vibration recording was validated for signal availability, finite values, and non-zero variance. The signals were then standardized using a global mean and standard deviation estimated exclusively from the healthy training recordings of bearings K001 and K002. The same training-derived statistics were subsequently applied to all validation, calibration, development, and evaluation recordings, thereby preventing information from held-out bearings from entering the preprocessing stage. Window generation was performed only after the bearing-level allocation had been fixed; consequently, windows originating from a particular bearing could not appear simultaneously in development and final-evaluation datasets.

The normalized vibration signals were converted into supervised forecasting samples using overlapping sliding windows. Each sample consists of a historical input sequence of length L, followed immediately by a prediction target containing the subsequent H samples.

During framework development, input lengths of $L \in \{128, 256, 512\}$ samples and forecasting horizons of $H \in \{1, 8, 16\}$ samples were investigated. The final configuration used an input length of 256 samples and a direct multi-step prediction horizon of 16 samples. Training windows were generated with a stride of 128 samples, providing greater overlap and therefore a larger number of training sequences, whereas a stride of 256 samples was used for validation and evaluation to reduce unnecessary redundancy. The potential leakage associated with overlapping windows arises primarily when highly similar windows from the same bearing are distributed across training and evaluation splits. In the present protocol, bearing-level allocation was completed before window generation; therefore, the different training and evaluation strides affect only the degree of within-bearing redundancy and do not introduce cross-bearing information leakage. With the final $L = 256$ and $H = 16$ configuration, the complete uncapped datasets contained 961,100 supervised windows. The distribution of these windows across the experimental roles is summarized in Table 3. The larger number of training windows results from the smaller training stride of 128 samples, compared with the stride of 256 samples used for validation and evaluation.

To reduce computational cost during the sequential development experiments, model-selection training was restricted to 10,000 training windows, while selected validation, calibration, and artificial-fault evaluations were restricted to budgeted subsets of 5000 windows. These caps were used only during development and configuration selection; the final healthy, real-fault, reproducibility, and reserved-bearing evaluations employed the corresponding uncapped datasets. This distinction allowed the framework to be developed

within the available computational budget while retaining large independent datasets for its final assessment.

Table 3. Number of supervised windows generated for each experimental role using the final temporal configuration.

Experimental Role	Number of Windows
Healthy training	320,443
Healthy validation	80,164
Healthy calibration	80,026
Healthy testing	80,116
Reserved healthy	80,119
Artificial-fault development	160,145
Real-fault evaluation	160,087
Total	961,100

3.3. Stacked LSTM Prediction Model

The prediction model learns the temporal evolution of healthy bearing vibration signals and performs direct multi-step forecasting. The historical vibration input at time step t is represented as

$$\mathbf{X}_t = [x_{t-L+1}, x_{t-L+2}, \ldots, x_t], \tag{1}$$

where $\mathbf{X}_t$ denotes the input sequence ending at time step t, L is the input sequence length, and x_t denotes the standardized vibration amplitude at time step t.

The predictor consists of two stacked unidirectional LSTM layers with 64 hidden units per layer, followed by a fully connected output layer. The complete output sequence of the first LSTM layer is passed to the second layer, and the hidden representation at the final input time step is mapped by the fully connected layer directly to the future prediction horizon. The architecture is illustrated in Figure 2.

The forecasting operation is expressed compactly as

$$\hat{\mathbf{Y}}_t = f_\theta(\mathbf{X}_t) = [\hat{x}_{t+1}, \hat{x}_{t+2}, \ldots, \hat{x}_{t+H}], \tag{2}$$

where f_θ denotes the stacked LSTM predictor parameterized by the learned parameters θ, $\hat{\mathbf{Y}}_t$ is the predicted future vibration sequence, H denotes the forecasting horizon, and $\hat{x}_{t+k}$ represents the predicted vibration amplitude k time steps ahead, for $k = 1, \ldots, H$.

3.4. Residual-Based Fault Detection

Following vibration forecasting, abnormal behaviour is quantified from the discrepancy between the predicted and measured future vibration sequences. For each forecasting window, let $\mathbf{Y}_t = [x_{t+1}, \ldots, x_{t+H}]$ denote the measured future sequence and let $\hat{\mathbf{Y}}_t$ denote the corresponding prediction defined in Section 3.3. The prediction residual is

$$\mathbf{e}_t = \mathbf{Y}_t - \hat{\mathbf{Y}}_t, \tag{3}$$

where $\mathbf{e}_t$ contains the prediction errors across the H-sample forecasting horizon.

The residual vector is reduced to a scalar anomaly score using the mean absolute error (MAE),

$$S_t = \frac{1}{H} \sum_{k=1}^{H} |x_{t+k} - \hat{x}_{t+k}|, \tag{4}$$

where S_t denotes the anomaly score associated with the forecasting window ending at time step t. Lower scores indicate closer agreement with the learned healthy dynamics, whereas larger scores represent greater deviation from the expected vibration behaviour.

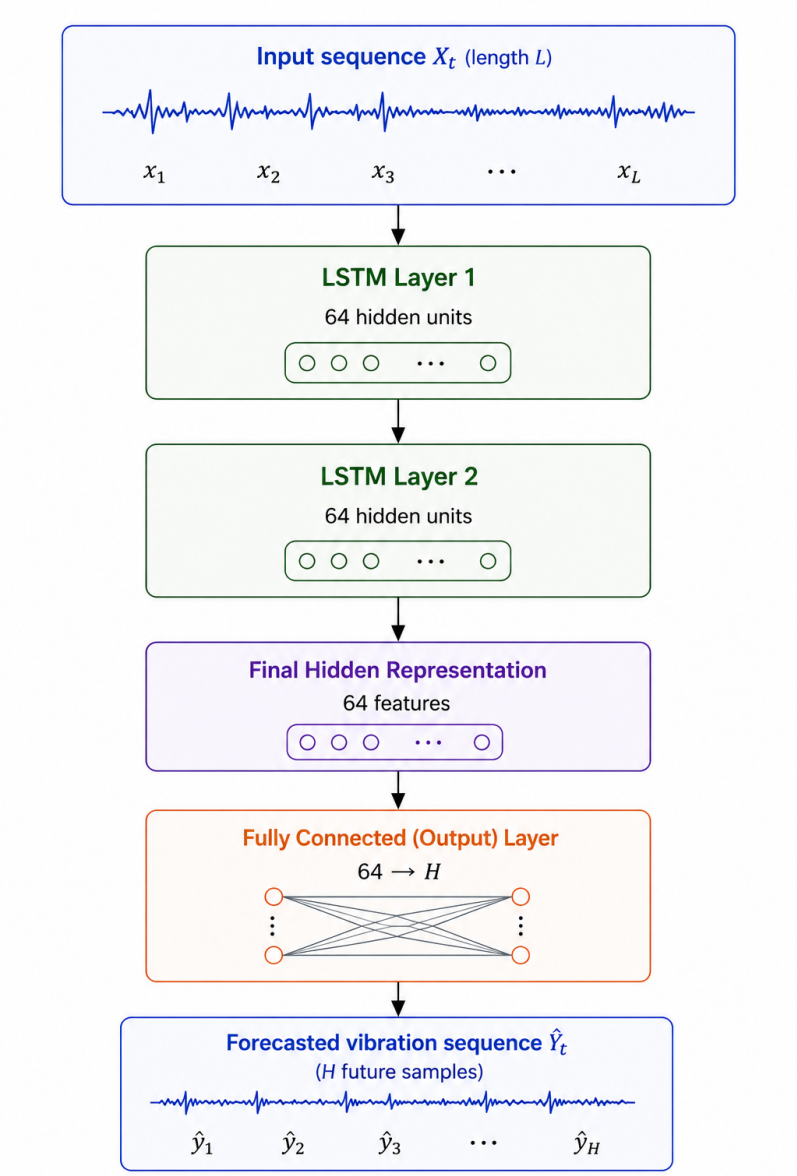

Figure 2. Stacked LSTM architecture for direct multi-step vibration forecasting.

The anomaly score is converted into a binary bearing-health decision using a threshold τ,

$$\hat{c}_t = \begin{cases} \text{Healthy}, & S_t \leq \tau, \\ \text{Faulty}, & S_t > \tau. \end{cases} \tag{5}$$

Here, $\hat{c}_t$ denotes the predicted bearing condition. The anomaly-scoring formulation and threshold configuration were selected during the sequential detector-development stages described in Section 3.5.

3.5. Sequential Framework Development Strategy

The bearing fault detection framework was developed sequentially, with one design component investigated at each stage while previously selected components were kept fixed.

This provided a controlled development path from vibration forecasting to residual-based fault detection. The predefined order follows the dependency of the prediction-to-decision pipeline: the temporal prediction configuration is established before the predictor architecture, residual scoring, and final threshold are selected. The components are not assumed to be strictly independent; rather, the stage-wise procedure provides a computationally tractable alternative to exhaustive joint optimization, with each selection made conditional on the components fixed at preceding stages.

Because the sequential development experiments were repeated across only three independent model initializations, the resulting comparisons were used for configuration selection rather than for claims of statistical superiority. Paired Wilcoxon signed-rank comparisons across the confirmation runs did not establish statistically significant differences at the 0.05 level for the principal predictor-development comparisons. Accordingly, small differences in prediction error are interpreted conservatively, and the limited number of independent runs restricts the strength of formal statistical inference.

The framework development is summarized in Table 4. The ablation studies examined four design components: input sequence length, forecasting horizon, prediction architecture, and residual-based anomaly scoring method. Following these experiments, the decision threshold was calibrated as a separate operating-point selection stage using the finalized predictor and anomaly score. Detailed candidate comparisons for the four ablation studies are provided in Appendix A.

Table 4. Sequential stages used during framework development.

Stage	Component	Selection Basis
1	Input sequence length	Healthy prediction performance
2	Forecasting horizon	Healthy prediction performance
3	Prediction architecture	Healthy prediction performance
4	Anomaly scoring method	Fault discrimination performance
5	Decision threshold	False-alarm and fault-sensitivity trade-off

All framework-development stages used only the designated development data described in Section 3.2. The healthy test bearing K005, reserved healthy bearing K006, and real-damage bearings KI04 and KA04 were excluded from configuration selection and introduced only during the corresponding evaluation experiments. The complete framework was therefore fixed before assessment on these held-out bearings.

3.6. Performance Evaluation and Experimental Implementation

The proposed framework was evaluated using both threshold-dependent classification metrics and threshold-independent ranking metrics. Following thresholding of the anomaly scores, detection performance was assessed using precision, recall, F1-score, false-positive rate (FPR), and specificity. Precision and recall quantify the reliability and sensitivity of fault detection, respectively, while F1-score summarizes their balance. FPR and specificity were additionally reported to characterize the detector's behaviour on healthy data. The terms TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Threshold-independent discrimination was evaluated using the receiver operating characteristic area under the curve (ROC–AUC) and precision–recall area under the curve (PR–AUC). These metrics assess the ability of the anomaly score to distinguish healthy and faulty observations independently of the selected decision threshold.

The framework was implemented in Python version 3.10.9 using the PyTorch version 2.6.0+cpu deep learning library. The predictor was trained using the Adam optimizer with

a learning rate of 1×10^{-3}, zero weight decay, mean squared error (MSE) loss, and a batch size of 256. Training was limited to a maximum of 12 epochs, with gradient clipping at a maximum norm of 1.0. Early stopping with a patience of three epochs was applied to validation MSE, with an improvement of at least 10^{-6} required to reset the patience counter; the checkpoint achieving the lowest validation MSE was retained. The LSTM hidden-state dimension was 64, with the selected stacked architecture using two recurrent layers.

To limit the computational cost of framework development, candidate configurations were first compared in preliminary screening experiments. For the input-length, forecasting-horizon, and prediction-architecture stages, the most promising configurations identified during screening were subsequently evaluated across three independent random initializations. Final selection among the retained configurations was based primarily on mean validation RMSE, with MAE reported as a secondary prediction-error measure. The preliminary candidate comparisons are reported in Appendix A, while the corresponding multi-initialization comparisons are presented in Section 4.

All framework-development experiments followed the sequential development strategy described in Section 3.5. After the complete prediction-to-decision configuration had been fixed, robustness to stochastic model initialization was assessed across three independent random initializations. The resulting performance is reported using the mean and standard deviation across the independent runs.

4. Results and Discussion

4.1. Selection of the Input Sequence Length

Following the preliminary comparison reported in Appendix A.1, the retained input lengths of 128 and 256 samples were evaluated across three independent random initializations. The final selection was based on mean validation RMSE, with MAE used as a secondary performance measure.

Table 5 summarizes the results. The 256-sample configuration achieved a lower mean RMSE than the 128-sample configuration (1.0911 versus 1.0998) and lower RMSE variability (0.0284 versus 0.0606). Although the 128-sample configuration produced a marginally lower mean MAE (0.5881 versus 0.5893), the 256-sample configuration provided better performance according to the primary RMSE criterion.

Table 5. Retained input sequence lengths across three random initializations. Bold indicates the best value for each metric; mean RMSE was the selection criterion.

Input Length	Mean RMSE	Std. RMSE	Mean MAE	Std. MAE
128	1.0998	0.0606	**0.5881**	0.0206
256	**1.0911**	**0.0284**	0.5893	**0.0081**

Figure 3a shows the mean validation RMSE and its variability across the three initializations. The difference in mean RMSE between the two configurations was modest; therefore, the result is interpreted as a configuration choice under the present experimental setting rather than evidence of a general advantage of longer input sequences. Based on the predefined RMSE criterion, $L = 256$ was retained for the subsequent experiments.

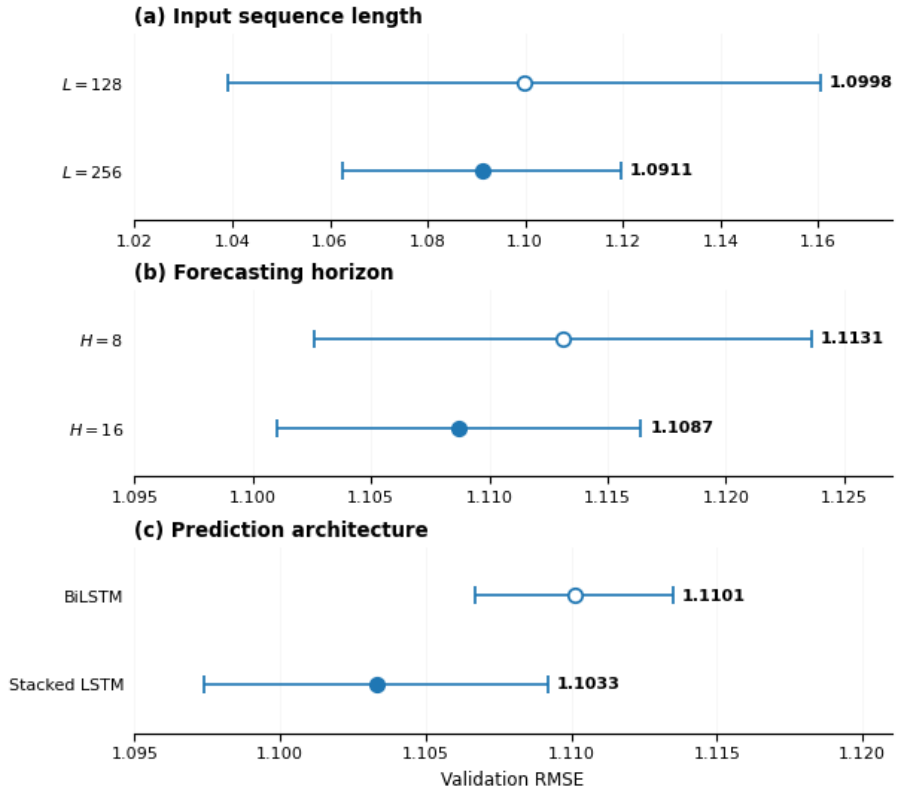

Figure 3. Predictor-development comparisons for (**a**) input sequence length, (**b**) forecasting horizon, and (**c**) prediction architecture. Markers indicate mean validation RMSE across three independent random initializations, with error bars representing standard deviation. Filled markers denote the configurations retained for the subsequent development stage.

4.2. Selection of the Forecasting Strategy

Following the preliminary forecasting-horizon comparison reported in Appendix A.2, the retained direct multi-step configurations with $H = 8$ and $H = 16$ were evaluated across three independent random initializations using the selected input length $L = 256$. Mean validation RMSE was used as the primary selection criterion.

As summarized in Table 6, $H = 16$ achieved a lower mean validation RMSE than $H = 8$ (1.1087 versus 1.1131), with lower RMSE variability (0.0077 versus 0.0105). It also achieved a lower mean MAE (0.5881 versus 0.5980), although $H = 8$ exhibited lower MAE variability.

Table 6. Retained forecasting strategies across three random initializations. Bold indicates the best value for each metric; mean RMSE was the selection criterion.

Forecast Strategy	Mean RMSE	Std. RMSE	Mean MAE	Std. MAE
Direct Multi-step ($H = 8$)	1.1131	0.0105	0.5980	**0.0031**
Direct Multi-step ($H = 16$)	**1.1087**	**0.0077**	**0.5881**	0.0097

Figure 3b shows the corresponding mean RMSE and variability. Although longer forecasting horizons are generally expected to increase prediction difficulty, $H = 16$ produced a modest improvement over $H = 8$ under the present direct multi-step formulation. This result is therefore interpreted as specific to the investigated data and forecasting configuration rather than as evidence that longer forecasting horizons are generally easier to predict. Based on the predefined mean-RMSE criterion, $H = 16$ was retained for the subsequent experiments.

4.3. Selection of the LSTM-Based Prediction Model

Following the candidate architecture comparison reported in Appendix A.3, StackedLSTM and BiLSTM were retained for further evaluation using the selected input length ($L = 256$) and forecasting horizon ($H = 16$). Both architectures were evaluated across three independent random initializations under identical training and evaluation settings, with mean validation RMSE used as the primary selection criterion.

As summarized in Table 7, StackedLSTM achieved a lower mean RMSE than BiLSTM (1.1033 versus 1.1101) and a lower mean MAE (0.5776 versus 0.5913). BiLSTM exhibited slightly lower variability in both metrics; however, the average prediction error was lower for StackedLSTM.

Table 7. Retained LSTM architectures across three random initializations. Bold indicates the best value for each metric; mean RMSE was the selection criterion.

Architecture	Mean RMSE	Std. RMSE	Mean MAE	Std. MAE
BiLSTM	1.1101	**0.0034**	0.5913	**0.0029**
Stacked LSTM	**1.1033**	0.0059	**0.5776**	0.0031

Figure 3c shows the corresponding mean validation RMSE and variability. Although the difference between the two architectures was modest, StackedLSTM achieved the lower mean prediction error under the investigated configuration. Based on the predefined mean-RMSE criterion, StackedLSTM was therefore selected as the final prediction architecture for the subsequent anomaly-detection stages.

4.4. Selection of the Residual-Based Anomaly Scoring Method

Following selection of the StackedLSTM predictor, the residual-based anomaly scoring formulations reported in Appendix A.4 were evaluated using the development data. Among the evaluated scoring formulations, MAE provided the strongest overall performance across the considered ranking and classification metrics. The corresponding performance is summarized in Table 8.

Table 8. Performance of the selected MAE residual-based anomaly scoring method.

PR–AUC	ROC–AUC	F1	Recall	Precision
0.6167	0.5873	0.5305	0.4933	0.5737

MAE achieved the highest value for each of the reported metrics among the evaluated scoring formulations. This consistently stronger performance across both ranking and classification metrics indicates that averaging the absolute prediction residuals provided the most effective fault discrimination among the scoring formulations considered in this study. MAE was therefore selected as the residual-based anomaly score and fixed for the subsequent threshold-selection and final-evaluation stages.

4.5. Threshold Selection for Fault Detection

Following selection of MAE as the residual-based anomaly score, four thresholding families were evaluated: global percentile, global median absolute deviation (MAD), global standard deviation (STD), and operating-condition-specific percentile thresholds. Multiple parameter values within each family were assessed using the development data across the independently trained predictors.

A healthy-validation false-positive rate (FPR) of 5% was used as a study-defined operating objective for controlling false alarms. This criterion and the associated selection procedure were fixed during preliminary workflow development before the final

experiments. Among configurations satisfying this objective, development F1-score and artificial-fault recall were used to guide selection. If no evaluated configuration satisfied the objective, the configuration with mean FPR closest to 5% was retained.

For each candidate rule, the threshold parameters were estimated from the healthy calibration bearing K004, while the resulting false-positive rate used for threshold selection was evaluated separately on the healthy validation bearing K003. Thus, the reported development FPRs represent performance on K003 under thresholds calibrated from K004, rather than the false-positive rate of the calibration bearing itself.

None of the evaluated configurations achieved the 5% FPR objective; therefore, the predefined fallback criterion was applied. Table 9 compares the most conservative configuration from each threshold family. Global MAD with $k = 5.0$ produced the lowest mean healthy-validation FPR (0.1506) and was the closest evaluated configuration to the specified operating objective.

Table 9. Most conservative configuration from each thresholding family. Bold indicates the selected configuration with the lowest mean healthy-validation FPR.

Threshold Method	Parameter	Mean FPR
Global percentile	99.9	0.1681
Condition percentile	99.9	0.1779
Global STD	4.0	0.1822
Global MAD	**5.0**	**0.1506**

The selected MAD configuration produced the lowest healthy-validation FPR among the evaluated threshold candidates. Global MAD with $k = 5.0$ was therefore fixed and applied without further adjustment during final evaluation.

4.6. Final Evaluation on Previously Unseen Bearings

After completion of the sequential optimization procedure, the finalized framework was evaluated without further adjustment using healthy bearing K005 and the previously unseen real-damage bearings KI04 and KA04. All predictor, anomaly-score, and threshold configurations remained fixed during this evaluation.

Table 10 summarizes the final detection performance. The detector achieved a precision of 0.9971 and, on the designated healthy test bearing K005, a specificity of 0.9987 and a false-positive rate (FPR) of 0.0013. The anomaly scores also achieved ROC–AUC and PR–AUC values of 0.8385 and 0.9200, respectively. In contrast, recall was 0.2257, resulting in an F1-score of 0.3681 and indicating limited fault sensitivity at the selected operating threshold.

Table 10. Final detection performance on the designated healthy test bearing K005 and previously unseen real-damage bearings. The false-positive rate on the completely reserved healthy bearing K006 is additionally reported to indicate healthy-domain generalization.

Metric	Value
Precision	0.9971
Recall	0.2257
F1-score	0.3681
Healthy K005 FPR	0.0013
Reserved healthy K006 FPR	0.1795
Specificity (K005)	0.9987
ROC–AUC	0.8385
PR–AUC	0.9200

The substantially lower FPR observed on K005 should be interpreted in the context of the bearing-level evaluation protocol. The threshold was fixed during development using K004 for calibration and K003 for healthy-validation assessment, whereas K005 was a previously unseen healthy bearing and was not used for further threshold adjustment. The difference between the 15.06% development healthy-validation FPR on K003 and the 0.13% final healthy-test FPR on K005 therefore reflects variation in the healthy residual-score distribution between bearings rather than re-calibration on the test data.

The K006 false-positive rate is included in Table 10 to make the observed healthy-domain generalization limitation explicit; K006 was not included in the calculation of the remaining final-evaluation metrics. A detailed analysis of K006 is provided in Section 4.8.

The confusion matrix in Figure 4 shows that only 104 of 80,116 healthy windows were incorrectly classified as anomalous, whereas 36,137 of 160,087 real-fault windows were detected. Correspondingly, the false-negative rate was 0.7743, indicating that a substantial proportion of fault windows remained undetected at the selected operating threshold. The low false-positive count and high false-negative rate reflect the conservative operating behaviour of the fixed threshold.

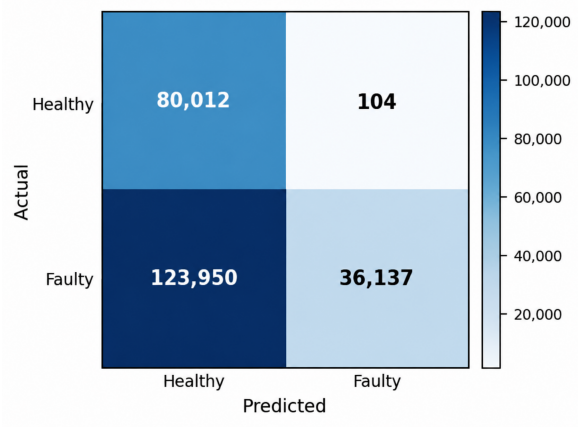

Figure 4. Confusion matrix for the final evaluation on previously unseen healthy and real-damage bearings.

The threshold-independent discrimination performance is shown in Figure 5. The ROC–AUC of 0.8385 and PR–AUC of 0.9200 indicate that the continuous anomaly scores retain substantial discrimination between healthy and faulty windows despite the lower recall obtained at the selected binary decision threshold.

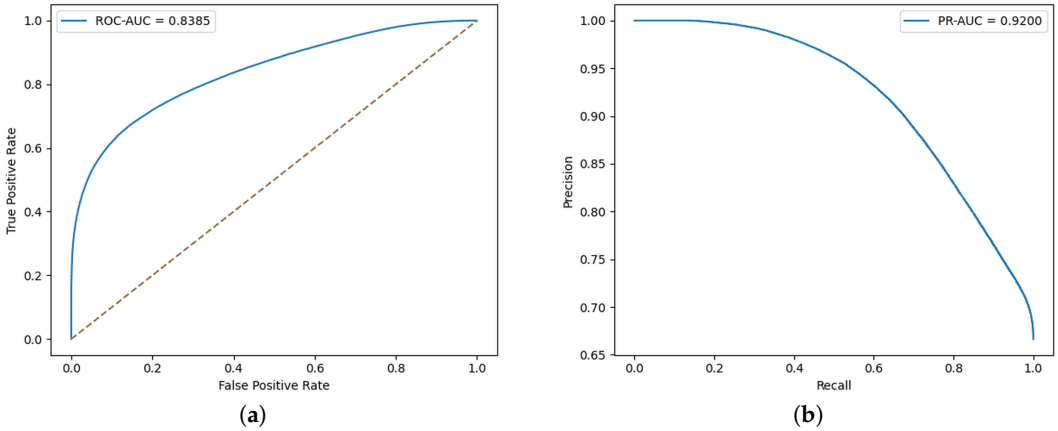

Figure 5. Threshold-independent discrimination performance during final evaluation: (**a**) ROC curve and (**b**) Precision–Recall curve.

Figure 6 further shows the anomaly-score distributions for healthy K005 and the previously unseen real-fault bearings. The distributions exhibit substantial separation but also overlap around the selected decision region, consistent with the difference between the strong ranking metrics and the lower threshold-dependent recall.

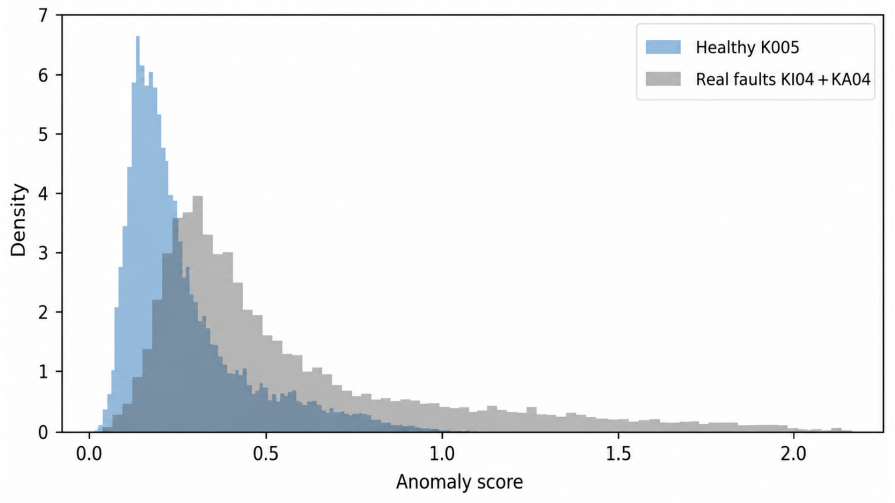

Figure 6. Anomaly-score distributions for healthy K005 and previously unseen real-damage bearings.

Performance varied across operating conditions, as summarized in Table 11. Healthy K005 false-positive rates remained below 0.23% across all four conditions, ranging from 0.020% to 0.225%. In contrast, real-fault detection rates ranged from 6.570% to 29.208%. The highest detection rate occurred under N15_M07_F10, while N09_M07_F10 produced the lowest, indicating that fault detectability remained sensitive to operating condition.

Table 11. Detection performance across the evaluated operating conditions.

Operating Condition	Healthy K005 FPR (%)	Real-Fault Detection Rate (%)
C1	0.020	6.570
C2	0.145	28.178
C3	0.225	26.360
C4	0.130	29.208

Overall, the designated final evaluation demonstrates a very low false-alarm rate on K005 and substantial anomaly-ranking capability on the unseen evaluation bearings, but also reveals limited fault sensitivity at the selected threshold. The substantially higher false-positive rate observed on the reserved healthy bearing K006 further indicates that this low false-alarm behaviour does not generalize uniformly across unseen healthy bearings. The contrast between the high PR–AUC and relatively low recall indicates that threshold calibration and healthy-domain generalization remain important limitations of the finalized detector.

4.7. Robustness Across Independent Random Initializations

To assess reproducibility, the finalized framework was independently trained and evaluated across three independent random initializations while all selected design components remained fixed. This experiment therefore examined the sensitivity of the final detector to stochastic model initialization and training.

Table 12 summarizes the results. Performance remained highly consistent across the three runs, with F1-scores ranging from 0.3639 to 0.3674 and precision exceeding 0.997

for every run. Healthy false-positive rates also remained close to 0.12%, while only small variations were observed in PR–AUC and ROC–AUC.

Table 12. Final detection performance across three independent random initializations.

Run	F1	Recall	Precision	FPR	PR–AUC	ROC–AUC
1	0.3639	0.2225	0.9972	0.001248	0.9200	0.8385
2	0.3674	0.2252	0.9972	0.001248	0.9197	0.8377
3	0.3669	0.2248	0.9974	0.001173	0.9214	0.8417
Mean	0.3661	0.2242	0.9973	0.001223	0.9204	0.8393
Std.	0.0019	0.0014	0.0001	0.000043	0.0009	0.0021

The small standard deviations across all reported metrics indicate that the final performance was relatively insensitive to random initialization under the investigated training procedure. In particular, the consistency of both threshold-dependent and threshold-independent metrics supports the reproducibility of the reported detector behaviour across three independent random initializations.

4.8. Generalization to the Reserved Healthy Bearing

To assess healthy-domain generalization, bearing K006 was reserved throughout framework development and introduced only after the complete prediction-to-decision pipeline had been finalized. K006 therefore played no role in predictor training, detector calibration, or configuration selection.

Table 13 compares the anomaly-score statistics of the calibration bearing K004, healthy test bearing K005, and reserved healthy bearing K006. K004 and K005 produced relatively similar score distributions and low false-positive rates of 0.04% and 0.13%, respectively. In contrast, K006 exhibited a higher mean anomaly score (0.6476) and a false-positive rate of 17.95%, indicating substantially reduced detector robustness on the reserved healthy bearing.

Figure 7 shows that the K006 anomaly-score distribution is shifted towards larger values and exhibits a longer upper tail than those of K004 and K005. This shift is consistent with the substantially higher false-positive rate observed for the reserved bearing.

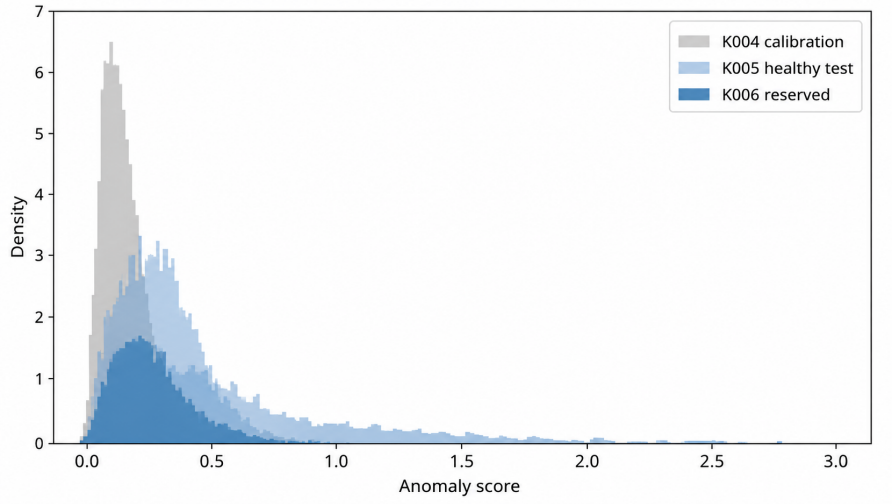

Figure 7. Comparison of anomaly-score distributions for the calibration bearing (K004), healthy test bearing (K005), and reserved healthy bearing (K006).

Table 13. Anomaly-score statistics across healthy bearings. Bold indicates the largest value for each metric, highlighting the shift for K006.

Bearing	Mean Score	95th Percentile	99th Percentile	False-Positive Rate
K004 (Calibration)	0.3109	0.5693	0.7081	0.0004
K005 (Healthy Test)	0.3444	0.6192	0.7776	0.0013
K006 (Reserved)	**0.6476**	**1.5391**	**2.1049**	**0.1795**

The K006 false-positive rate was also examined separately across the four operating conditions. As shown in Figure 8, the rates ranged from 15.81% to 20.12%, indicating that the degradation was not confined to a single operating condition.

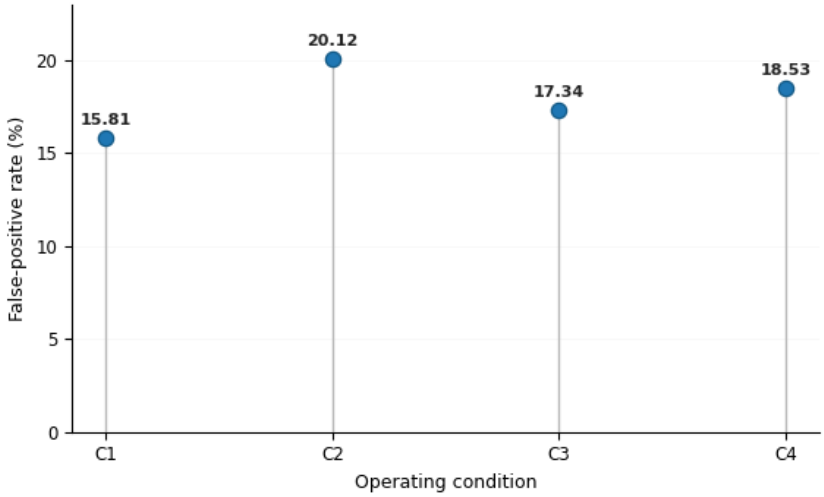

Figure 8. False-positive rate of the reserved healthy bearing (K006) across the evaluated operating conditions.

The K006 recordings also exhibited higher average signal standard deviation and RMS values than K004 and K005 across the evaluated operating conditions, supporting the interpretation that the increased anomaly scores were associated with healthy-domain variation. These results reveal an important limitation of the fixed detector: although false-positive rates remained very low for K004 and K005, performance deteriorated substantially on a healthy bearing that was completely excluded from framework development. This finding highlights the importance of representing healthy-bearing variability during calibration and motivates future investigation of broader healthy training data or adaptive thresholding strategies.

4.9. Comparison with Contemporary Time-Series Models

To assess whether the selected recurrent architecture remains competitive with more recent time-series models, PatchTST and iTransformer were evaluated as external forecasting baselines. These architectures were selected because they represent contemporary patch-based and inverted-Transformer approaches and have also been considered in recent bearing-diagnosis research [23,26,30,31]. For a controlled comparison, only the forecasting model was changed; the healthy-only prediction-residual detection framework was otherwise kept aligned with the proposed method, including the input window and prediction horizon, residual-based anomaly scoring, calibration procedure, and held-out bearing evaluation. Each architecture was evaluated over three independent runs, and Table 14 reports the mean and standard deviation across these runs.

Table 14. Comparison with contemporary time-series models. Bold indicates the best performance for each metric.

Model	ROC-AUC	PR-AUC	Precision	Recall	Healthy FPR
Stacked LSTM	**0.8393 ± 0.0021**	**0.9204 ± 0.0009**	**0.9973 ± 0.0001**	0.2242 ± 0.0014	**0.00122 ± 0.00004**
PatchTST	0.8264 ± 0.0054	0.9132 ± 0.0022	0.9959 ± 0.0004	0.2235 ± 0.0021	0.00183 ± 0.00018
iTransformer	0.8254 ± 0.0030	0.9146 ± 0.0017	0.9961 ± 0.0003	**0.2480 ± 0.0056**	0.00195 ± 0.00016

Under the aligned evaluation protocol, the Stacked LSTM achieved the highest mean ROC-AUC (0.8393), PR-AUC (0.9204), and precision (0.9973), together with the lowest healthy false-positive rate (0.00122). PatchTST produced similar recall to the Stacked LSTM but slightly lower ranking performance, whereas iTransformer achieved the highest recall (0.2480) while exhibiting lower ROC-AUC and a higher healthy false-positive rate. The three architectures therefore demonstrated broadly comparable anomaly-detection performance under the aligned evaluation protocol, with different trade-offs across the reported metrics. Given the small differences and the variability observed across the three independent runs, these results are interpreted descriptively and do not establish statistical superiority of one architecture over the others.

5. Conclusions

This study presented a healthy-only bearing fault detection framework based on multi-step stacked-LSTM forecasting and residual-based anomaly detection, with the complete prediction-to-decision pipeline developed through a controlled sequential procedure. Rather than formulating bearing monitoring as a supervised fault-classification problem, the proposed approach learns healthy temporal behaviour from vibration sequences and identifies abnormal behaviour from the resulting prediction residuals. A central aspect of the methodology was the sequential development of the complete prediction-to-decision pipeline, in which the input sequence length, forecasting horizon, prediction architecture, residual-based anomaly score, and decision threshold were evaluated in successive stages and frozen after selection. Bearing-level separation was maintained throughout development so that the real-damage bearings and an additional reserved healthy bearing remained unseen during framework optimization.

The finalized framework demonstrated strong anomaly-ranking performance on previously unseen healthy and real-damage bearings, achieving ROC–AUC and PR–AUC values of 0.8385 and 0.9200, respectively, together with a healthy false-positive rate of 0.13%. However, recall at the selected operating threshold was only 22.57%, showing that the conservative global threshold favoured suppression of false alarms at the expense of detecting a substantial proportion of real-fault windows. These results therefore establish that the residual scores contain useful fault-discrimination information, while also showing that effective anomaly ranking alone is insufficient to guarantee strong threshold-dependent fault sensitivity. The controlled comparison with contemporary forecasting architectures showed broadly comparable performance among the Stacked LSTM, PatchTST, and iTransformer under the same healthy-only prediction-residual protocol, with different trade-offs across anomaly ranking, precision, recall, and healthy false-positive rate.

The additional experiments provide important context for this result. Performance was highly consistent across three independent model initializations, with a mean F1-score of 0.3661 and a standard deviation of 0.0019, indicating that the observed detector behaviour was not dependent on a favourable random initialization. Fault detectability nevertheless varied across operating conditions, demonstrating that changes in operating regime influence the magnitude and separability of prediction residuals. More importantly,

evaluation on the completely reserved healthy bearing K006 produced a false-positive rate of 17.95%, compared with 0.13% for the designated healthy test bearing K005. This result reveals that a globally calibrated residual threshold can be sensitive to previously unseen healthy-domain variation even when excellent specificity is obtained on the original healthy test distribution.

These findings highlight an important distinction between anomaly-score discrimination and the final threshold-dependent detection decision. The high PR–AUC obtained on real faults indicates that useful fault-related information is present in the learned residual scores, whereas the comparatively low recall shows that a conservative global threshold does not fully exploit this separation. At the same time, the reserved-bearing experiment demonstrates that reducing the threshold solely to increase fault sensitivity could increase false alarms when the healthy distribution changes. Consequently, threshold calibration and healthy-domain generalization emerge as important challenges for translating prediction-based bearing anomaly detection into robust condition-monitoring systems.

Future work should therefore investigate adaptive or operating-condition-aware thresholding strategies capable of balancing fault sensitivity and false-alarm control under changing healthy distributions. Extending healthy-behaviour learning across a broader range of bearings and operating regimes may further improve robustness to bearing-to-bearing variation. Evaluation on additional run-to-failure and industrial datasets would also be valuable for determining whether the observed behaviour generalizes beyond the Paderborn dataset. Finally, incorporating temporal persistence or recording-level decision rules, rather than treating individual prediction windows independently, may provide a more practically meaningful basis for maintenance decisions.

Overall, the results demonstrate that sequential development of the prediction-to-decision pipeline provides a transparent and reproducible framework for healthy-only bearing fault detection, while also identifying threshold calibration and healthy-domain shift as key directions for further development.

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 research received no external funding.

Data Availability Statement: The Paderborn Bearing Dataset analyzed in this study is publicly available from the Paderborn University Bearing Data Center. No new datasets were generated during this study. The implementation code will be made publicly available on GitHub upon publication.

Acknowledgments: The authors gratefully acknowledge support from the Advanced Food Innovation Centre (AFIC) and Sheffield Hallam University.

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

Appendix A. Ablation Studies

The framework configuration was established sequentially using only the development data described in Section 3.2. Candidate configurations were compared while previously selected components were kept fixed, allowing the influence of each design choice to be assessed separately. The final evaluation bearings were excluded from all configuration-selection experiments.

The ablation studies considered four components of the framework: input sequence length, forecasting horizon, LSTM architecture, and residual-based anomaly scoring

method. The following subsections summarize the preliminary candidate comparisons used to identify the configurations retained for subsequent confirmation and framework-development stages.

Appendix A.1. Input Sequence Length

Three candidate input sequence lengths were evaluated under an identical UniLSTM one-step forecasting configuration as a preliminary configuration screening. The screening results are summarized in Table A1.

Table A1. Preliminary screening of candidate input sequence lengths.

Input Length	Validation RMSE	Outcome
128	1.0930	Retained
256	1.1360	Retained
512	1.1408	Not retained

The 128- and 256-sample configurations were retained for repeated evaluation. Their three-run comparison and the final input-length selection are reported in Section 4.1.

Appendix A.2. Forecasting Horizon

Following the input-length analysis, three forecasting horizons were compared: one-step forecasting ($H = 1$) and direct multi-step forecasting with horizons of ($H = 8$) and ($H = 16$). All configurations used an input sequence length of $L = 256$ and otherwise identical model and training settings. The preliminary screening results are summarized in Table A2.

Table A2. Preliminary screening of candidate forecasting horizons.

Forecasting Horizon	Validation RMSE	Outcome
$H = 1$	1.1360	Not retained
$H = 8$	1.1174	Retained
$H = 16$	1.0940	Retained

The two direct multi-step configurations, $H = 8$ and $H = 16$, were retained for further evaluation. Their final comparison and forecasting-horizon selection are reported in Section 4.2.

Appendix A.3. Candidate LSTM Architectures

Following selection of the temporal configuration, six LSTM-family architectures were compared using the selected input length ($L = 256$) and forecasting horizon ($H = 16$), with the remaining training and evaluation settings kept unchanged. The preliminary comparison is summarized in Table A3.

Table A3. Preliminary comparison of candidate LSTM architectures.

Architecture	RMSE	MAE	Outcome
StackedLSTM	1.0916	0.5728	Retained
BiLSTM	1.0926	0.5747	Retained
UniLSTM	1.0940	0.5775	Not retained
CNN-LSTM	1.0978	0.5838	Not retained
Encoder–Decoder LSTM	1.1007	0.5899	Not retained
Residual LSTM	1.1124	0.6052	Not retained

StackedLSTM and BiLSTM achieved the lowest prediction errors and were therefore retained for further evaluation. Their final comparison and prediction-model selection are reported in Section 4.3.

Appendix A.4. Residual-Based Anomaly Scoring Methods

Following selection of the StackedLSTM predictor, six residual-based anomaly scoring methods were compared: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Residual Energy, Mahalanobis Distance, and Maximum Absolute Residual (MaxAbs). The comparison is summarized in Table A4.

Table A4. Comparison of the evaluated residual-based anomaly scoring methods. Bold indicates the best performing scoring method

Method	PR–AUC	ROC–AUC	F1	Recall	Precision
MAE	**0.6167**	**0.5873**	**0.5305**	**0.4933**	**0.5737**
RMSE	0.5700	0.5457	0.5009	0.4739	0.5313
MSE	0.5700	0.5457	0.5009	0.4739	0.5313
Residual Energy	0.5700	0.5457	0.5009	0.4739	0.5313
Mahalanobis	0.5141	0.4943	0.4415	0.4076	0.4817
Maximum Absolute Residual	0.5210	0.4979	0.4208	0.3777	0.4750

MAE achieved the highest PR–AUC, ROC–AUC, F1-score, recall, and precision among the evaluated scoring methods and was therefore selected as the residual-based anomaly score for the subsequent threshold analysis. The corresponding final selection is discussed in Section 4.4.

References

1. Zhang, S.; Zhang, S.; Wang, B.; Habetler, T.G. Deep learning algorithms for bearing fault diagnostics—A comprehensive review. *IEEE Access* **2020**, *8*, 29857–29881. [CrossRef]
2. Li, J.; Luo, W.; Bai, M. Review of research on signal decomposition and fault diagnosis of rolling bearing based on vibration signal. *Meas. Sci. Technol.* **2024**, *35*, 092001. [CrossRef]
3. 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]
4. Czeczot, G.; Rojek, I.; Mikołajewski, D.; Sangho, B. AI in IIoT management of cybersecurity for industry 4.0 and industry 5.0 purposes. *Electronics* **2023**, *12*, 3800. [CrossRef]
5. Md, W.; Tanzila, M.N.; Tasnim, H.S.; Shirin, S.; Jan, T.; Barros, A. Enhancing IIoT Security Using Digital Twins in Industry 5.0: A Systematic Literature Review. *Information* **2026**, *17*, 209. [CrossRef]
6. Eren, L.; Ince, T.; Kiranyaz, S. A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier. *J. Signal Process. Syst.* **2019**, *91*, 179–189. [CrossRef]
7. 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]
8. Chen, X.; Zhang, B.; Gao, D. Bearing fault diagnosis base on multi-scale CNN and LSTM model. *J. Intell. Manuf.* **2021**, *32*, 971–987. [CrossRef]
9. Li, X.; Ma, Z.; Yuan, Z.; Mu, T.; Du, G.; Liang, Y.; Liu, J. A review on convolutional neural network in rolling bearing fault diagnosis. *Meas. Sci. Technol.* **2024**, *35*, 072002. [CrossRef]
10. Wu, C.; Zheng, S. Fault diagnosis method of rolling bearing based on MSCNN-LSTM. *Comput. Mater. Contin.* **2024**, *79*, 4395–4411. [CrossRef]
11. Zamanzadeh Darban, Z.; Webb, G.I.; Pan, S.; Aggarwal, C.; Salehi, M. Deep learning for time series anomaly detection: A survey. *ACM Comput. Surv.* **2024**, *57*, 1–42. [CrossRef]
12. Malhotra, P.; Ramakrishnan, A.; Anand, G.; Vig, L.; Agarwal, P.; Shroff, G. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv* **2016**, arXiv:1607.00148.
13. Malhotra, P.; Vig, L.; Shroff, G.; Agarwal, P. Long short term memory networks for anomaly detection in time series. In Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015), Bruges, Belgium, 22–24 April 2015.

14. Thill, M.; Däubener, S.; Konen, W.; Bäck, T.; Barancikova, P.; Holena, M.; Horvat, T.; Pleva, M.; Rosa, R. Anomaly Detection in Electrocardiogram Readings with Stacked LSTM Networks. In Proceedings of the 19th Conference Information Technologies-Applications and Theory (ITAT 2019), Donovaly, Slovakia, 20–24 September 2019; pp. 17–25.
15. Oh, D.Y.; Yun, I.D. Residual error based anomaly detection using auto-encoder in SMD machine sound. *Sensors* **2018**, *18*, 1308. [CrossRef] [PubMed]
16. Gonzalez-Muniz, A.; Diaz, I.; Cuadrado, A.A.; Garcia-Perez, D.; Perez, D. Two-step residual-error based approach for anomaly detection in engineering systems using variational autoencoders. *Comput. Electr. Eng.* **2022**, *101*, 108065. [CrossRef]
17. Shamim, M.M.R.; Nuruzzaman, M.; Ferdus, Z.; Ahmed, M.R.; Anward, A.N.; Tooneer, M.; Khan, J.U.; Hossen, K. Leakage-Robust Evaluation and Data-Scale Sensitivity of Attention-Enhanced Multi-Task Learning for Joint Fault Diagnosis and Remaining Useful Life Estimation. *arXiv* **2026**, arXiv:2607.16493.
18. Wosiak, A.; Sumiński, M.; Żykwińska, K. Impact of temporal window shift on EEG-based machine learning models for cognitive fatigue detection. *Algorithms* **2025**, *18*, 629. [CrossRef]
19. 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 European Conference of the Prognostics and Health Management Society, Bilbao, Spain, 5–8 July 2016.
20. Geetha, G.; Geethanjali, P. An efficient method for bearing fault diagnosis. *Syst. Sci. Control Eng.* **2024**, *12*, 2329264. [CrossRef]
21. Mudasir, M.; Asiri, Y.; Ameer, I.; Reshan, M.S.A.; Almansour, H.; Awan, K.A.; Shaikh, A. RBC-AD: Conformal anomaly detection with explicit false-alarm control for the Tennessee Eastman Process. *Connect. Sci.* **2026**, *38*, 2707826. [CrossRef]
22. Quan, R.; Liang, W.; Wang, J.; Li, X.; Chang, Y. An enhanced fault diagnosis method for fuel cell system using a kernel extreme learning machine optimized with improved sparrow search algorithm. *Int. J. Hydrog. Energy* **2024**, *50*, 1184–1196. [CrossRef]
23. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv* **2022**, arXiv:2211.14730.
24. Zhang, Y.; Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In Proceedings of the The Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
25. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Proceedings of the 39th International Conference on Machine Learning (ICML 2022), Baltimore, MD, USA, 17–23 July 2022; pp. 27268–27286.
26. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. Itransformer: Inverted transformers are effective for time series forecasting. *Proc. Int. Conf. Learn. Represent.* **2024**, *2024*, 11116–11140.
27. Goswami, M.; Szafer, K.; Choudhry, A.; Cai, Y.; Li, S.; Dubrawski, A. Moment: A family of open time-series foundation models. *arXiv* **2024**, arXiv:2402.03885.
28. Liu, Y.; Zhang, H.; Li, C.; Huang, X.; Wang, J.; Long, M. Timer: Generative pre-trained transformers are large time series models. *arXiv* **2024**, arXiv:2402.02368.
29. Zhu, X.; Carpentier, L.; Verbeke, M. When Foundation Models are One-Liners: Limitations and Future Directions for Time Series Anomaly Detection. In Proceedings of the Fourteenth International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026.
30. Ko, S.; Lee, S. Multi-patch time series transformer for robust bearing fault detection with varying noise. *Appl. Sci.* **2025**, *15*, 1257. [CrossRef]
31. Yan, Y.; Shang, T.; Zeng, K.; Yang, S.; Cheng, H.; Quan, W. ACSE-RNformer: Amplitude-Calibrated Sequence Embedding and Response-Normalized Transformer for Vibration-Based Rotating Machinery Fault Diagnosis. *Sensors* **2026**, *26*, 5275. [CrossRef] [PubMed]
32. Zhou, M.; Chen, Y.; Sun, S. L-PromptCTRL: Learnable prompting for contextual residual anomaly detection. *Digit. Signal Process.* **2025**, *171*, 105841. [CrossRef]
33. Ding, Y.; Cao, Y.; Miao, Q.; Zhao, X.; Ge, Y.; Hu, C. UFOCC: Contamination-Robust few-shot one-class learning for online detection of Bearing incipient faults. *Measurement* **2026**, 120402. [CrossRef]
34. Nguyen, H.D.; Tran, K.P.; Thomassey, S.; Hamad, M. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. *Int. J. Inf. Manag.* **2021**, *57*, 102282. [CrossRef]
35. Oshida, T.; Murakoshi, T.; Zhou, L.; Ojima, H.; Kaneko, K.; Onuki, T.; Shimizu, J. Development and implementation of real-time anomaly detection on tool wear based on stacked LSTM encoder-decoder model. *Int. J. Adv. Manuf. Technol.* **2023**, *127*, 263–278. [CrossRef]
36. Hasani, R.M.; Wang, G.; Grosu, R. An automated auto-encoder correlation-based health-monitoring and prognostic method for machine bearings. *arXiv* **2017**, arXiv:1703.06272.
37. Neupane, D.; Bouadjenek, M.R.; Dazeley, R.; Aryal, S. Label-free Industrial Fault Detection via Adversarial Inverse Reinforcement Learning: A System for Run-to-Failure Prognostics. *arXiv* **2026**, arXiv:2607.22987.

38. Hu, C.; Wu, J.; Sun, C.; Chen, X.; Nandi, A.K.; Yan, R. Unified flowing normality learning for rotating machinery anomaly detection in continuous time-varying conditions. *IEEE Trans. Cybern.* **2024**, *55*, 221–233. [CrossRef] [PubMed]
39. Sangiorgio, M.; Dercole, F. Robustness of LSTM neural networks for multi-step forecasting of chaotic time series. *Chaos Solitons Fractals* **2020**, *139*, 110045. [CrossRef]
40. He, X.; Shi, S.; Geng, X.; Yu, J.; Xu, L. Multi-step forecasting of multivariate time series using multi-attention collaborative network. *Expert Syst. Appl.* **2023**, *211*, 118516. [CrossRef]
41. Marzouk, M.; Mansouri, M.; Kahloul, A.A.; Kermani, M.; Sakly, A.; Nounou, H. Towards Intelligent Fault Diagnosis in Renewable Energy Systems: Integrating Data-Driven and Model-Based Approaches under Variable Environmental Conditions. *IEEE Access* **2026**, *14*, 115386–115423. [CrossRef]
42. Zaidi, S.H.H.; Shenfield, A.; Zhang, H.; Ikpehai, A. Self-Supervised CNN–Transformer Anomaly Detection for Bearing Health Monitoring. *Processes* **2026**, *14*, 2452. [CrossRef]