U.Porto Journal of Engineering, 12:2 (2026) 132-155
ISSN 2183-6493
DOI: 10.24840/2183-6493_0012-002_003840

Received: 10 July, 2025
Accepted: 22 September, 2026
Published: 21 October, 2026

A Hybrid CNN-BiLSTM-Multihead Attention Framework for Robust Broken Tooth Fault Diagnosis in Gearboxes Under Noise Conditions

Canan Tastimur
Department of Computer Engineering, Faculty of Engineering-Architecture, Erzincan Binali Yildirim University, Erzincan 24100, Türkiye (ctastimur@erzincan.edu.tr) ORCID 0000-0002-3714-6826

Author Keywords

Gearbox fault diagnosis, Vibration signal analysis, Hybrid deep learning, Bidirectional long short-term memory, Multi-head self-attention.

Type: Research Article

Open Access
Peer Reviewed
CC BY

Abstract

Early detection of defects in rotary machinery is critical to preventing system downtime and cost losses. This study proposes a hybrid deep learning model combining 1D CNN, Bidirectional LSTM, and Multi-Head Attention mechanism to classify healthy and fractured tooth conditions using vibration signals obtained from the SpectraQuest simulator. The model, which works directly on raw signals, eliminates manual feature extraction while achieving noise resilience through rich data enhancement techniques such as amplitude scaling and frequency masking. Training results show that proposed model achieved a training accuracy of 99.06%, a validation accuracy of 98.33%, and a test accuracy of 98.00%. The relatively minor differences between the training, validation, and test accuracies indicate consistent performance across the evaluated data splits and suggest that the model maintains good generalization without substantial overfitting. These achievements have been supported by 5-fold cross-validation and are presented in detail and comparatively through durability tests performed at different noise levels (5-20 dB SNR) and ablation studies revealing the contribution of architectural components.

1. Introduction

With Industry 4.0, smart manufacturing systems are rapidly developing. This has made it necessary to increase the safety of rotating machinery. Since failures in critical equipment such as electric motors, bearings, gearboxes, pumps, and turbines in rotating machinery can cause the machine to stop, gearbox fault diagnosis is crucial to prevent production system downtime. Broken teeth, wear, and other faults in gearboxes, widely used in numerous sectors such as mining, automotive, aviation, wind energy, iron and steel production, and robotics systems, are diagnosed using traditional methods, which can lead to overlooking the fault and often require expert knowledge, resulting in limited performance (Guo et al., 2021). In recent years, with the development of deep learning-based models, this technology has also shown superior success in fault diagnosis.

Gearboxes in rotating machinery are among the most critical power transmission elements, safely transmitting mechanical power at varying speeds. Gearboxes are widely used in wind turbines, electric motors, industrial robots, and conveyor systems. Continuous changes in load within a gearbox can lead to various mechanical failures due to friction, fatigue, and vibration. Gearbox failures are initially difficult to detect because they generate vibration signals with low amplitude (Kumar et al., 2020). As failure progresses, the amplitude of the vibration signal increases, making it easier to detect. However, as the failure progresses, maintenance costs

increase and equipment lifespan decreases. Therefore, early detection of failures before they worsen is of immense importance.

This study proposes a novel hybrid deep learning architecture that is lightweight, modular, and computationally efficient, working directly on raw vibration data to reliably determine healthy and faulty operating conditions from gearbox vibration signals. Unlike traditional approaches that involve manual feature extraction and complex signal processing steps, the proposed model automatically learns distinguishing features end-to-end from raw vibration signals. The first stage of the model, a one-dimensional Convolutional Neural Network (1D-CNN), efficiently extracts local temporal patterns, abrupt changes, and short-term fault characteristics in the vibration signals. Subsequently, using bidirectional Long Short-Term Memory (Bi-LSTM) layers, not only past signal information but also information from future time steps is evaluated, thus enabling more successful modeling of long-term temporal dependencies. In addition, the Multi-Head Self-Attention mechanism integrated into the model adaptively learns the significance levels of different time steps along the signal, highlighting the representational components that contribute most to fault diagnosis. The novelty of the proposed approach lies not in the individual use of 1D-CNN, Bi-LSTM, or Multi-Head Self-Attention, but in their task-oriented integration into a unified end-to-end gearbox diagnosis framework. The CNN captures local temporal patterns, the Bi-LSTM models bidirectional temporal dependencies, and the attention mechanism adaptively emphasizes diagnostically relevant representations. This coordinated design combines complementary temporal information while retaining a lightweight architecture and eliminating manual feature engineering. Thus, local features learned by CNN, global temporal relationships modeled by Bi-LSTM, and distinctive information selected by the attention mechanism are efficiently combined within a single integrated architecture. Ablation testing was also carried out by independently removing the Bi-LSTM and Multi-head components. Extensive experiments were conducted with healthy and broken gear vibration signal data obtained from different load levels. The model classification performance metrics, complexity matrix, ROC curve, and feature space visualizations were also supported. The experimental results obtained have shown that the proposed model, through ablation studies, reveals the effectiveness of each module, demonstrating that our model learns more distinctive features and has a strong generalization ability.

The main contributions of this approach are listed below:

1. Signal representations are automatically extracted from raw vibration signals without the need for manual feature extraction.
2. Better representations of gear failures are provided by paying attention to the relationships between long-time representations, distinctive information at the critical time step, and temporal local patterns.
3. The individual contribution of the proposed temporal modeling and attention components is systematically investigated through controlled component-wise ablation experiments, providing direct evidence of their complementary roles.
4. It has shown to be an effective solution for real-time industrial predictive maintenance applications by offering high accuracy, strong generalization ability, and low computational cost under different load conditions.

2. Literature Review

Aurangabadkar et al. (2023) performed gearbox fault diagnosis using different machine learning approaches such as Decision Tree Classifier, Random Forest Classifier, and Stochastic

Gradient Boosting. Noise was removed from the signals in the dataset, and they achieved the highest success with Stochastic Gradient Boosting, reaching a test accuracy of 93.9%. Sohaib et al. (2022) obtained discriminative representations from raw vibration signals using multi-task learning by acquiring data under variable load and speed conditions for gearbox fault diagnosis. Thanks to multi-task learning, it was possible to learn diverse types of faults simultaneously. Test accuracy of 99.93% was achieved. Cheng et al. (2024) performed four different types of gearbox fault classification using a hybrid architecture of transfer learning and channel attention mechanism. With EfficientNetV2 and channel attention, a test accuracy rate of 99.75% was achieved. The effectiveness of the proposed architecture in the study was evaluated with various classification performance metrics. According to the results obtained from the experiments, it was determined that the model has generalization ability. Xu et al. (2022) used a rotational frequency search algorithm to determine the actual rotational speed of the gearbox. This algorithm was shown to be a robust approach with noise and variable operating conditions during the testing phase.

Lu et al. (2025) derived ten different gearbox faults using Continuous Wavelet Transform time-frequency. They achieved a 99.4% success rate by running the vision transformer architecture under different operating conditions. When they tested the model under noisy conditions, they found that the model was robust to noise, achieving 93.0% even at 5dB of noise. Sekko et al. (2023) proposed a noise-tolerant method for early-stage gearbox fault diagnosis. Unlike other methods, they modeled faulty vibration signals without adding noise. They proposed a least squares-based deconvolution method. Bhadane et al. (2014) proposed an automated gearbox fault diagnosis approach based on Motor Current Signature Analysis. They also established a hybrid structure using Discrete Wavelet Transform (DWT) and Fast Fourier Transform (FFT) in the model. Kumar et al. (2023) applied preprocessing to raw vibration data with Calculus Enhanced Energy Operator with three different fault classes in gearbox fault diagnosis. After extracting features from the signals, fault types were classified with Multi-Class Support Vector Machine. Their proposed solution achieved 96.90% test accuracy. They stated that this model, with its non-parametric structure, can be used in industrial solutions.

Wang et al. (2025) proposed using transmission fault signals instead of vibration signals in gearbox fault diagnosis. These transmission fault signals are more stable and more resistant to noise. They established a 6-degree-of-freedom dynamic gearbox model. A lightweight 1D Convolutional Neural Network (1D-CNN) was used to classify three different gearbox failures as cracked, worn, and healthy. In the SNR test, 40.3% accuracy was achieved at 0dB with the classical vibration signal, while this rate increased to 99.25% with transmission failure signals. Han et al. (2022) combined the advantages of CNN and transformer for the diagnosis of gearbox failure under high noise, completing the learning by feeding local features and long-range dependencies with the channel attention mechanism. In the architecture they proposed, eight different gearbox failures were classified, achieving accuracy between 84.67–98.21% at different SNR levels.

Saufi et al. (2020) classified three different gearbox faults using a Stacked Spare Autoencoder by converting vibration signals into time-frequency images using a small amount of data for gearbox fault diagnosis. Their proposed model achieved 99% accuracy. Li et al. (2016) proposed an innovative method based on Deep Random Forest Fusion multi-sensor fusion using acoustic emission and vibration signals together for eleven different gearbox faults. For comparison purposes, the proposed model was tested with k-neighbor nearest (K-NN), SVM, and other scenarios, and the hybrid combination of these modules showed the highest success rate at 97.68%. Chen et al. (2015) identified the model with the highest success rate over

twenty different scenarios by testing twelve different gearbox faults with classical CNN using different hyperparameters and number of layers. They achieved 98.2% test accuracy.

Zheng et al. (2002) Liu et al. proposed a method for gearbox fault diagnosis using Morlet CWT by improving the Continuous Wavelet Transform (CWT). Liu et al. (2006) proposed an innovative solution based on Empirical Mode Decomposition and Hilbert Spectrum for early-stage external crack faults of the gearbox. Lin and Zuo (2003) comparatively tested DB4-based Decimated DWT and Harmonic Wavelet-based Undecimated DWT methods for detecting early gearbox tooth cracks in noisy vibration signals. By developing an Adaptive Morlet Wavelet Filter and separating pulse signals from noisy signal data, they concluded that Harmonic DWT showed more successful results. Fan et al. (2025) established a hybrid model of improved residual connections, transformer, and CNN for gearbox fault classification. They achieved an accuracy rate of 96.21% with noise-free vibration data. Inturi et al. (2021) proposed a multi-level fault diagnosis approach to simultaneously determine the type, location, and severity of gearbox faults. Fault classification was performed using Decision Tree by separating vibration signals with Wavelet Transform. Inference was then made from the combined features using Adaptive Neural Fuzzy Inference System. This resulted in an accuracy rate of 92%.

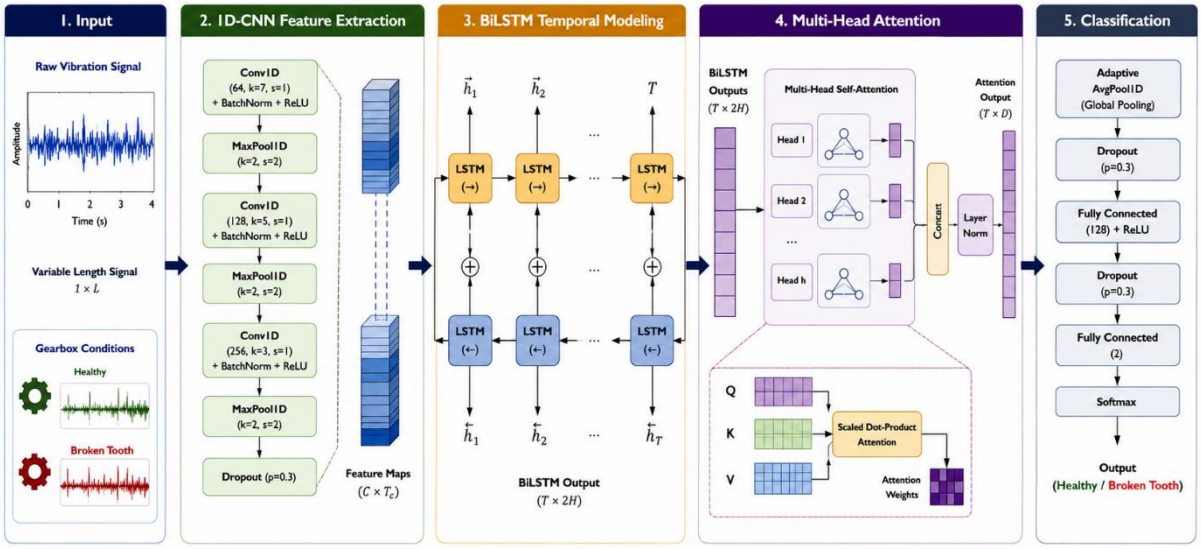
Refs.	Method	Dataset	Performance
Aurangabadkar et al., 2023	Decision Tree, Random Forest, Stochastic Gradient Boosting	Gearbox vibration signals	93.9% Accuracy
Sohaib et al., 2022	Multi-task Deep Learning on raw vibration signals	Variable load & variable speed gearbox dataset	99.93% Accuracy
Lu et al., 2025	Continuous Wavelet Transform (CWT) + Vision Transformer	Ten gearbox fault dataset under different operating conditions	99.4% Accuracy, 93.0% at 5 dB SNR
Kumar et al., 2023	Calculus Enhanced Energy Operator + Multi-class SVM	Three-class gearbox vibration dataset	96.90% Accuracy
Wang et al., 2025	Transmission Fault Signals + Lightweight 1D-CNN	Three-class gearbox dataset (Crack, Wear, Healthy)	99.25% at 0 dB SNR, compared with 40.3% using vibration signals
Han et al., 2022	CNN + Transformer + Channel Attention	Eight-class gearbox fault dataset	84.67–98.21% Accuracy depending on SNR
Saufi et al., 2020	Continuous Wavelet Transform + Stacked Sparse Autoencoder	Three-class gearbox dataset	99.0% Accuracy
Li et al., 2016	Deep Random Forest Fusion + Multi-sensor Fusion (AE + Vibration)	Eleven-class gearbox fault dataset	97.68% Accuracy
Fan et al., 2025	Improved Residual CNN + Transformer	Noise-free gearbox vibration dataset	96.21% Accuracy
Inturi et al., 2021	Wavelet Transform + Decision Tree + Adaptive Neuro-Fuzzy Inference System (ANFIS)	Multi-level gearbox fault dataset	92.0% Accuracy

Table 1: Studies in literature on gearbox defect diagnosis

3. Materials and Method

This study proposes hybrid architecture for end-to-end classification of gearbox faults by automatically extracting features from raw vibration signals. The hybrid structure consists of 1D-CNN, multi-head self-attention, and Bi-LSTM modules. This allows for the simultaneous

learning and modeling of both local and long-term dependencies. In gearbox fault diagnosis, classical methods require expert knowledge in addition to manually extracting frequency-domain, time-domain, and time-frequency domain-based features. Furthermore, classical diagnostic methods have limited generalization capabilities to different operating conditions. The proposed method overcomes these limitations by directly performing an automatic feature extraction process from raw vibration signals and eliminating the need for manual feature engineering. The first step of the proposed architecture involves extracting vibration amplitude variations and short-range local variations from gearbox signals at multiple levels using 1D-CNN blocks with different kernel sizes. Through these blocks, high-level representations are transferred to the Bi-LSTM module. Thanks to the Bi-LSTM layer, it comprehensively evaluates the time-dependent evolution of gearbox signals by learning patterns from past temporal information as well as evaluating future patterns. In the final stage, a multi-headed attention mechanism focuses on the valuable time steps in different feature representations, contributing to the model's learning of distinguishing features. By evaluating both local vibration characteristics and long-term temporal relationships together, the classification of healthy and fractured teeth is ensured with high accuracy. The comprehensive flow of the proposed architecture is shown in Figure 1.



Figure

Figure 1: The overview of the proposed model for gearbox defect diagnosis

3.1. CNN-based Local Feature Extraction

The first stage of the model, the Gearbox, uses 1D-CNN to extract distinctive features from raw vibration signals. Since the Gearbox vibration signal is a time-dependent one-dimensional array, the computational cost is low. The raw vibration signal is transferred to the 1D-CNN model without the need for any transformation. The architecture includes three consecutive convolution blocks. The proposed three consecutive convolution blocks offer a balance between computational cost and network depth. In the first convolution, the model learns the low-frequency components in the raw vibration signal; in the second convolution, mid-level features are obtained from the signals; and in the final convolution layer, the model learns the finer details at higher frequencies. This allows for the generation of abstract and distinctive features.

The kernel sizes in the three convolutional blocks in the input raw vibration signal (1) are 7, 5, and 3. In this convolution with a large core filter, the general characteristics of the raw signal

were studied with a large area (2) and (3). Learning medium-scale vibration patterns with filter size 5 is done in (4) and (5) and learning local fault characteristics with the final filter size 3 is done in (6) and (7).

$$X = [x_1, x_2, \dots, x_N] \tag{1}$$

$$F_1 = \mathrm{ReLU}\left(\mathrm{BN}(w_1 * \mathrm{X} + b_1)\right) \tag{2}$$

$$P_1 = \mathrm{MaxPool}\left(F_1\right) \tag{3}$$

$$F_2 = \mathrm{ReLU}\left(\mathrm{BN}(w_2 * P_1 + b_2)\right) \tag{4}$$

$$P_2 = \mathrm{MaxPool}\left(F_2\right) \tag{5}$$

$$F_3 = \mathrm{ReLU}\left(\mathrm{BN}(w_3 * P_2 + b_3)\right) \tag{6}$$

$$P_3 = \mathrm{MaxPool}\left(F_3\right) \tag{7}$$

$$F_{CNN} = f_{CNN}(\mathrm{X}) \tag{8}$$

The CNN feature map (8) obtained at the end of the three convolution blocks is also obtained and this high-level feature representation is transferred to Bi-LSTM.

3.2. Bidirectional LSTM for Temporal Dependency Learning

To learn long-term temporal relationships, the representations obtained from the CNN layers are transferred to Bi-LSTM. CNNs are limited by their inability to fully model the features at distant time steps, which also restricts the learning of fault characteristics. With the Bi-LSTM structure, the input feature representation is processed both forward and backward, producing two different representations for each time step. The classification success of the model is also increased by using inference information from both past and future information. At each time step, the LSTM cell regulates the information flow through the input gate (9), forget gate (10), and output gate (13). The candidate cell state (11) is obtained with the input gate and forget gate, and the new cell state in (12) is obtained by adding the added information to the cell state.

$$f_t = \sigma\left(\left(w_f \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} + b_f\right)\right) \tag{9}$$

$$i_t = \sigma\left(\left(w_i \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} + b_i\right)\right) \tag{10}$$

$$\hat{c}_t = tanh\left(\left(w_c \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} + b_c\right)\right) \tag{11}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \tag{12}$$

$$o_t = \sigma\left(\left(w_o \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} + b_o\right)\right) \tag{13}$$

Forward LSTM carries information from the past to the future (14) and backward LSTM processes the sequence in the reverse direction (15) learning contextual information from future instances. Then, by combining the outputs from both directions (16), a richer representation is created for each time step.

$$\vec{h}_t = LSTM_f(x_t, \vec{h}_{t-1}) \quad (14)$$

$$\overleftarrow{h}_t = LSTM_f(x_t, \overleftarrow{h}_{t-1}) \quad (15)$$

$$H_t = [\vec{h}_t \ ; \ \overleftarrow{h}_t] \quad (16)$$

3.3. Multi-Head Self-Attention

A multi-head self-attenuation mechanism was used to ensure that all time steps in the incoming feature map of the proposed model contribute equally to the classification process. In Gearbox, fractured tooth failures appear as short, impulsive pulses at specific time intervals, rather than characteristic and continuously manifesting changes.

Thanks to self-attenuation, it is possible to learn the relationship of each time step with all other time steps. Since a single-head self-attenuation can analyze only single subspace, better representation ability than vibration signals was obtained in this study with a multi-head self-attenuation mechanism.

4. Experimental Results

4.1. Dataset Description

In this study, the public gearbox fault dataset (Brjapon, 2026) was used. This dataset consists of a total of 20 vibration signals, 10 healthy and 10 faulty, from a setup with four vibration sensors. A total of 8,084,476 raw samples were obtained from all vibration signals, each signal consisting of segments of 2048 sample length and 1024 sample steps, however, a rule was found that there should be a maximum of 40 segments in all files.

File-level group-based data splitting was applied from 800 balanced samples with three-channel tensor three-dimensional input, each with dimensions of 128 × 128. The characteristics of the dataset used and the preprocessing steps applied are as shown in Table 2. Examining Figure 2, the faulty vibration signal picks at different time steps and differs significantly from the healthy signal.

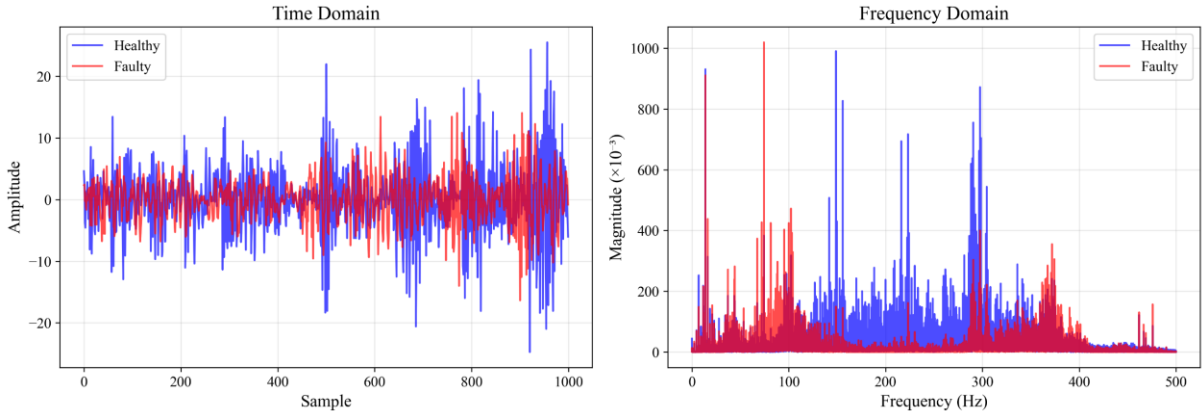


Figure 2: The healthy and broken tooth fault of gearbox.

Category	Description
Signal files	20 (10 Healthy, 10 Broken Tooth)
Sensors	4
Load range	0–90%
Raw samples	8,084,476
Segment length	2,048
Stride	1,024
Segments/file	≤40
Total segments	800 (400/400)
Representation	Multi-resolution
Input size	3 × 128 × 128
Classes	2

Table 2: Dataset characteristics and preprocessing configuration.

4.2. Experimental Setup

In the proposed architecture, feature extraction is applied to the raw vibration signals without manual feature extraction, and features are extracted from the deep learning-based model. The performance obtained in identifying healthy and faulty signals produced by the model has been measured with various metrics. Since input data in deep learning architectures must be of fixed length, raw vibration signals are separated into smaller signals using a sliding window approach. The window length is 512 units and the window sliding amount is assigned as 256 units. Thus, overlap is achieved between successive windows, and faulty features are transferred to different time steps. To test the model's robustness against different conditions through data augmentation and noise addition, random amplitude scaling, time-shifting, limited time compression/expansion, frequency masking, and low-probability pulse noise were added to the signal data. To prevent data leakage, all signal segments were grouped together and then split into 70/15/15 groups. The experimental adjustments of the model are shown in Table 2.

Category	Parameter	Value
Data	Input type	Vibration
	Segment length	512
	Step length	256
Train	Train/ Validation / Test	70 / 15 / 15
	Division method	Group-based Split
	Batch Size	32
	Epoch	25
	Optimizer	Adam
	Loss	Cross Entropy
Hardware	GPU	NVIDIA RTX 4070 Laptop GPU
	Framework	Pytorch

Table 3: Experimental settings of the proposed model.

4.3. K-Fold Cross-Validation Results

A five-fold cross-validation (CV) was applied to evaluate the generalization ability of the proposed model. In each fold, the dataset was divided into training and test subsets, the model was retrained in each iteration and evaluated on the test set. This approach reduces

random performance variations that might arise from a single training-test split, thus more accurately revealing the true generalization success of the model. Figure 3 shows the changes in accuracy rates over the train- validation, loss curve, and f1-score curves. Table 4 provides detailed results of model performance.

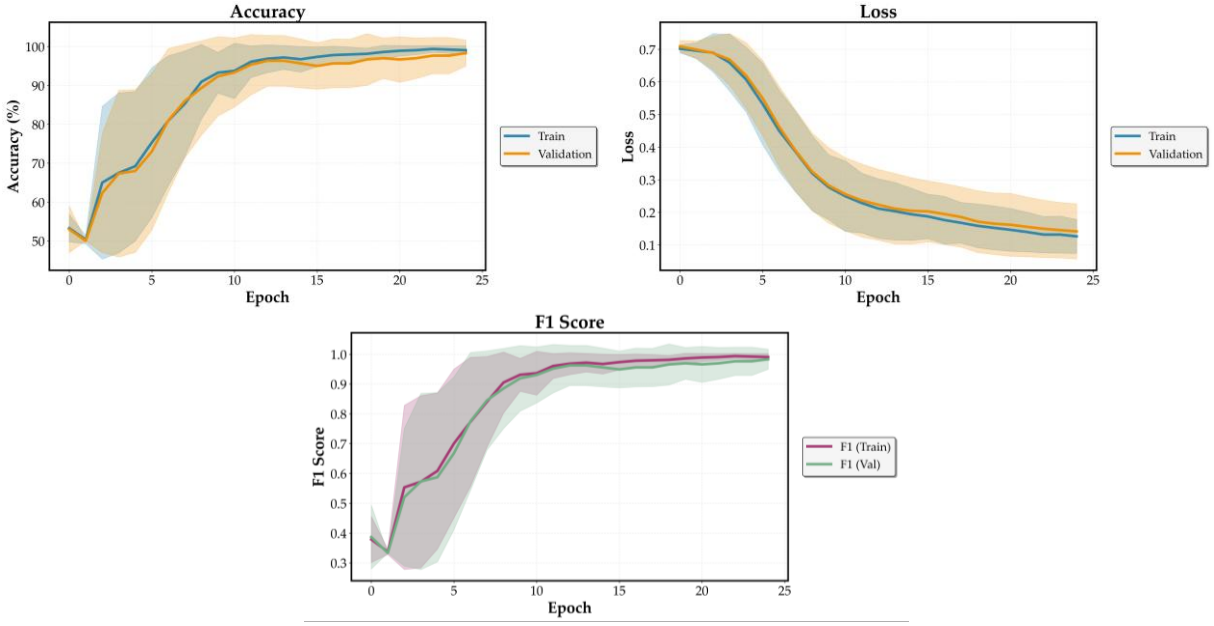


Figure 3: Performance curves of the proposed model using 5-fold CV.

Metric	Train	Validation	Test
Accuracy	0.9906 ± 0.0128	0.9833 ± 0.0373	0.9800 ± 0.0328
Precision	0.9910 ± 0.0124	0.9857 ± 0.0319	0.9822 ± 0.0282
Recall	0.9907 ± 0.0128	0.9833 ± 0.0373	0.9800 ± 0.0328
F1-Score	0.9906 ± 0.0128	0.9832 ± 0.0375	0.9799 ± 0.0330

Table 4: Performance results of the proposed model using 5-fold CV.

4.4. Experimental Evaluation on the SEU Gearbox Dataset

To demonstrate the generalization capability of the proposed model under different fault conditions, a publicly available gearbox dataset, namely the SEU dataset (Shao et al., 2018), was employed. Five distinct gearbox conditions were investigated, including Chipped, Health, Miss, Root, and Surface. The dataset was acquired with a frequency limit of 2000 Hz, yielding 1600 spectral lines and 1024 blocks per recording. Each recording comprises a total of 4,194,304 data samples. To ensure a fair and robust comparison across the five gearbox conditions, a stratified 5-fold CV strategy was employed. In each fold, the model was trained and evaluated using different subsets of the dataset, while the final performance was obtained by aggregating the results across all five folds. Detailed results are presented in Table 5 and Figure 4. The findings demonstrate that the proposed model achieves consistently high performance across the five folds, supporting its generalization capability under different gearbox fault conditions. The confusion matrix and ROC curve results are illustrated in Figure 5 and Figure 6, respectively. These results further demonstrate the performance of the proposed model on both the two-class and multi-class gearbox datasets.

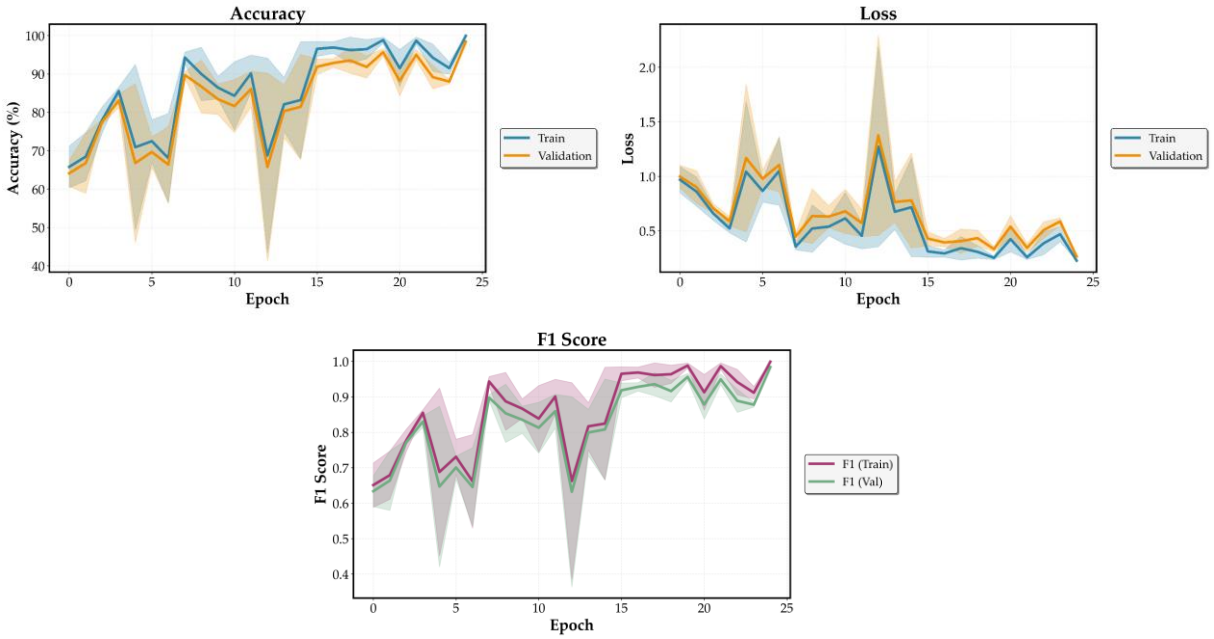


Figure 4: Comparative classification performance of proposed models on the SEU dataset.

Metric	Train	Validation	Test
Accuracy	0.9992±0.0012	0.9840±0.0042	0.9825±0.0049
Precision	0.9992±0.0012	0.9841±0.0043	0.9827±0.0047
Recall	0.9992±0.0012	0.9840±0.0042	0.9825±0.0049
F1-Score	0.9991±0.0012	0.9840±0.0043	0.9825±0.0049

Table 5: Performance results of proposed model on SEU dataset using 5-fold CV.

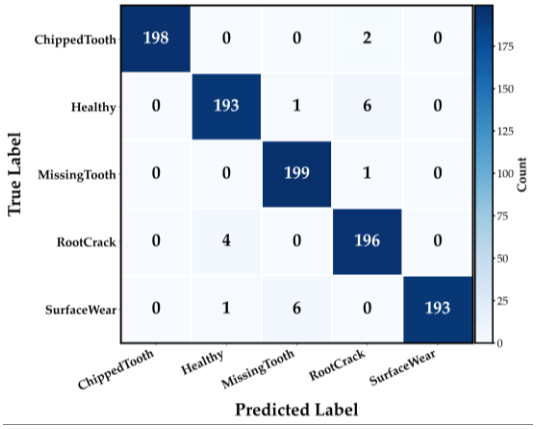


Figure 5: Confusion matrix of proposed models on the SEU dataset.

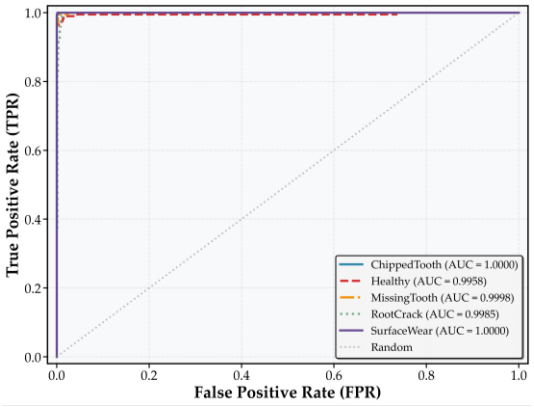
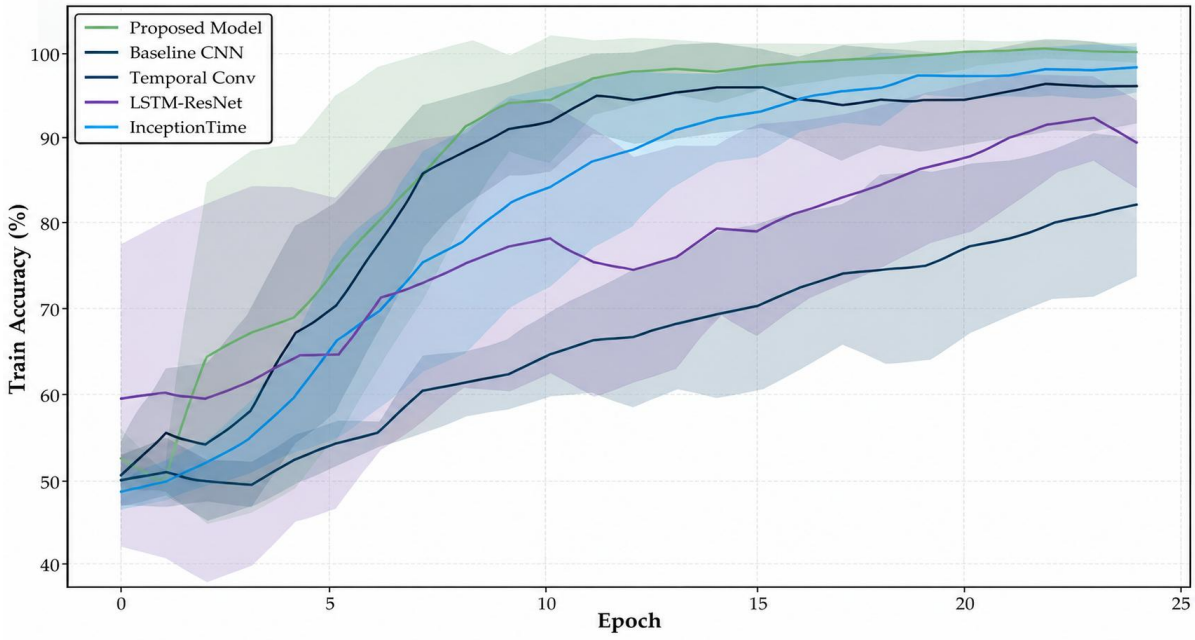


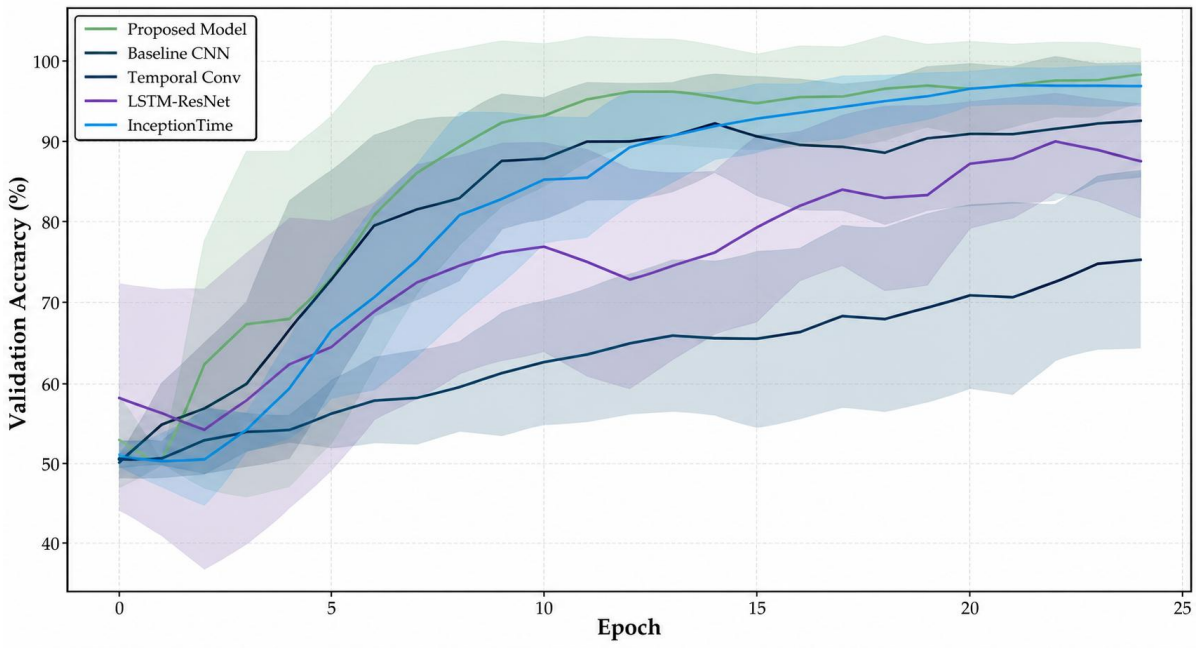
Figure 6: ROC curve of proposed models on the SEU dataset.

4.5. Comparison with Existing Deep Learning Models

To prove the effectiveness of the proposed model on the gearbox dataset, it was compared with various deep learning models. Baseline CNN, Temporal Convolutional Attention (TCA), ResNet-LSTM, and InceptionTime 1D models were used for the comparison. To ensure a fair comparison, all models were trained with the same dataset and hyperparameters. The results are presented comparatively in Table 6 and Figure 7, showing accuracy, precision recall, and f1 score comparisons for training validation and testing. Figure 8 shows the comparative confusion matrix results of our model against these models, Figure 9 shows the ROC curve comparisons, and Figure 10 presents the models' ability to distinguish classes from each other using t-SNE.



a) Training accuracy

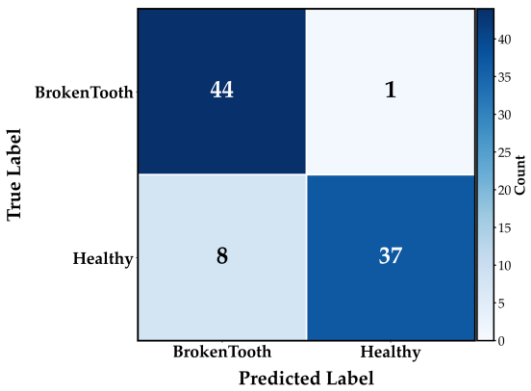


b) Validation accuracy

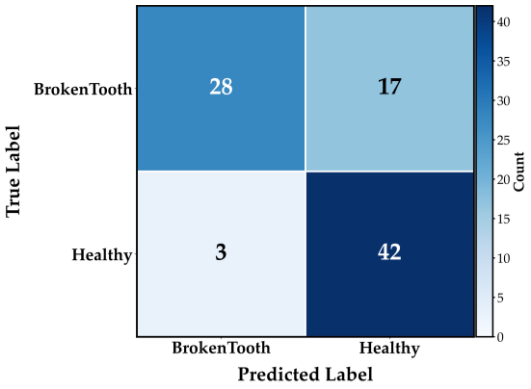
Figure 7: Comparative classification performance of comparison models.

Metric	Baseline CNN	Temporal Conv	LSTM-ResNet	InceptionTime	Proposed Model
Train Accuracy	0.9531±0.0465	0.8188±0.0869	0.8891±0.0534	0.9734±0.0295	0.9906±0.0128
Validation Accuracy	0.9267±0.0805	0.7533±0.1227	0.8767±0.0796	0.9700±0.0274	0.9833±0.0373
Test Accuracy	0.9444±0.0689	0.8000±0.0766	0.9089±0.0751	0.9533±0.0355	0.9800±0.0328
Train Precision	0.9598±0.0381	0.8370±0.0869	0.8974±0.0464	0.9752±0.0279	0.9910±0.0124
Validation Precision	0.9426±0.0589	0.7729±0.1219	0.8870±0.0800	0.9727±0.0241	0.9857±0.0319
Test Precision	0.9553±0.0530	0.8216±0.0812	0.9154±0.0726	0.9564±0.0341	0.9822±0.0282
Train Recall	0.9530±0.0464	0.8178±0.0874	0.8869±0.0547	0.9731±0.0294	0.9907±0.0128
Validation Recall	0.9267±0.0805	0.7533±0.1227	0.8767±0.0796	0.9700±0.0274	0.9833±0.0373
Test Recall	0.9444±0.0689	0.8000±0.0766	0.9089±0.0751	0.9533±0.0355	0.9800±0.0328
Train F1-Score	0.9527±0.0471	0.8156±0.0878	0.8874±0.0550	0.9734±0.0296	0.9906±0.0128
Validation F1-Score	0.9249±0.0831	0.7477±0.1258	0.8758±0.0798	0.9699±0.0275	0.9832±0.0375
Test F1-Score	0.9435±0.0705	0.7968±0.0766	0.9084±0.0756	0.9532±0.0356	0.9799±0.0330

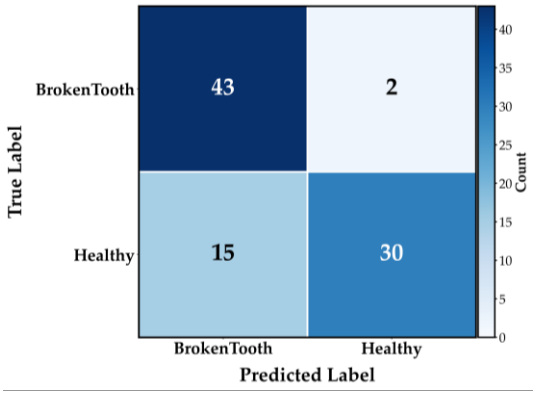
Table 6: Comparison of classification performance comparison models.



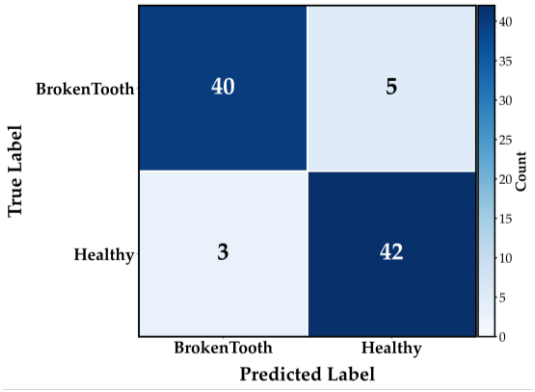
a) Baseline CNN



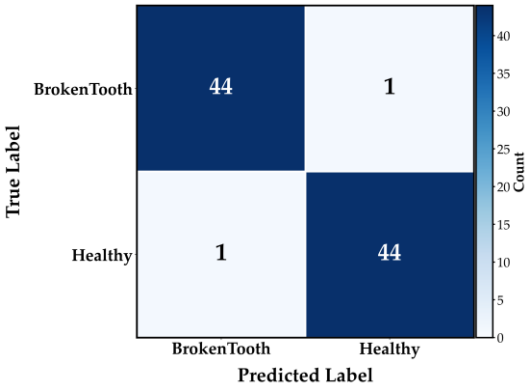
b) Temporal Conv



c) LSTM-ResNet

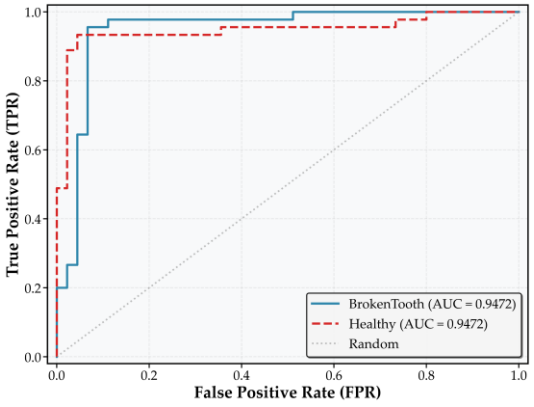


d) InceptionTime1D

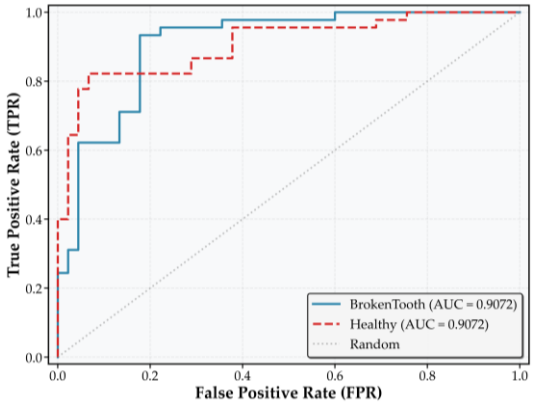


e) Proposed model

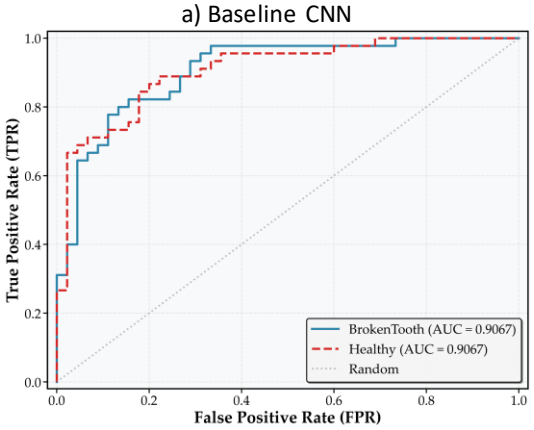
Figure 8: Confusion matrices of comparison benchmark models.



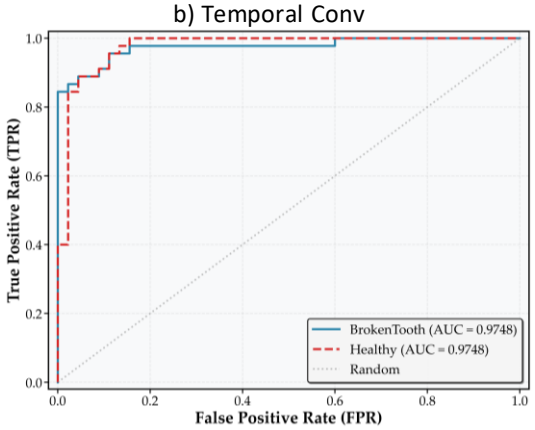
a) Baseline CNN



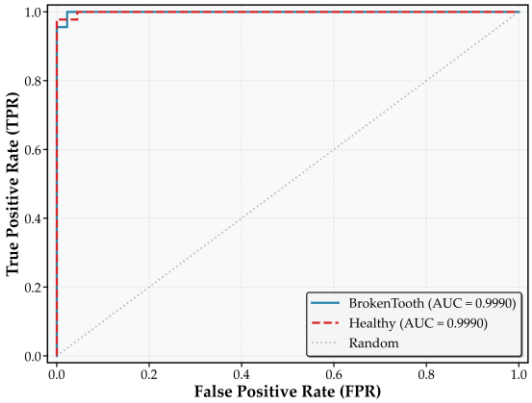
b) Temporal Conv



c) LSTM-ResNet

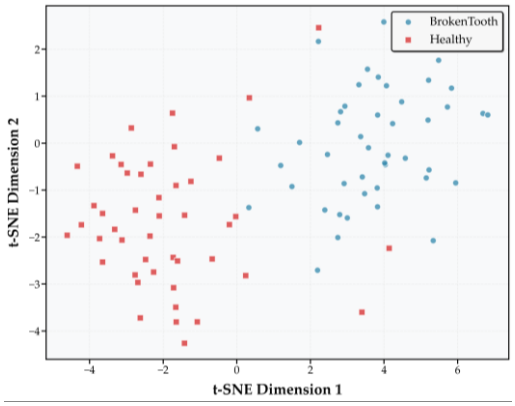


d) InceptionTime1D

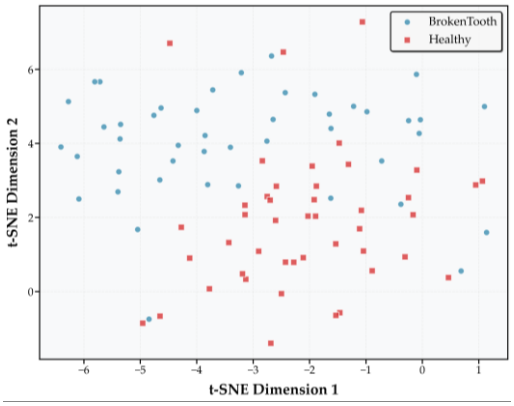


e) Proposed model

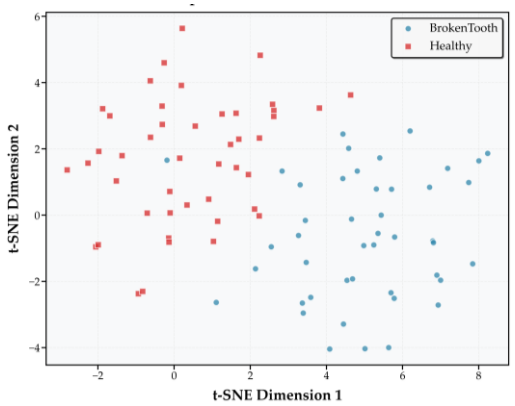
Figure 9: Comparative ROC curves of comparison models.



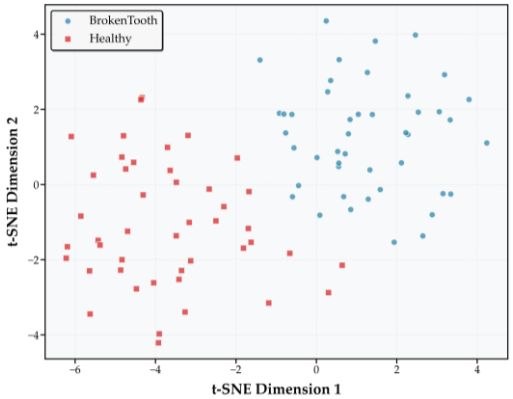
a) Baseline CNN



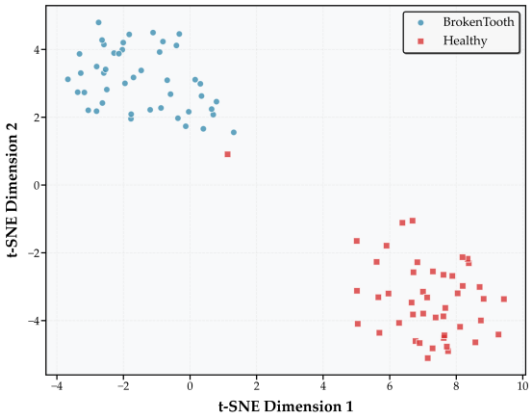
b) Temporal Conv



c) LSTM-ResNet



d) InceptionTime1D



e) Proposed model

Figure 10: Comparative t-SNE visualization of the comparison models.

As illustrated in Figure 8, the confusion matrices reveal that the Baseline CNN, Temporal Convolutional Attention, LSTM-ResNet, InceptionTime 1D, and the proposed model produced 9, 20, 17, 8 and 2 misclassified samples, respectively. Among all evaluated models, the proposed hybrid architecture achieved the lowest number of misclassifications, indicating its superior classification capability and enhanced robustness in distinguishing healthy and broken gear conditions.

A similar trend can be observed in the ROC-AUC analysis presented in Figure 9. In the Figure 10, the t-SNE visualizations illustrate the superiority of the proposed model in feature representation learning. Compared with the competing approaches, the proposed model produces more compact intra-class feature distributions while simultaneously increasing the separation between the healthy and faulty classes.

Model	Params (M)	FLOPs (G)	Inference (ms)
Baseline CNN	0.006	0.009	1.19
Temporal Conv	0.007	0.011	1.27
LSTM-ResNet	0.007	0.011	1.30
Proposed model	0.010	0.011	1.35

Table 7: Computational characteristics of the evaluated models.

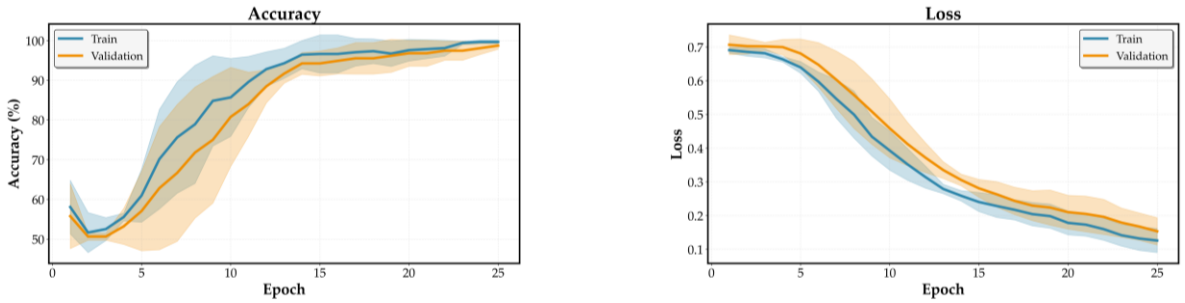
Table 7 shows that the proposed model maintains a comparable computational cost to the benchmark models while achieving improved classification performance. With only 0.010 M parameters, 0.011 G FLOPs, and 1.35 ms inference time, the proposed model introduces only a limited computational overhead. These results indicate that the performance gains are achieved without a substantial increase in computational complexity.

4.6. Reproducibility and Stability Analysis

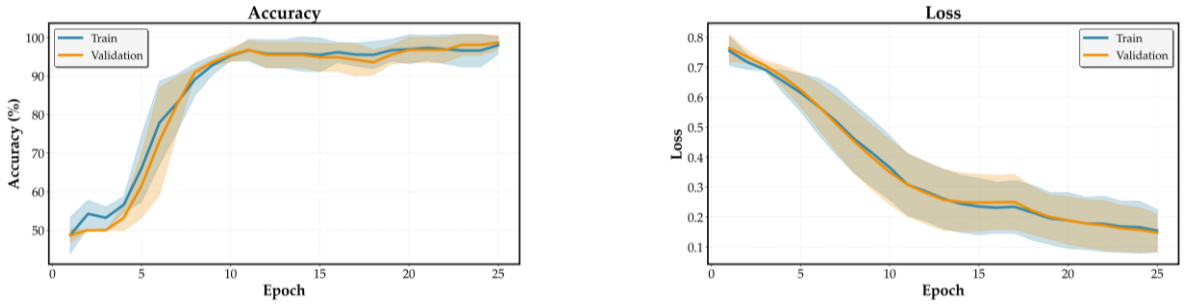
To evaluate the reproducibility and stability of the proposed model, it was retrained by restarting the process with three randomly generated seeds. To obtain fair comparison results, the comparison of the three different seeds was initiated using the same preprocessing optimization and training parameters. The results are presented comparatively in Table 8. Accuracy and loss curves using 5-fold CV are presented for each seed setting in Figure 11 and Figure 12.

Metric	Seed-1	Seed-2	Seed-3
Train Accuracy	0.9965±0.0060	0.9800±0.0281	0.9722±0.0318
Validation Accuracy	0.9872±0.0111	0.9872±0.0222	0.9936±0.0111
Test Accuracy	0.9927±0.0127	0.9634±0.0317	0.9707±0.0336
Train Precision	0.9969±0.0053	0.9801±0.0281	0.9745±0.0264
Validation Precision	0.9877±0.0107	0.9881±0.0206	0.9938±0.0107
Test Precision	0.9919±0.0141	0.9664±0.0298	0.9716±0.0344
Train Recall	0.9961±0.0067	0.9816±0.0253	0.9718±0.0345
Validation Recall	0.9872±0.0111	0.9872±0.0222	0.9936±0.0111
Test Recall	0.9936±0.0111	0.9626±0.0334	0.9722±0.0289
Train F1-Score	0.9965±0.0061	0.9800±0.0282	0.9717±0.0323
Validation F1-Score	0.9872±0.0111	0.9872±0.0222	0.9936±0.0111
Test F1-Score	0.9926±0.0129	0.9625±0.0325	0.9704±0.0337

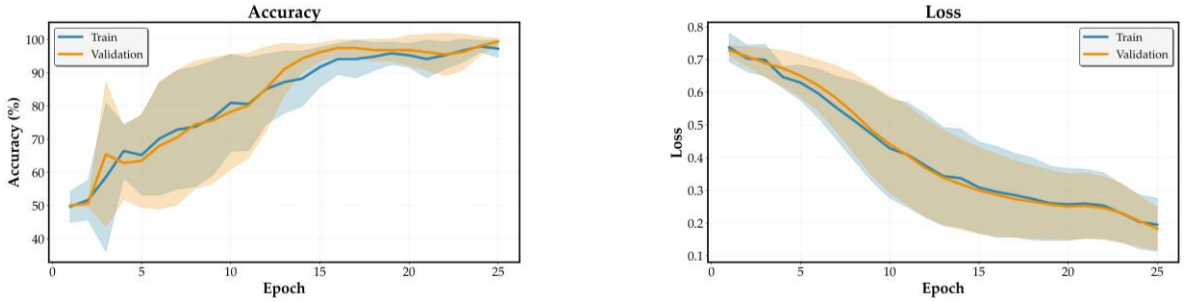
Table 8: Performance stability of the proposed model across different random seeds.



(a) Seed- 1

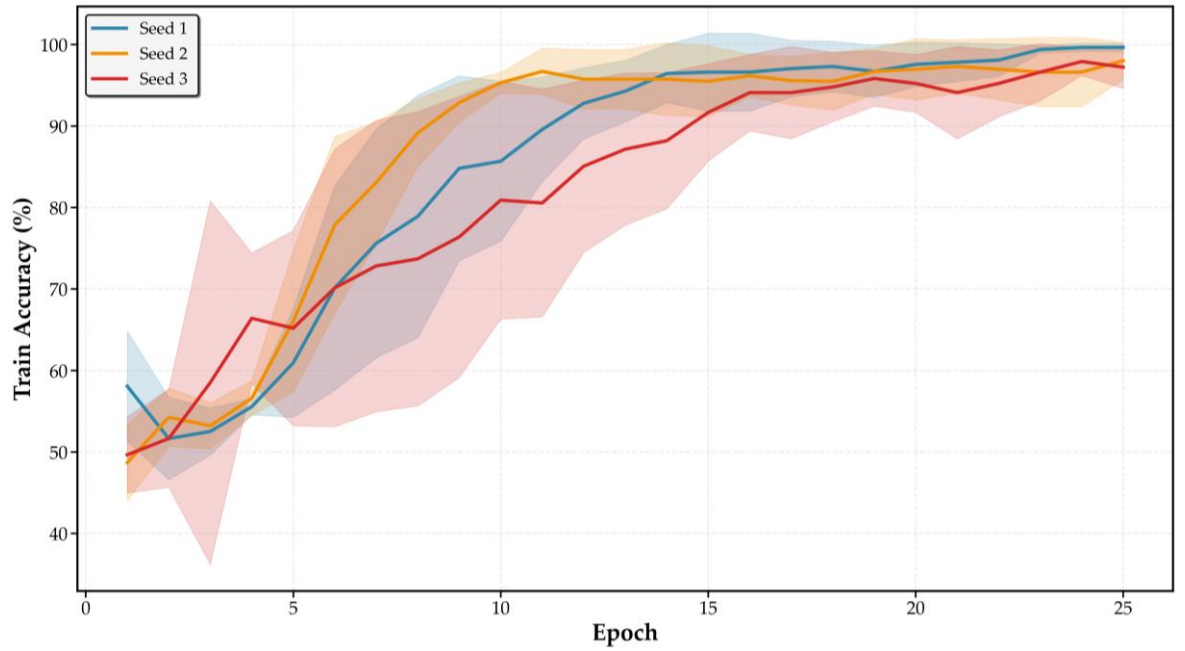


(b) Seed- 2

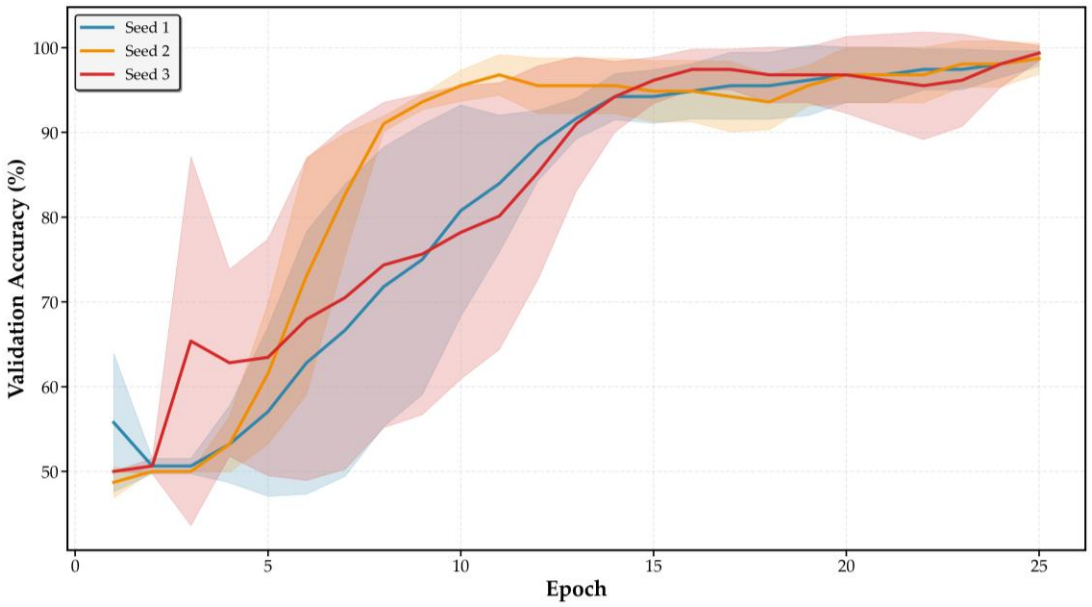


(c) Seed- 3

Figure 11: 5-fold CV performance across different random seeds.



(a) Training



(b) Validation

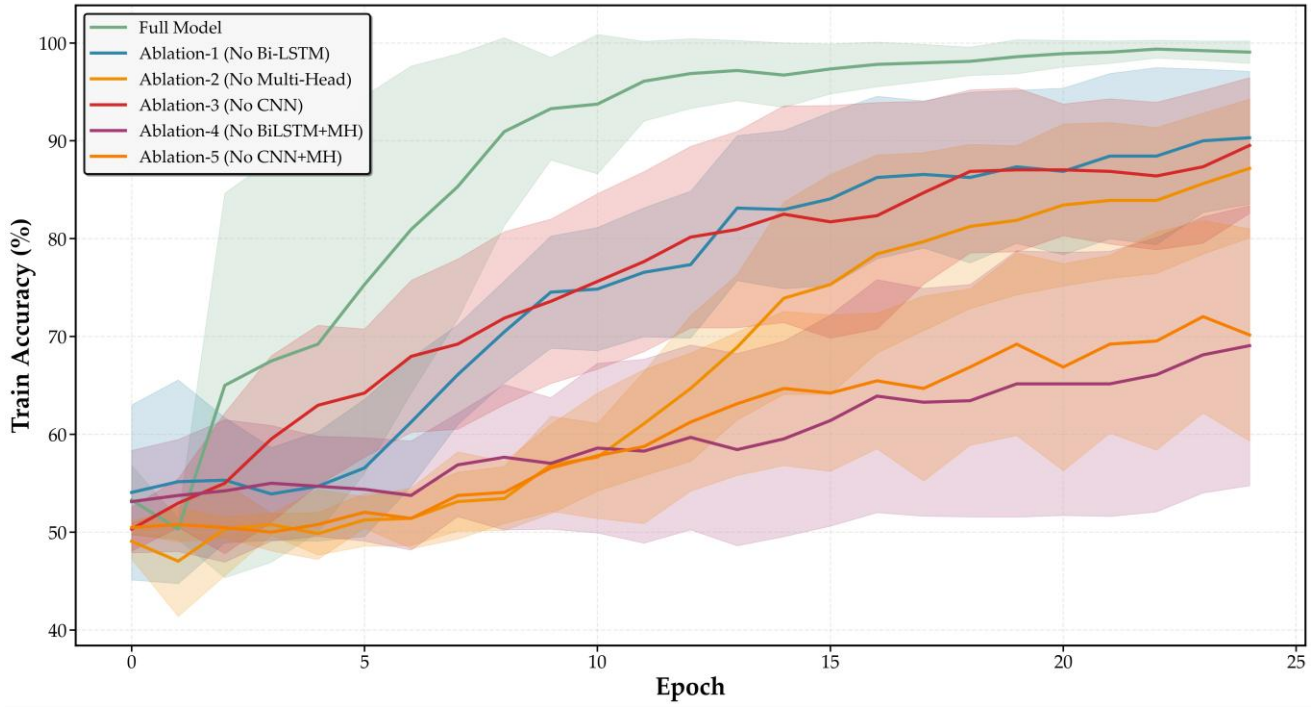
Figure 12: (a) Training and (b) Validation accuracy comparison across different random seeds using 5-fold CV.

4.7. Ablation Studies

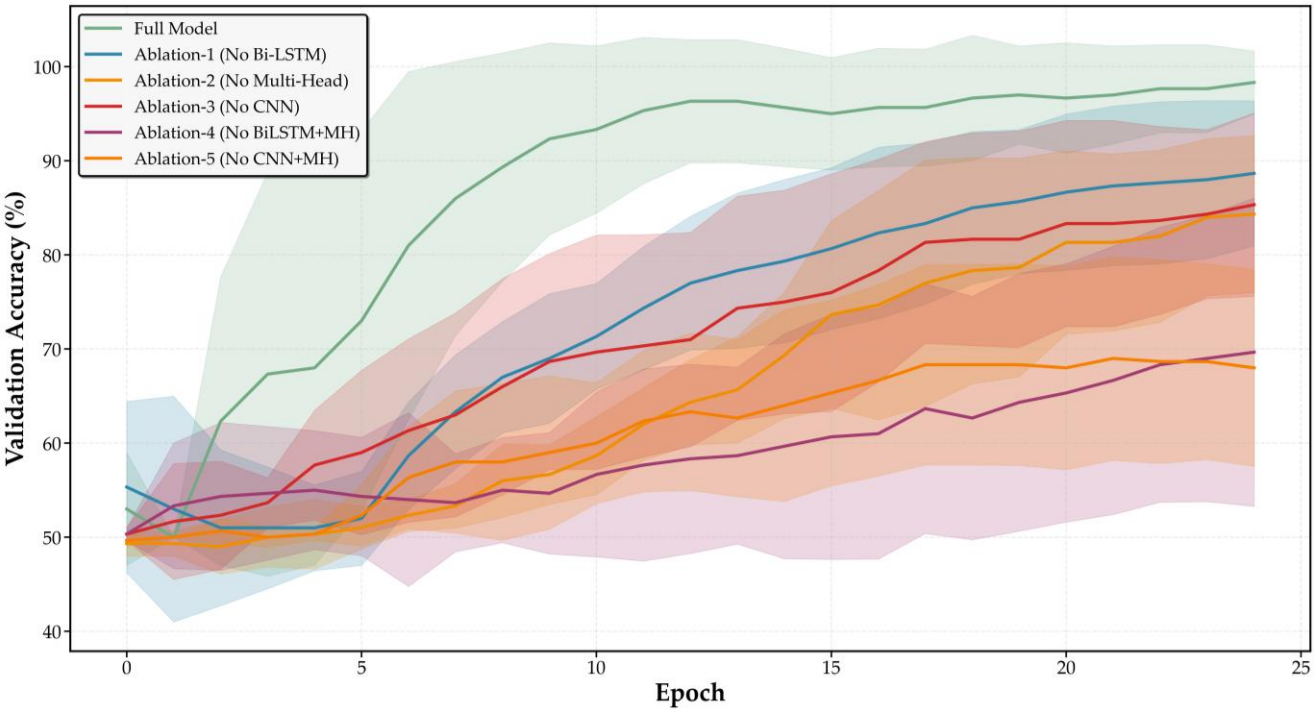
In this section, an ablation study was conducted to examine the contribution of the components to the proposed model. The model was run by successively shutting down each component in the model, and the resulting classification metrics, confusion matrix, ROC curve, and t-SNE outputs were compared. Table 9 and Figure 13 present a comparative analysis of the ablation results in terms of classification metrics. Figure 14 shows the confusion matrix, Figure 15 shows the ROC curves, and Figure 16 shows the t-SNE outputs comparatively.

Metric	Ablation-1 No Bi-LSTM	Ablation-2 No Multi-Head	Ablation-3 No CNN	Ablation-4 No Bi-LSTM+MH	Ablation-5 No CNN+MH	Full Model
Train Acc	0.9031±0.0759	0.8719±0.0792	0.8953±0.0775	0.6906±0.1600	0.7016±0.1212	0.9906±0.0128
Val Acc	0.8867±0.0861	0.8433±0.0932	0.8533±0.1089	0.6967±0.1831	0.6800±0.1169	0.9833±0.0373
Test Acc	0.8822±0.1041	0.8533±0.0944	0.8844±0.0788	0.6333±0.1718	0.7044±0.1151	0.9800±0.0328
Train Prec	0.9239±0.0557	0.8990±0.0543	0.9180±0.0528	0.7617±0.1278	0.6913±0.2496	0.9910±0.0124
Val Prec	0.9060±0.0727	0.8726±0.0787	0.8880±0.0741	0.6655±0.2720	0.6721±0.2473	0.9857±0.0319
Test Prec	0.9077±0.0761	0.8879±0.0557	0.9073±0.0536	0.6123±0.2638	0.6896±0.2464	0.9822±0.0282
Train Rec	0.9033±0.0753	0.8660±0.0848	0.8918±0.0808	0.6859±0.1642	0.7024±0.1173	0.9907±0.0128
Val Rec	0.8867±0.0861	0.8433±0.0932	0.8533±0.1089	0.6967±0.1831	0.6800±0.1169	0.9833±0.0373
Test Rec	0.8822±0.1041	0.8533±0.0944	0.8844±0.0788	0.6333±0.1718	0.7044±0.1151	0.9800±0.0328
Train F1	0.9010±0.0780	0.8657±0.0859	0.8914±0.0825	0.6491±0.2107	0.6598±0.1884	0.9906±0.0128
Val F1	0.8843±0.0887	0.8388±0.0966	0.8463±0.1185	0.6561±0.2357	0.6360±0.1804	0.9832±0.0375
Test F1	0.8780±0.1095	0.8468±0.1036	0.8813±0.0831	0.5882±0.2150	0.6636±0.1854	0.9799±0.0330

Table 9: Five-fold CV results of the component-wise ablation study.

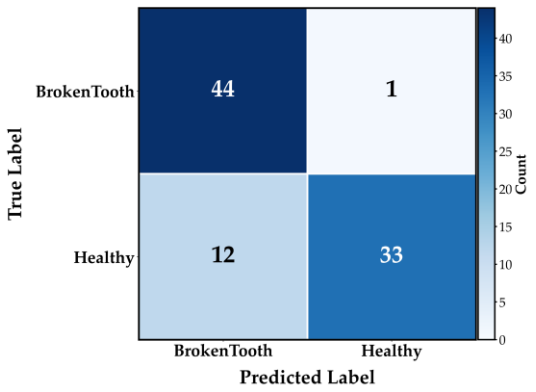


(a) Training accuracy

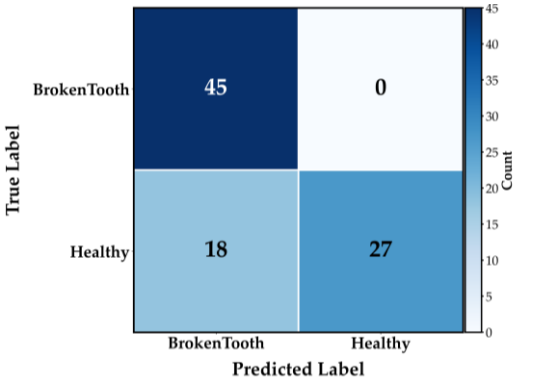


(b) Validation accuracy

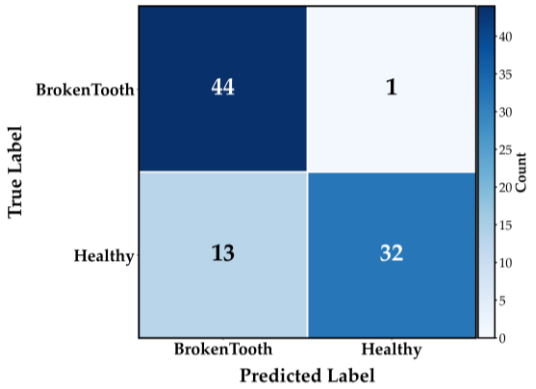
Figure 13: Comparative classification performance of ablation studies.



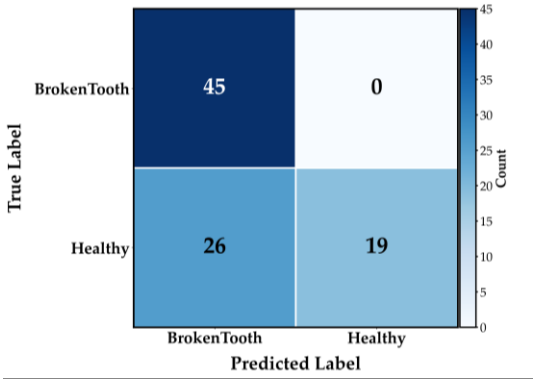
a) Ablation-1



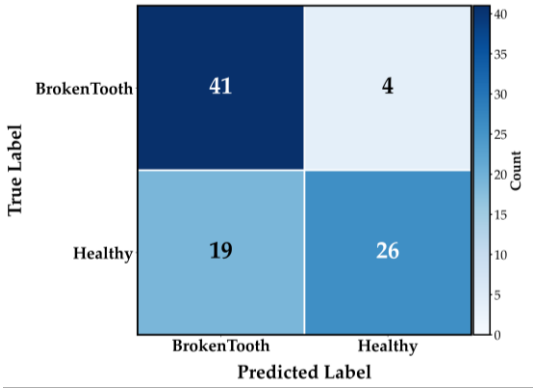
b) Ablation-2



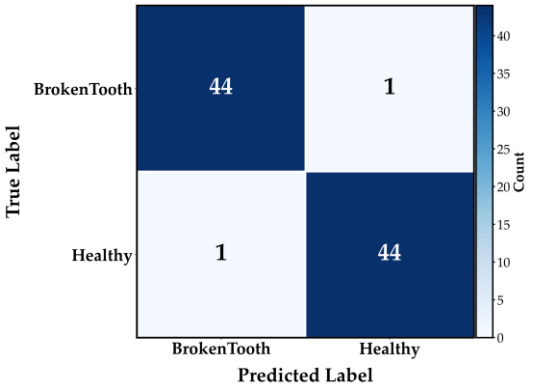
c) Ablation-3



d) Ablation-4

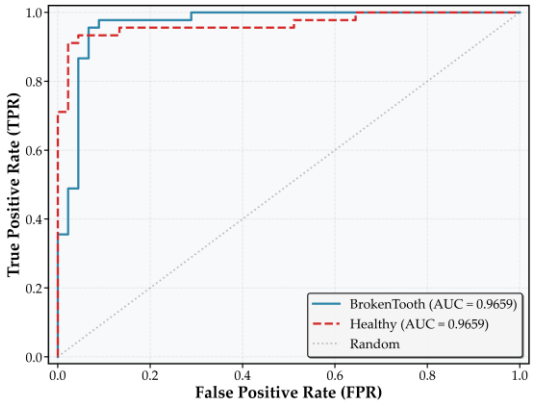


e) Ablation-5

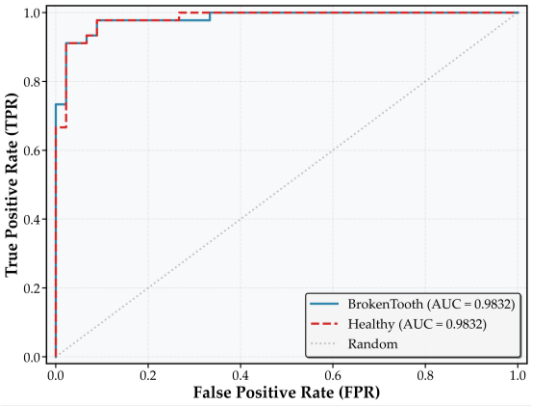


f) Full model

Figure 4: Comparative confusion matrices of the ablation studies.



a) Ablation-1



b) Ablation-2

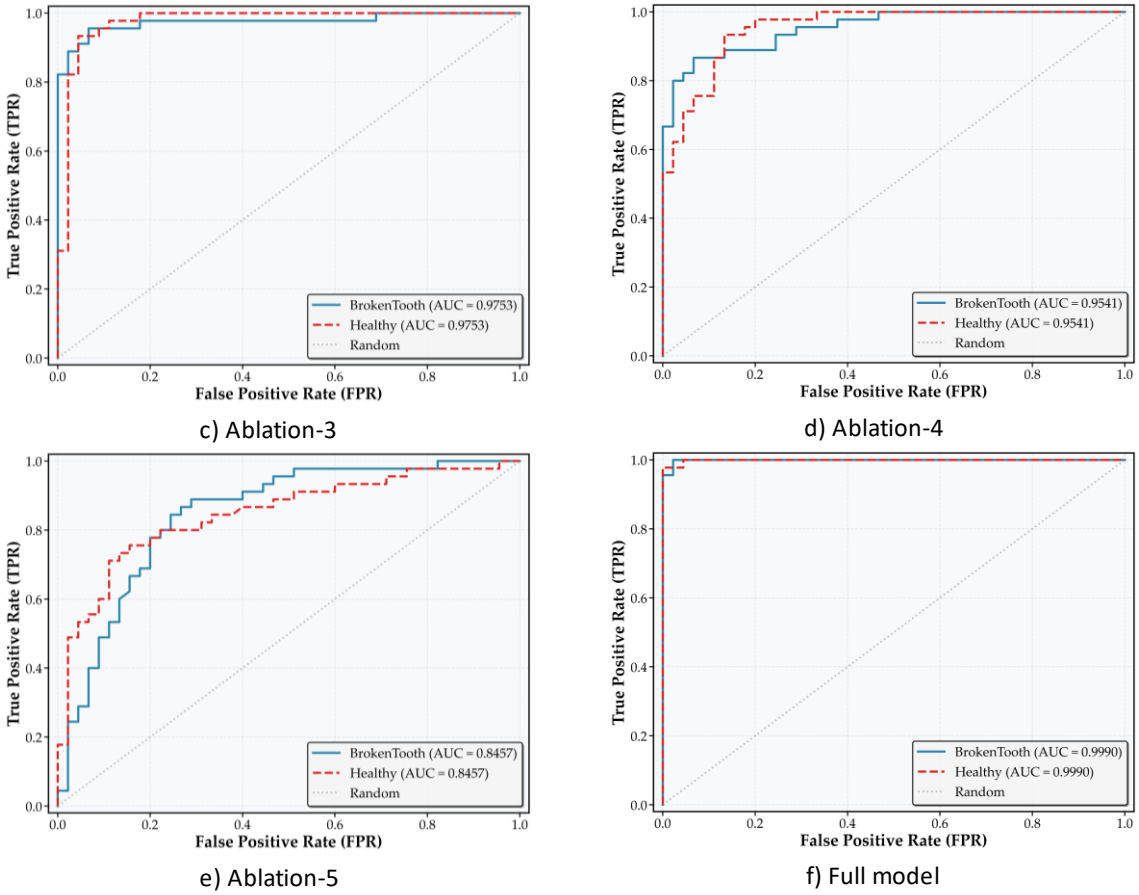
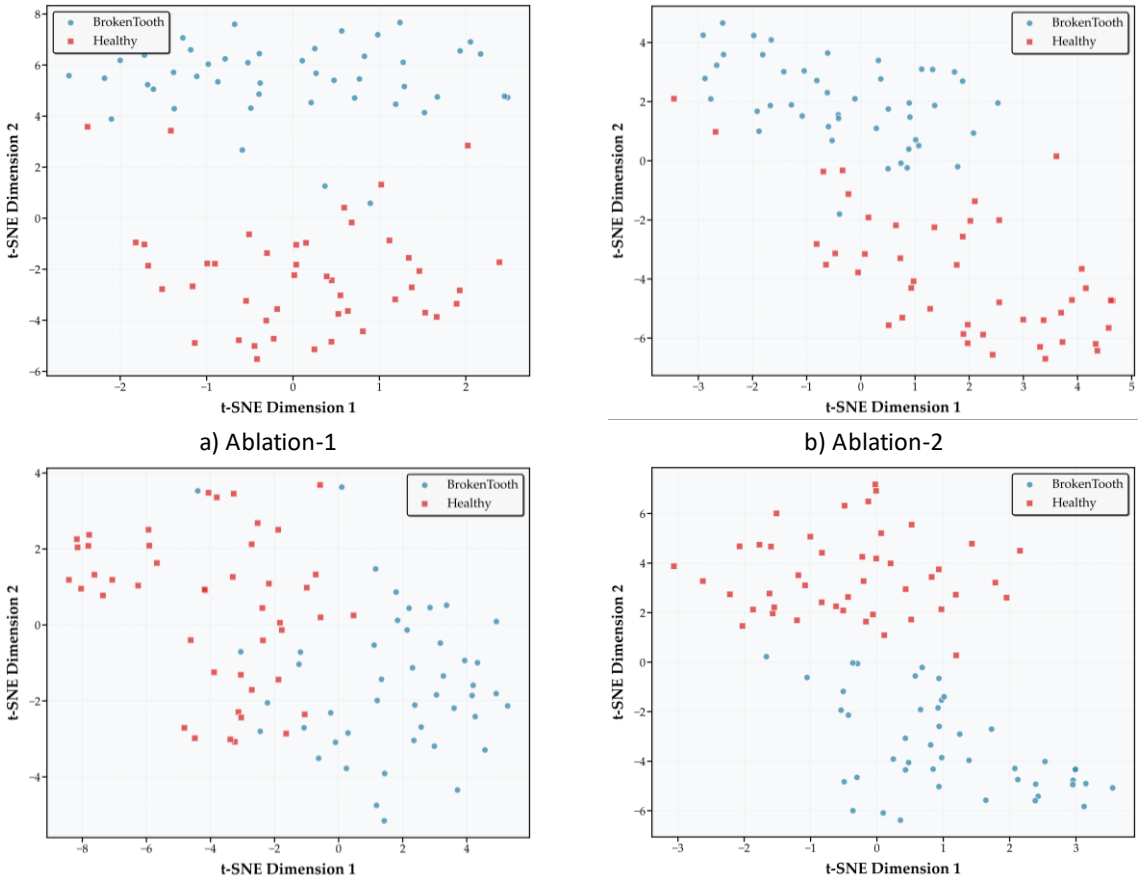


Figure 55: Comparative ROC curves of the ablation studies.



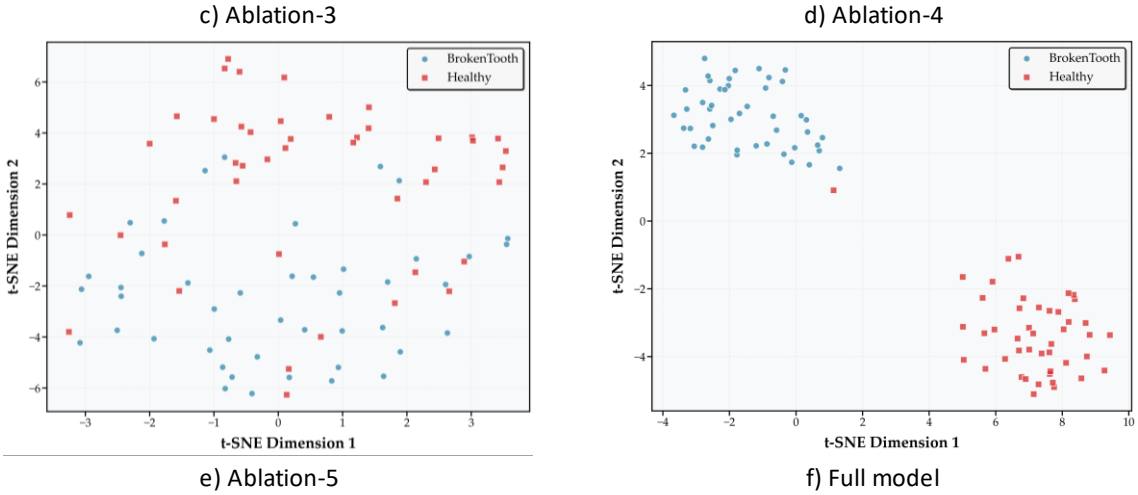


Figure 6: Comparative t-SNE visualization of the ablation studies.

As shown in Figure 13 and Table 9, the ablation study was designed to quantify the contribution of the individual components of the proposed architecture. The results demonstrate that removing any of the major components leads to a noticeable degradation in overall classification performance. The removal of the multi-head component (Ablation-2) reduces the test accuracy from 98.00% ± 3.28% for the proposed model to 85.33% ± 9.44%, while removing the CNN component (Ablation-3) decreases the test accuracy to 88.44% ± 7.88%. Similarly, removing the Bi-LSTM component (Ablation-1) results in a test accuracy of 88.22% ± 10.41%. These results indicate that the CNN, Bi-LSTM, and Multi-Head components each provide complementary information for the classification task.

The largest performance degradation is observed when Bi-LSTM and Multi-Head components are simultaneously removed (Ablation-4), for which the test accuracy decreases substantially to 63.33% ± 17.18%, compared with 98.00% ± 3.28% for the proposed model. The corresponding test F1-score also decreases from 0.9799 ± 0.0330 to 0.5882 ± 0.2150. Removing CNN together with Multi-Head (Ablation-5) similarly reduces the test accuracy to 70.44% ± 11.51% and the test F1-score to 0.6636 ± 0.1854. These results demonstrate that the combined removal of multiple components causes a considerably greater performance loss than the removal of a single component.

Small variations among the individual metrics are expected because of stochastic optimization, parameter initialization, and data sampling. Therefore, the contribution of each component should be interpreted by considering the overall performance pattern rather than a single metric. The proposed model achieves the highest performance across the reported training, validation, and test metrics, with a test accuracy of 98.00% ± 3.28%, test precision of 0.9822 ± 0.0282, test recall of 0.9800 ± 0.0328, and test F1-score of 0.9799 ± 0.0330. Its AUC also reaches 1.0000 ± 0.0000, compared with 0.9665 ± 0.0328, 0.9795 ± 0.0140, 0.9611 ± 0.0544, 0.7431 ± 0.2196, and 0.8660 ± 0.0579 for Ablation-1 through Ablation-5, respectively. Overall, these results confirm that the proposed architectural components provide complementary contributions, while their joint presence yields the most consistent overall classification performance.

When examining the confusion matrices in Figure 14, the proposed model produces the highest classification performance, whereas the ablated configurations exhibit progressively more classification errors, particularly when multiple components are removed. The most pronounced degradation is observed for Ablation-4 (No Bi-LSTM+MH), which is consistent with its lower test accuracy of 63.33% ± 17.18%. Ablation-5 (No CNN+MH) also shows a

considerable decrease, achieving 70.44% ± 11.51% test accuracy. These results further support the importance of the complementary feature-learning and temporal/attention components of the proposed architecture.

The t-SNE visualizations in Figure 16 provide an additional qualitative view of the learned feature representations. The proposed model exhibits well-separated class distributions with compact intra-class groupings and clearer inter-class boundaries. Compared with the ablated configurations, this visualization suggests that integrating the CNN, Bi-LSTM, and Multi-Head components enables the model to learn more discriminative and structured representations. The improved feature organization is consistent with the quantitative performance reported in Table 9.

4.8. Robustness Analysis

In real industrial environments, vibration signals are inevitably contaminated by various sources of noise, such as environmental disturbances, mechanical vibrations, and sensor measurement errors. Therefore, it is essential to evaluate the robustness of the proposed model not only on clean vibration signals but also under noisy operating conditions. To this end, the proposed model was evaluated during the testing phase using noisy vibration signals with signal-to-noise ratio (SNR) levels of 5, 10, 15, and 20 dB. Gaussian noise was artificially added to the test signals, and its intensity was adjusted according to the corresponding SNR level. The 5 dB condition represents the strongest noise level, while higher SNR values correspond to lower noise intensity. For each SNR level, Accuracy, Precision, Recall, and F1-score were calculated to comprehensively assess the classification performance. The quantitative results obtained are summarized in Table 10, while the variation in classification accuracy with respect to different SNR levels is illustrated in Figure 17.

SNR (dB)	Accuracy	Precision	Recall	F1-Score
5	0.9081	0.8468	0.9121	0.8782
10	0.9465	0.9084	0.9485	0.9280
15	0.9650	0.9353	0.9709	0.9528
20	0.9671	0.9409	0.9703	0.9554

Table 10: Robustness analysis of the proposed model under various noise conditions.

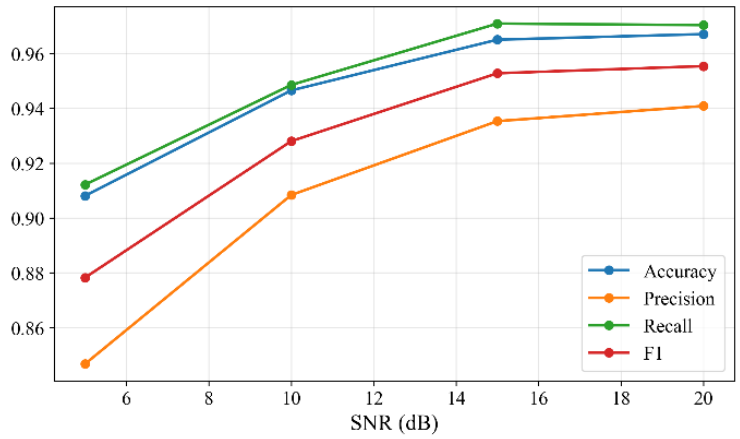


Figure 17: Classification performance of the proposed model under different SNR levels.

As can be observed from Table 10, the proposed model maintains an elevated level of classification performance even under severe noise conditions. Although the performance

gradually improves as the SNR increases, the degradation at lower SNR levels remains limited. Even at the most challenging condition of 5 dB, the proposed method achieves an accuracy of 90.81%, demonstrating its strong robustness against noise.

5. Conclusions

In this study, innovative deep learning architecture is proposed for gearbox fault diagnosis. Both temporal and spatial characteristics were effectively learned from raw vibration data through automatic feature extraction. The effectiveness and superiority of the proposed model were demonstrated through comparisons with various experiments. In addition to comparing the proposed model with various deep learning models in the literature, an ablation study was also conducted to demonstrate the effectiveness of each component of the model. Model performance was evaluated in terms of classification metrics, ROC curves, confusion matrices, and t-SNE visualization. The proposed method offers a reliable method for gearbox fault diagnosis thanks to its high accuracy, strong generalization ability, and noise resistance. In the future, the model will be evaluated on different gearbox types, varying operating conditions, and multiple fault scenarios.

References

Aurangabadkar, Ajinkya, Saurabh Kolhe, Safalya Shahakar, Harsh Dubey, Shankul Bhandare, and Pramod Lanjewar. 2023. “Fault Diagnosis of Gearbox Using Machine Learning Approach.” *International Journal of Innovative Research in Engineering* 4 (2): 608–15. https://doi.org/10.59256/ijire.2023040234.

Bhadane, D. B., Raj Kumar Shivankr, and P. K. Sharma. 2014. “Fault Diagnosis of Gear Box by Using MCSA Technique.” *International Journal of Engineering and Technical Research* 2 (12): 12–13. https://www.erpublication.org/published_paper/IJETR022910.pdf.

Brjapon. n.d. “Gearbox Fault Diagnosis.” *Kaggle*. Accessed July 10, 2026. https://www.kaggle.com/datasets/brjapon/gearbox-fault-diagnosis.

Chen, Z., C. Li, and R.-V. Sanchez. 2015. “Gearbox Fault Identification and Classification with Convolutional Neural Networks.” *Shock and Vibration* 2015: 390134. https://doi.org/10.1155/2015/390134.

Cheng, Xuemin, Shuihai Dou, Yanping Du, and Zhaohua Wang. 2024. “Gearbox Fault Diagnosis Method Based on Lightweight Channel Attention Mechanism and Transfer Learning.” *Scientific Reports* 14: 743. https://doi.org/10.1038/s41598-023-50826-6.

Fan, Caizi, Yongchao Zhang, Hui Ma, and Zeyu Ma. 2025. “A Novel Deep Denoising Model Integrating Transformer and Time–Frequency Loss for Gearbox Fault Diagnosis.” *Advanced Engineering Informatics* 66: 103400. https://doi.org/10.1016/j.aei.2025.103400.

Guo, Lei, Yongjian Wang, Jinxing Ju, and Zhengyong Liu. 2021. “Fault Diagnosis Method of Gearbox Based on Sound Pressure Signal Analysis.” *Journal of Physics: Conference Series* 1748 (4): 042044. https://doi.org/10.1088/1742-6596/1748/4/042044.

Han, Songyu, Haidong Shao, Junsheng Cheng, Xingkai Yang, and Baoping Cai. 2023. “ConvFormer-NSE: A Novel End-to-End Gearbox Fault Diagnosis Framework under Heavy Noise Using Joint Global and Local Information.” *IEEE/ASME Transactions on Mechatronics* 28 (1): 340–49. https://doi.org/10.1109/TMECH.2022.3199985.

Inturi, Vamsi, N. Shreyas, Karthick Chetti, and G. R. Sabareesh. 2021. “Comprehensive Fault Diagnostics of Wind Turbine Gearbox through Adaptive Condition Monitoring Scheme.” *Applied Acoustics* 174: 107738. https://doi.org/10.1016/j.apacoust.2020.107738.

Kumar, Anil, C. P. Gandhi, Yuqing Zhou, Rajesh Kumar, and Jiawei Xiang. 2020. "Latest Developments in Gear Defect Diagnosis and Prognosis: A Review." *Measurement* 158: 107735. https://doi.org/10.1016/j.measurement.2020.107735.

Kumar, Vikash, Subrata Mukherjee, and Somnath Sarangi. 2023. "Integrated Approach of Gearbox Fault Diagnosis." *arXiv* preprint arXiv:2308.14174. https://arxiv.org/abs/2308.14174.

Li, Chuan, René-Vinicio Sanchez, Grover Zurita, Mariela Cerrada, Diego Cabrera, and Rafael E. Vásquez. 2016. "Gearbox Fault Diagnosis Based on Deep Random Forest Fusion of Acoustic and Vibratory Signals." *Mechanical Systems and Signal Processing* 76–77: 283–93. https://doi.org/10.1016/j.ymssp.2016.02.007.

Lin, J., and M. J. Zuo. 2003. "Gearbox Fault Diagnosis Using Adaptive Wavelet Filter." *Mechanical Systems and Signal Processing* 17 (6): 1259–69. https://doi.org/10.1006/mssp.2002.1507.

Liu, B., S. Riemenschneider, and Y. Xu. 2006. "Gearbox Fault Diagnosis Using Empirical Mode Decomposition and Hilbert Spectrum." *Mechanical Systems and Signal Processing* 20 (3): 718–34. https://doi.org/10.1016/j.ymssp.2005.02.003.

Lu, Y., S. Shao, X. Yang, W. Zheng, Y. Zhao, and Y. Cheng. 2025. "Gearbox Fault Diagnosis Based on Multi-Source Information Fusion." In *Proceedings of the 2025 International Conference on Equipment Intelligent Operation and Maintenance (ICEIOM)*, 178–82. Piscataway, NJ: IEEE.

Saufi, Mohd Syahril Ramadhan, Zair Asrar Ahmad, M. Salman Leong, and Meng Hee Lim. 2020. "Gearbox Fault Diagnosis Using a Deep Learning Model with Limited Data Sample." *IEEE Transactions on Industrial Informatics* 16 (10): 6263–71. https://doi.org/10.1109/TII.2020.2967822.

Sekko, Edgard, Cécile Capdessus, Marie Sbai, and Emmanuel Attal. 2023. "Noise Robust Gearbox Defect Diagnosis." In *Proceedings of Surveillance, Vibrations, Shock and Noise*.

Sohaib, Muhammad, Shahid Munir, M. M. Manjurul Islam, Jungpil Shin, Faisal Tariq, S. M. Mamun Ar Rashid, and Jong-Myon Kim. 2022. "Gearbox Fault Diagnosis Using Improved Feature Representation and Multitask Learning." *Frontiers in Energy Research* 10: 998760. https://doi.org/10.3389/fenrg.2022.998760.

Wang, Siliang, Naige Wang, Anil Kumar, and Jianlong Wang. 2025. "Gear Fault Classification and Diagnosis Based on Gear Transmission Errors: Theoretical and Experimental Research." *Machines* 13 (12): 1093. https://doi.org/10.3390/machines13121093.

Xu, Lei, Tiantian Wang, Jingsong Xie, Jun Yang, and Guangjun Gao. 2022. "A Mechanism-Based Automatic Fault Diagnosis Method for Gearboxes." *Sensors* 22 (23): 9150. https://doi.org/10.3390/s22239150.

Zheng, H., Z. Li, and X. Chen. 2002. "Gear Fault Diagnosis Based on Continuous Wavelet Transform." *Mechanical Systems and Signal Processing* 16 (2–3): 447–57. https://doi.org/10.1006/mssp.2002.1482.