Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhanced Gear Fault Net Using Squeeze Group and Depthwise Convolutions with Grad-CAM

DURGA PRASAD CHARAKANAM

durgaprasadch16@gmail.com

Government Polytechnic https://orcid.org/0009-0005-8676-4734

Tirumala Rao Bennala

JNTUK UCEV: Jawaharlal Nehru Technological University Kakinada University College of Engineering Vizianagaram

Research Article

Keywords: Convolutional Neural Network (CNN), Deep learning, gear fault detection, Constant-Q Transform (CQT), Receptive Field Attention, Squeeze-Aggregated Excitation

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10450043/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Enhanced Gear Fault Net Using Squeeze Group and Depthwise Convolutions with Grad-CAM

Durga Prasad Charakanam[1*] and Tirimula Rao Benala[2]

[1*]Department of Electronics and Communication Engineering, Government Polytechnic Pillaripattu, Nesanur, Tirupati, 517583, Andhra Pradesh, Tirupati.

[2] Department Of Information Technology, JNTU-GV College of Engineering, Gurajada, Vizianagaram, 535003, Andhra Pradesh, Visakhapatnam.

*Corresponding author(s). E-mail(s): durgaprasadch16@gmail.com;
Contributing authors: btirimula.it@jntugvcev.edu.in;

Abstract

Gearboxes provide a stable transmission ratio and are essential in wind turbines and circulation systems. Reducing downtime and maintenance costs prevents costly equipment failures and mitigates safety risks. Early gear fault identification is essential to achieving these goals. Conventional methods, such as wavelet transforms and Fast Fourier Transform (FFT), exhibit limitations in frequency resolution due to fixed time-frequency trade-offs, particularly under non-stationary conditions. In this study, the Constant Q Transform, combined with deep learning, outperformed traditional signal processing techniques in analyzing statistical features extracted from both the time domain and the frequency domain of vibration signals in gearboxes. Incorporating attention mechanisms, Inception blocks, grouped convolutions, Receptive Field Attention, and Squeeze-Aggregated Excitation in the proposed deep learning model enhances feature extraction and enables accurate classification. The proposed method achieves an accuracy of 96.03%, a precision of 97.16%, and a specificity of 97.28% in distinguishing between healthy and broken tooth gearbox conditions. The experimental results demonstrate that the proposed method outperforms existing state-of-the-art techniques. Additionally, Grad-CAM analysis is applied to validate the model's decision-making process effectively, ensuring that the model focuses on relevant regions of the input data and enhancing the interpretability and reliability of the results.

Keywords: Convolutional Neural Network (CNN), Deep learning, gear fault detection, Constant-Q Transform (CQT), Receptive Field Attention, Squeeze-Aggregated Excitation

1 Introduction

Gearboxes are compact, efficient, and provide a stable transmission ratio, making them essential in wind turbines and circulation systems, but they are prone to faults due to complex designs and varying loads [1]. Fault Diagnosis aids in fault detection, classification, and severity assessment, enabling efficient maintenance and enhancing reliability [2]. Wind turbines, often located in remote areas, face high operational and maintenance costs, accounting for 25%-30% of total expenses, and challenges in minimizing downtime. Gearboxes, a critical component comprising 18% of turbine costs, experience the highest failure rates due to their complex structures and harsh operating conditions, necessitating effective fault diagnosis [3].

Downtime effects are more problematic for offshore wind turbines due to longer repair times. Electrical and electronic subassemblies experience shorter downtimes but fail more frequently, approximately every 2–2.5 years, whereas mechanical subassemblies, such as the drive train, fail less frequently, approximately every 19 years, but have longer downtimes. Early gear fault identification has the potential to reduce downtime and associated costs. The study by S. Faulstich et al. investigates how downtimes vary across different subassemblies based on frequent wind turbine (WT) failures and the duration of associated downtimes, particularly for electrical and electronic subassemblies. Only 5% of total downtime is due to minor failures, which constitute 75% of total failures in onshore wind turbines, while major failures, which account for 25% of total failures, contribute to 95% of total downtime. The study is based on the WMEP survey (1989–2006), which collected data from 1,500 onshore WTs, including 64,000 maintenance reports over 15,357 turbine-years, providing a robust dataset on reliability and availability. The analysis relies on retrospective data from WMEP reports [4]. Faults in gears exhibit two types of effects: dynamic effects, which are related to gear meshing, and structural effects, which are related to the signal transmission path. As FM increases, all seeded faults, including spalling, gear cracks, and pinion cracks, can be detected. The most pronounced effect in FM Out (tangential axis) is attributed to the fault location of the pinion crack. The study by Alexander Bliznyuk et al. investigates vibration signatures based on sideband features of Frequency Modulation (FM) [5].

The failure and repair cost curve in Figure 1 highlights vibration-based early fault detection's effectiveness in reducing maintenance costs compared to other methods. Early fault diagnosis, especially through vibration analysis, dramatically reduces maintenance expenses and prevents costly equipment failures. The vibration signals, which are non-stationary in nature, are analyzed in the order domain, and traditional statistical condition indicators are used to enhance signal analysis. Machine learning and deep learning approaches outperformed exclusive traditional signal processing techniques, such as wavelet transforms and Fast Fourier Transform (FFT), in handling the

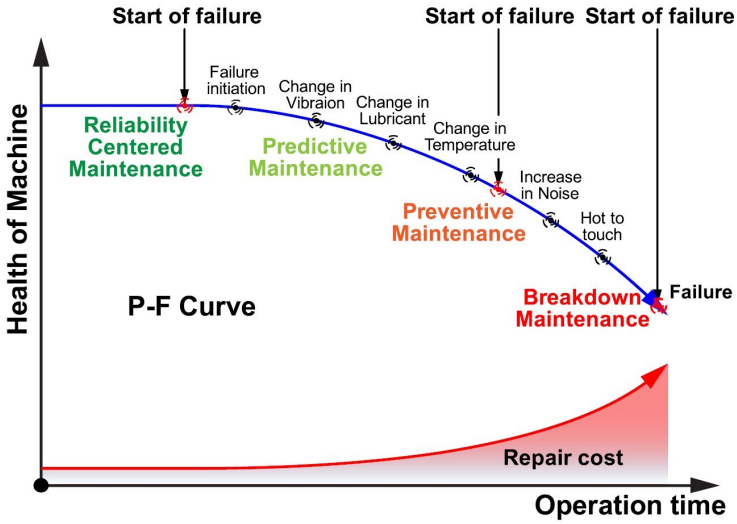

Fig. 1 Failure and Repair Cost Curve

non-stationary characteristics of vibration signals for fault detection without requiring manual feature engineering. A subset of machine learning, deep learning focuses on learning hierarchical feature representations directly from raw data. To prevent downtime, costly repairs, and safety risks, early fault detection was studied by ZhiQiang Chen et al., who trained a CNN for reliable fault diagnosis. This approach was based on a combination of a feature extraction method using statistical features extracted from both the time domain and the frequency domain [6]. The time-domain features included standard deviation, kurtosis, and skewness, while the frequency-domain features were derived from FFT band RMS values.

A Hanning window was applied for FFT processing, with the signal split into 251 frequency bands for RMS calculation. These frequency-domain features were then combined with statistical features, and the signals were pre-processed accordingly. In a study by Yunus Emre Karabacak et al., a model was developed using CSP-extracted features based on vibration and acoustic data and employing Artificial Neural Networks (ANNs) for classifying healthy and faulty worm gears, including wear, pitting, and tooth breakage [7]. The classification performance for binary fault detection demonstrated that ANNs outperformed other classifiers, such as k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM), in fault detection accuracy, while acoustic data exhibited superior performance in multiclass fault classification. As per the study by Sanjay Kumar et al., in the time domain, vibration analysis is highly effective for fault detection, particularly using kurtosis, which pinpoints defect locations. In frequency domain analysis, defect detection is performed through signal processing

and computing, which is the need of the hour. The current growing trend for bearing diagnostics is automated data interpretation (non-contact monitoring) machine learning, and deep learning techniques.

For specific defects such as spalls or missing teeth, techniques like artificial neural networks (ANNs) combined with wavelet transforms improve fault detection accuracy [8]. As per the review by Dhiraj Neupane et al., in the context of Industry 4.0, popular deep learning algorithms, including auto-encoders, Convolutional Neural Networks (CNNs), Deep Belief Networks (DBNs), Recurrent Neural Networks (RNNs)/Long Short-Term Memory (LSTM), Generative Adversarial Networks (GANs), and Reinforcement Learning (RL), are used for bearing fault detection in smart manufacturing [9]. However, auto-encoders are ineffective in detecting early-stage faults, while GANs struggle to reach Nash equilibrium, highlighting the overall superior performance of deep learning approaches.

The key contributions of this paper are as follows:

- This study proposes a novel fault detection model based on the Constant-Q Transform (CQT) spectrogram, which effectively captures detailed time-frequency features and demonstrates superior performance compared to traditional methods such as the Short-Time Fourier Transform (STFT) and the Continuous Wavelet Transform (CWT).
- The proposed model integrates attention mechanisms, Inception blocks, grouped convolutions, Receptive Field Attention, and Squeeze-Aggregated Excitation to enhance feature extraction and improve classification accuracy.
- A series of performance comparison experiments are conducted with other related methods.

The study is structured as follows: Section 2 reviews related literature, Section 3 explains the Constant-Q transform and the proposed network, Section 4 discusses the experimental results, and Section 5 concludes the study.

2 Literature Review

Gear fault detection signals are non-stationary, constantly changing due to dynamic operating conditions, fault progression, environmental noise, and transient events. Also, these signals are noise-sensitive, hindering fault detection under varying load conditions. This variability challenges traditional signal processing methods that assume consistent signal characteristics. Understanding the non-stationary nature of these signals is crucial for designing robust detection and diagnosis systems. To address these complexities, time-frequency analysis techniques are used to extract fault features from non-stationary vibration signals.

2.1 Methods Based On Time-frequency representation

Time-frequency representation (TFR) excels in analysing non-stationary signals, capturing transient information crucial for gear fault diagnosis. Time-domain signals are converted to the cycle-domain to detect and distinguish between different types of gear faults, such as spalling and cracking. To address the challenge of difficult feature

extraction, signal processing methods combined with deep learning models are used for fault diagnosis. Integrating deep learning models can enhance adaptability, improving fault diagnosis accuracy under limited data scenarios. Techniques like CWT, STFT, and FSST convert signals to TFR, enabling effective deep learning (DL) input while minimizing information loss in convolutional processes.

Nguyen et al. [10] introduced a Transfer Learning-based method using TFR images and a fine-tuned Wavelet ResNet-50 for efficient gear fault diagnosis. Similarly, Kim et al. [11] proposed a DNN-based gear defect detection system using frequency band sums for real-time diagnosis with 95.5% accuracy, reducing computational demands. Meanwhile, Patil et al. [12] proposed a model for gear tooth failure detection using the YAMNet network through transfer learning. By extracting Mel spectrograms from noise data and fine-tuning the pre-trained network, the model was developed for fault detection. However, the study's limitation is that only a single type of gear fault was explored, and the generalizability of the model requires further validation.

Furthermore, the study by Ravikumar et al. [13] introduces a fault diagnosis model combining multiscale CNN with residual learning and LSTM layers (MDRL-SLSTM) to classify gearbox faults effectively. It achieved high classification accuracy of 94.33% for gears using vibration signal data from an IC engine gearbox. CNNs were employed for feature extraction, with residual learning addressing vanishing gradient challenges, and LSTM layers handling sequential data dependencies. Furthermore, Hu et al. [14] employ discrete wavelet decomposition, Markov transition fields, and an improved VGG-based CNN with data augmentation to achieve 88.33% gear fault classification accuracy. However, the challenges in classifying gearbox in the above studies is that faults arise from reliance on specific setups, difficulty distinguishing similar fault types, and dependence on frequency bands, limiting applicability to other machinery. Additionally, the small dataset and focus on a single fault type further restrict the generalizability and effectiveness of the classification methods.

2.2 Methods Based On Architecture

To address the challenge of difficult feature extraction, signal processing methods combined with deep learning models are used for fault diagnosis. Integrating deep learning models can enhance adaptability, improving fault diagnosis accuracy under limited data scenarios. In the study by Imran Ahmed et al., a deep learning approach was proposed using a 6-layer deep autoencoder framework, which can learn representative features from normal vibration data for anomaly detection and condition monitoring. Additionally, the autoencoder is used for dimensionality reduction and feature learning. This framework leverages vibrational signal analysis, incorporating both time-domain and frequency-domain characteristics such as Power Spectral Density, to analyze vibration signals and identify faults, achieving an overall accuracy of 91%. The study is based on a dataset collected using SpectraQuest's simulator equipped with four vibration sensors [15]. In the study by P. Dutta et al., the proposed GearFaultNet model is a lightweight 1D Convolutional Neural Network (CNN) designed for detecting gearbox faults using preprocessed vibration data. Standard preprocessing steps, including z-score and min-max normalization, are applied, along with 50% overlapping augmentation per channel to increase the training data, followed by global

min-max normalization per segment. The model is optimized using the Adam optimization algorithm. However, a limitation of the study is that the dataset is derived from a single simulator [16].

In the study by Kefan Cheng et al., a proposed model, the combined CNN-LSTM model augmented with an attention mechanism (CNN-LSTM-Attention), was utilized to extract spatial features from gearbox vibration data using CNN and temporal features using LSTM. Additionally, an attention mechanism was applied to the LSTM outputs to assign weights based on feature importance. The study was conducted on the Southeast University (SEU) gearbox dataset. However, the generalizability of the current study might be limited to the specific dataset and conditions tested [17]. In the study by Yong Yao et al., an Attention-Based Multi-Scale CNN was proposed as a novel acoustic-based diagnosis (ABD) method, integrating multi-scale 1D convolutions, a Temporal Attention Mechanism, and a stacked 2D CNN classifier to analyze acoustic signals under different working conditions. The CNN was utilized for automatic feature extraction from time-series acoustic signal data, leveraging the concept of multi-scale processing to capture information at different temporal resolutions. The attention mechanism, based on the multi-scale structure, adaptively focused the network on relevant fault pattern information using Cross-Entropy Loss [18]. The proposed model achieved high accuracy, specifically 93.3% and 82.8% in two cross-domain evaluation scenarios. However, the study had certain limitations, as the experiments were conducted in a controlled semi-anechoic chamber, which does not reflect real-world noisy industrial environments.

Additionally, the research focused solely on single gear faults, including wear, pitting, and fracture, in a specific two-stage gearbox, without testing the method's effectiveness on other common component faults such as bearing faults, multi-fault scenarios, or coupling faults. In the study by Rishikesh Magar et al. [19], a custom CNN model, FaultNet, was proposed for time-series classification, where a convolutional neural network (CNN) is utilized. The model employs stacked mean and median channels alongside the raw signal and incorporates Wavelet Packet Decomposition (Haar wavelet) for time-frequency feature extraction. By augmenting the raw vibration signal with additional channels, the approach leverages a greater number of stacked channels, including the raw signal, mean channel, and median channel, to extract more differentiating features compared to a single-channel input, thereby improving performance. However, a limitation of the model is observed in the slight confusion between ball fault (0.18mm) and outer race fault (0.36mm), indicating a minor difficulty in distinguishing these specific fault types and sizes perfectly within that dataset.

Studies related to gearbox fault classification face major issues, including limitations due to reliance on specific datasets, sensor quality, and preprocessing requirements. Additional challenges include a lack of adaptability to unseen datasets, sensitivity to wavelet decomposition levels, and the constraint of methods being often limited to binary classification. Furthermore, reduced generalizability arises due to reliance on lab-scale datasets, while high computational requirements also pose significant limitations that need to be addressed. Challenges such as potential overfitting, performance variability in different conditions, and reduced effectiveness in noisy environments further necessitate continued research in this domain. To address the

problems identified in earlier literature, this study proposes a novel network that integrates cutting-edge innovations such as modified Inception blocks, Receptive Field Attention, and Grouped Convolutions. These enhancements improve feature extraction, computational efficiency, and scalability compared to traditional architectures. The architecture leverages Modified Inception Blocks for multi-scale feature extraction, Receptive Field Attention for dynamic spatial and channel recalibration, and Grouped Convolutions for efficient parallel feature learning. Also, in this study, we use CQT for spectral analysis, transforming vibration signals into a two-dimensional time-frequency spectrogram images where feature extraction is based on frequency bins and harmonic relationships. This study helps significantly reduce downtime and improve operational efficiency by detecting early signs of mechanical failure and ensuring timely interventions.

3 Materials and Methodology

3.1 SpectraQuest's Gearbox Fault Dataset

This study utilizes a vibrational dataset generated from SpectraQuest's Gearbox Fault Diagnostics Simulator (GFDS), a tool designed for replicating industrial gearboxes to facilitate research in diagnostics and prognostics. Figure **??** depicts SpectraQuest's Gearbox Prognostics Simulator (GPS), a specialized system designed for simulating and analyzing gearbox fault conditions to support prognostics research and development. The GFDS simulator offers a versatile platform for researching gearbox health monitoring, dynamics, and vibration-based diagnostics. It includes vibrational data from four sensors placed around a gearbox under various load conditions (0% to 90%). It supports various gearbox configurations, gear ratios (1 to 6), and bearing types (rolling or sleeve bearings). The simulator collects data through multiple sensors, such as accelerometers affixed to the gearbox body to capture vibrations in all three directions. The dataset enables the comparison of the gearbox's performance under normal and faulty conditions, facilitating the development of machine learning models aimed at predictive maintenance and fault diagnosis. The dataset is categorized into two classes: Healthy and Broken, with 10 different load conditions (0%-90%) for each class. Each condition is represented by data from four sensors, with a sampling frequency of 30 Hz. The total dataset comprises 20 files, with 4 files for each sensor. Each file contains between 85,000 to 120,000 samples, making it a robust resource for researchers to refine techniques for diagnosing and predicting faults in industrial machinery.

The data is accessible via the data.world repository and offers significant potential for improving gearbox fault diagnosis and maintenance strategies.

3.2 Case Western Reserve University Bearing Dataset

The Case Western Reserve University Bearing Data Center conducted an experiment using a 2-horsepower Reliance Electric motor, where acceleration data was measured at locations near to and remote from the motor bearings. The dataset contains test data for both normal and faulty bearings, with faults ranging from 0.007 inches to 0.040 inches in diameter, introduced separately at the inner raceway, rolling element

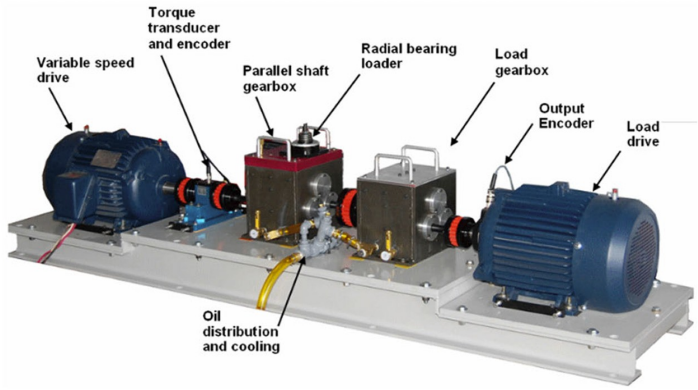

Fig. 2 SpectraQuest's Gearbox Prognostics Simulator (GPS)

(i.e., ball), and outer raceway. Motor bearings were seeded with faults using electro-discharge machining (EDM), after which the faulted bearings were reinstalled into the test motor, and vibration data was recorded for motor loads ranging from 0 to 3 horsepower, corresponding to motor speeds between 1797 and 1720 RPM. Data was collected for normal bearings, as well as for single-point drive end and fan end defects, at sampling rates of 12,000 samples per second and 48,000 samples per second. Specifically, data for drive end bearing experiments was collected at both 12,000 and 48,000 samples per second, while all fan end bearing data was collected at 12,000 samples per second. Each data file contains motor rotational speed, as well as fan and drive end vibration data. Vibration data was collected using accelerometers attached to the housing with magnetic bases, while speed and horsepower data were gathered using a torque transducer/encoder and recorded manually.

Gear faults can include normal wear, tooth wear (abrasive and interference), tooth pitting, gear cracks, and tooth breakage. Normal wear is expected from friction, while tooth wear affects the gear profile. Pitting occurs due to high-frequency stress and low surface hardness. Gear cracks result from fatigue or overload, and tooth breakage can be caused by extreme overload or impact. In this study, we focused only on the broken tooth type of gear fault. Figure 3 displays the different types of gear faults. In the first stage, sixty-second segments are generated from the input vibration signals. In the second stage, the CQ Transform is applied to convert each segment from a one-dimensional signal into a two-dimensional time-frequency representation.

During the third stage, each two-dimensional image is input into the deep learning model for training. Figure 4 illustrates the proposed model for gearbox fault detection, while Table 1 summarizes recent literature on approaches to the automatic diagnosis of gearbox faults.

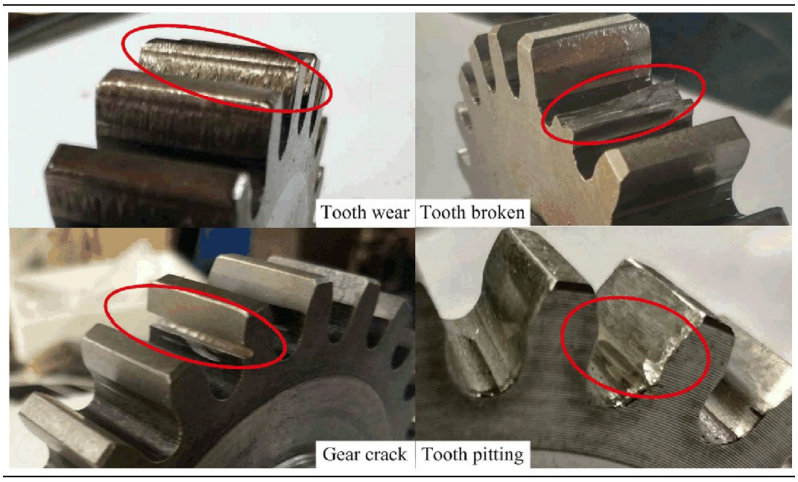

Fig. 3 Gear Fault Broken Condition

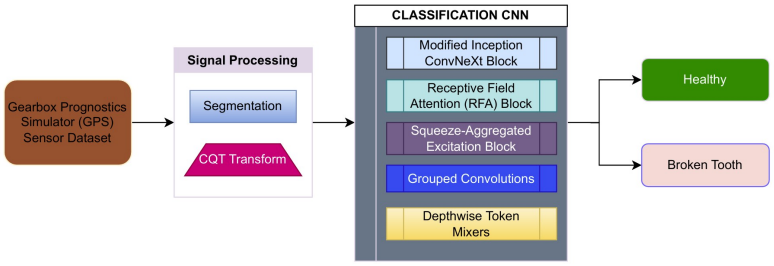

Fig. 4 Overview of the proposed model

3.3 Sensor Data Processing and Analysis Methodology

In this study, we process and analyze sensor data from the dataset to classify machine states as either Healthy or Faulty. This process involves calculating sample counts across different load conditions, estimating recording durations, and identifying the minimum sample size. Recording durations are estimated by combining sample counts with the sampling rate, facilitating the identification of incomplete or abnormally short recordings. Time series analysis is conducted using signal segmentation, where signals are divided into 60-second segments at a sampling rate of 30 Hz. The Constant-Q Transform (CQT), implemented via the librosa library, is applied to generate spectrograms, with specified parameters including the number of bins per octave, the minimal frequency, and the total number of bins. These spectrograms are resized, converted to RGB format, and labeled with the dataset type, sensor, file name, and segment index

for machine learning training. Gear conditions are categorized as either Healthy or Faulty.

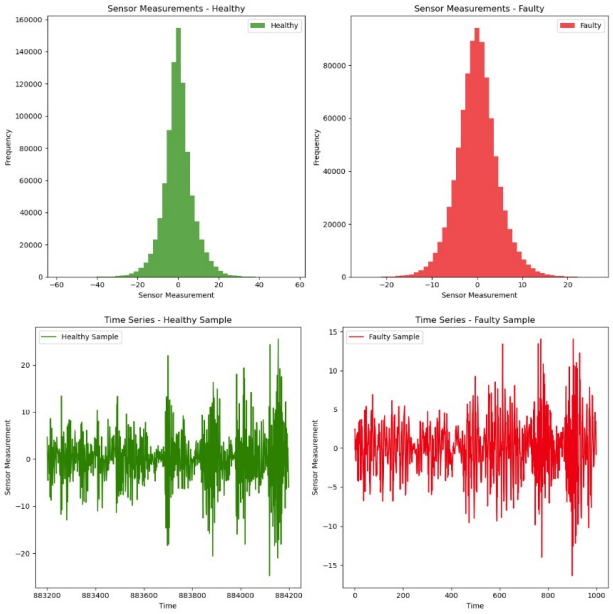

Fig. 5 Time series and Histograms of Sensors for Healthy and Faulty Samples

Data standardization and modular pipeline design enhance the quality and reliability of the datasets, ensuring their suitability for developing robust machine learning models and fault analysis. Standard deviation calculations are performed to quantify variability in sensor readings, aiding in the identification of inconsistencies or noise within the data.

Additionally, excluding specific sensors during analysis allows for the isolation of their impact on variability, facilitating the detection of faulty or miscalibrated sensors. This methodological approach ensures that the dataset accurately reflects system conditions by addressing potential sources of error. Finally, variability comparisons across different states, such as Healthy and Faulty, contribute to robust fault detection and enhance decision-making processes for system maintenance.

3.4 Constant Q Transform

Fault diagnosis involves vibration feature extraction using STFT, FFT, HT, WT, and envelope analysis [20]. Traditional spectrograms may face limitations in frequency resolution due to fixed time-frequency trade-offs, particularly in nonstationary conditions. The Constant-Q Transform (CQT) is a time-frequency representation method using a logarithmically spaced frequency scale, where each frequency bin has a constant ratio of frequency to bandwidth. The CQT aids in the early detection of faults under varying operational speeds and conditions. Compared to spectrograms (based

on STFT), CQT offers better frequency resolution at low frequencies, essential for identifying subtle fault characteristics in gearbox systems, logarithmic frequency scaling that matches the natural frequency characteristics of many mechanical systems, and enhanced sensitivity to fault features, particularly useful for detecting weak or early-stage faults. Time-frequency features extracted using CQT significantly improve the accuracy of fault classification models, such as those based on machine learning algorithms such as support vector machines and neural networks [21].

Mathematically, the CQT $C_Q(f)$ of a signal $x(t)$ is defined as the convolution of the signal with a set of complex sinusoidal filters, each centered at a specific frequency f and having a bandwidth determined by a quality factor Q. The transform is expressed as:

$$C_Q(f) = \int_{-\infty}^{\infty} x(t) \cdot g_{Q,f}(t) dt \tag{1}$$

$$f_n = f_0 \cdot r^n \tag{2}$$

In the study by Judith C. Brown et al., it is stated that transforming a standard DFT/FFT output using pre-calculated spectral kernels is a computationally more efficient method for obtaining a constant Q transform. This method involves pre-calculating spectral kernels that are applied to the FFT result, with the efficiency primarily stemming from the sparsity of these spectral kernels. The approach is based on the theory of the Discrete Fourier Transform (DFT), specifically leveraging Parseval's theorem to relate time-domain correlation to frequency-domain multiplication. Additionally, the constant Q transform often provides a better spectral representation for musical signals than the standard FFT [22]. As per the study by Lhoucine Bahatti et al., the research is based on sinusoidal modeling derived from Fourier analysis, extended for non-stationary signals using short-term analysis. It incorporates amplitude and frequency modulation (AM/FM) models with a quadratic phase [23]. The key theoretical components include the Constant Q Transform (CQT) concept for frequency-dependent window sizing and the generalized Reassignment Method for parameter estimation using STFT derivatives. Standard models struggle with non-stationarities common in music. By combining short-term sinusoidal modeling with CQT-based variable windowing and reassignment techniques, the study accurately estimates the time-varying parameters, including amplitude, frequency, and phase, of oriental music signals based on a weighting window using the Hann window.

3.5 Proposed deep learning architecture for fault detection in gearboxes

The CNN model begins with a convolutional layer containing 64 filters and a 7x7 kernel, followed by max-pooling to reduce the spatial dimensions. Batch normalization stabilizes the activations. Next, a modified Inception-ConvNeXt block is applied, consisting of convolutions (1x1, 3x3, 7x7) and max-pooling. The outputs are concatenated after pooling and batch normalization. Receptive Field Attention recalibrates both channel and spatial features. The Squeeze-and-Excitation block recalibrates channels using global average pooling, followed by a dense layer.

Grouped convolutions split the filters into groups and perform convolutions, increasing depth. Depth-wise token mixers use depth-wise convolutions to process

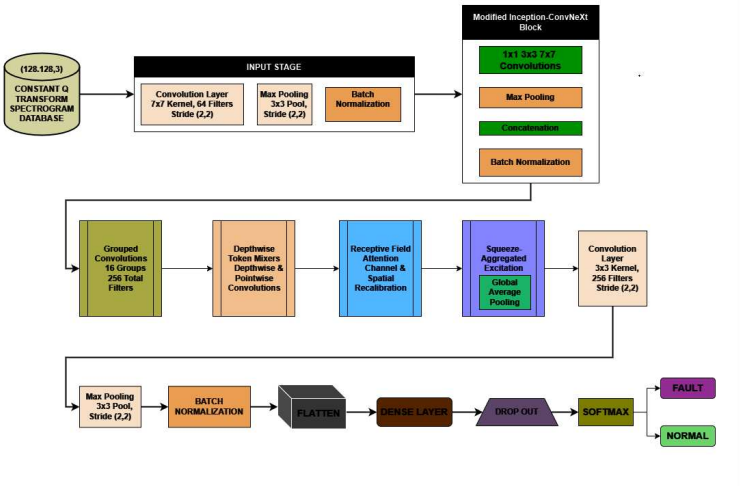

Fig. 6 Novel CNN with Inception Blocks, Grouped Convolutions, Receptive Field Attention, and Squeeze-Aggregated Excitation Mechanism

spatial information while maintaining depth. Additional convolution and pooling layers further reduce the spatial dimensions. Flattening converts the output to a 1D vector, followed by a dense layer for classification. The output layer produces class probabilities using softmax. The final output shape depends on the binary classification outcomes. This architecture combines advanced techniques to extract and refine features at various scales, ensuring effective feature representation and classification.

3.5.1 Modified Inception ConvNeXt

In the study by Mehmet Nergiz et al., the ConvNeXt architecture, which incorporates modernization principles inspired by Vision Transformers and Swin Transformers into a pure ConvNet design based on ResNet/ResNeXt, demonstrates superior classification performance in terms of Accuracy, AUC, F1-score, and Cohen's Kappa compared to standard CNN architectures (AlexNet, DenseNet, SqueezeNet, Inception v3, and VGG16) and Vision Transformers (ViTs) on the MHIST histopathological image dataset [24]. These findings indicate that design principles from Transformers can significantly enhance the performance of pure CNN architectures. The study utilized the AdamW optimizer and a cosine learning rate schedule.

Decomposing a large-kernel depthwise convolution into parallel branches, including a small square kernel, two orthogonal band kernels, and identity mapping along the channel dimension, reduces memory access costs and FLOPs compared to the original large kernel. This decomposition, based on ConvNeXt, enables a network built with it to achieve significantly higher throughput. In the study by Xinchao Wang et al., ConvNeXt is recognized for its modernized CNN macro-architecture and the use of large kernels as a starting point. The orthogonal band kernel branches $(1 \times k_b, k_b \times 1)$

are crucial for maintaining performance, likely by efficiently preserving a large receptive field. Additionally, Batch Normalization was preferred over Layer Normalization primarily due to its higher throughput [25].

In the study conducted by El Gannour et al. on the COVID-19 chest X-ray images dataset, a convolutional neural network (CNN) using transfer learning (TL) was implemented by concatenating features from multiple pre-trained CNN models, which improved the performance of COVID-19 prediction. The application of a parallel deep feature extraction approach via concatenation represents a novel technique. Six pre-trained CNNs—Xception, InceptionV3, InceptionResNetV2, ResNet50V2, MobileNetV2, and DenseNet201—were modified by adding custom dense layers at the output, and models were subsequently created by concatenating pairs of these modified models [26]. Efficient feature extraction involves Depthwise Convolution for computational efficiency, Multi-Scale Feature Extraction using varying kernel sizes (K_1, K_3, K_5) for diverse features, and Pointwise Convolution (1x1) for channel mixing. Stability is ensured by Layer Normalization, followed by GELU Activation for enhanced representation. The combined process can be expressed as:

$$Z_{\text{activated}} = \text{GELU}\left(\text{LayerNorm}\left([X * K_{\text{depthwise}}]_{(K_1,K_3,K_5)} * K_{\text{pointwise}}\right)\right) \quad (3)$$

This design combines the simplicity of ConvNeXt with the adaptability of Inception, ensuring a computationally efficient yet diverse feature representation.

3.5.2 Receptive field attention

Channel and spatial attention mechanisms are integrated to enhance feature extraction, a process that is carried out by the RFA Block. Through global pooling and advanced attention computation, it effectively emphasizes informative image regions, suppresses irrelevant feature noise, and enhances contextual understanding. This leads to improved contextual feature representation, reduced overfitting risk, and computational efficiency. These attributes make the RFA Block particularly effective in tasks such as image classification, object detection, and semantic segmentation. In the study by Minjie Tan et al., a novel module called the Dynamic Receptive Field (DRF) mechanism was proposed, which integrates elements from attention mechanisms and transformer-style relative position encoding, adapting them into a distance-based correlation model [27]. By incorporating the DRF module into CNN backbones, specifically ResNet50, ResNeXt50, ResNet101, and ResNeXt101, the mechanism enables global information-guided responses, as the neuron's dynamic receptive field mechanism can effectively capture both local and global contextual relations, thereby improving performance on object detection and image recognition tasks. In the study by Lin Zhou et al., a model was proposed using a receptive-field Pixel Attention Block (VAB) that incorporates depth-wise separable convolutions [28].

The use of depth-wise separable convolutions, both standard and dilated, in the attention branch—comprising channel attention (RCAN), spatial attention (PAN), combined attention (HAN), and self-attention (IPT, SwinIR)—improves Super-Resolution performance. Additionally, depth-wise separable convolutions are applied to reduce the number of parameters and overall model size [28]. In the study by

Xianghua Ma et al., the authors proposed FSRFNet models based on the concept that the receptive field size of visual cortical neurons and the additive effect of feature-selective attention and spatial attention can be utilized for implementing adaptive receptive field sizes via the proposed FSRF block. The proposed FSRFNet models leverage feature-selective and spatial receptive fields along with feature-selective and spatial attention mechanisms. In the study by Yong Yao et al., the authors proposed a model with an acoustic-based diagnosis (ABD) method, utilizing a multi-scale convolutional neural network (CNN) combined with a novel attention mechanism to adaptively extract relevant fault patterns and reduce sensitivity to data distribution shifts caused by varying conditions [29]. The multi-scale feature extraction module included three parallel 1D convolutional branches using different filter sizes. A temporal attention mechanism was applied to the outputs of the multi-scale branches, using 1×1 and subsequent convolutions to generate attention. A subsequent stacked CNN was used for feature learning and classification. A limitation of the study is that experiments were conducted in a controlled semi-anechoic chamber, potentially limiting the direct applicability to noisy industrial environments.

Channel-wise attention is computed using global average pooling, which calculates the mean, and global max pooling, which identifies the maximum value across spatial dimensions, highlighting key features. Spatial attention uses a 3×3 convolution followed by a sigmoid activation to generate attention maps that focus on significant spatial regions. The final output combines the input features with channel and spatial attention components, refining the feature representation for improved effectiveness.

3.5.3 Squeeze-Aggregated Excitation

The SE (Squeeze-and-Excitation) block enhances CNN performance with minimal computational overhead while improving feature representation. It achieves this by using Global Average Pooling to capture channel relationships, applying two fully connected layers with a ReLU activation function, and using a sigmoid function to generate normalized channel weights. In the study by Mahendran et al., a model with a Squeeze-and-Excitation (SE) module was proposed, which was modified to enhance the performance and feature learning capabilities, where channel-wise feature recalibration is used to enhance representational power by modeling interdependencies between channels; additionally, delayed squeeze/excite steps and an added internal layer were incorporated to better utilize global dependencies and enhance network learning, as SE-like modules can impact learning dynamics [30]. In another study by Mahendran N et al., the author proposed a model that induces global representations within channels using aggregated excitation through multi-branch linear layers after the squeezing operation, which enhanced the representational power of Convolutional Neural Networks (CNNs) compared to existing methods such as Squeeze-and-Excitation (SE) networks [31].

In this approach, a Squeeze Aggregated Excitation Network was used for learning global channel-wise representations between layers. The Squeeze-and-Excitation Networks (SEnet) module had increased cardinality in the fully connected layers to improve representational power, and the aggregated fully connected layers acted as a

method for learning global representations within the network architecture by introducing parallel processing paths (aggregation) in the excitation phase's fully connected layers. Additionally, aggregated transformations, as in ResNeXt, improved representational power through multi-branch architectures by increasing cardinality. The study proposed the combination of spatial, channel-wise, and the enhanced global representations within the network architecture. However, the experiments were conducted with computational constraints, limiting training to only 50 epochs. In the study by Justin Lo et al., a U-Net-based deep learning architecture was proposed by applying adaptive weights to feature channels using SE blocks for adaptive channel-wise feature recalibration and emphasizing these weighted features using attention mechanisms for focusing on salient spatial features, while also emphasizing the localization of relevant contextual information with preprocessing through intensity normalization [32]. The U-Net backbone was modified by adding SE blocks after max-pooling in the encoder and Attention Gates before concatenation in the decoder, utilizing 5×5 kernels and transposed convolutions. A limitation of the study is that the dataset size, consisting of 34 patients, is relatively small for deep learning.

3.5.4 Grouped Convolution

Grouped convolutions are an efficient technique to enhance the computational efficiency of a CNN by partitioning the input channels and then applying convolutions independently within each group. In the study by Perry Gibson et al., a new implementation algorithm for grouped convolutions, named Grouped Spatial Pack Convolutions (GSPC), was proposed, which modifies and extends spatial pack convolutions to handle groups and exploit data locality. This approach achieves lower inference times compared to existing implementations, improves execution speed, and better realizes the potential performance gains from reduced computation. The concept of grouped convolutions serves as a network compression technique by reducing parameters and multiply-accumulate operations (MACs). In the study by Yong Li et al., a dual attention mechanism, incorporating both channel and spatial attention, was proposed within grouped convolution blocks inspired by Res2Net to enhance the network's ability to focus on relevant features for sensor-based Human Activity Recognition (HAR). The multi-branch network architecture, which combines grouped convolutions and dual attention, achieves better performance on standard HAR datasets.

This approach significantly reduces computational complexity and memory usage while preserving the network's feature extraction capabilities. As a result, grouped convolutions are particularly effective in tasks like image classification, object detection, and real-time computer vision applications, enabling efficient parameter management and facilitating deeper, more complex architectures with performance comparable to standard convolutions.

3.5.5 Depthwise Token Mixers

In the study, Abdullah et al. [33] introduces a new computational block that replaces attention layers in Vision Transformers with a Network in Network structure, enhancing MLP-Mixer's static approach with dynamic token mixing for improved

image classification performance. Depthwise Token Mixers (DTMs) offer an innovative approach for efficiently mixing tokens in deep neural networks, particularly in transformer-based architectures. By combining depth-wise convolutions with advanced token mixing techniques, DTMs address the challenge of processing high dimensional feature representations with reduced computational overhead compared to traditional attention mechanisms. Also, as per the study Lu et al. [34], the DTM method applies depth-wise convolutions to independently mix tokens across channels, striking a balance between computational efficiency and sophisticated feature extraction. As a result, DTMs represent a significant advancement in the development of more accessible and performant neural network architectures for complex machine learning tasks. The mathematical framework of DTMs involves an input representation, where B denotes the batch size, T the number of tokens, and C the feature dimension. DTMs are particularly effective in computer vision, natural language processing, and high-dimensional feature learning. Their key advantages include significant computational efficiency, reduced parameter count, and scalability across high dimensional inputs. By enabling efficient token processing while maintaining rich representational capabilities, DTMs facilitate the creation of more accessible deep learning models suitable for diverse computational environments.

Grad-CAM computes class-specific feature importance by calculating gradients of target class predictions with respect to feature maps. These gradients are averaged across spatial dimensions to determine neuron importance weights. The final visualization is created by combining weighted feature activation maps with ReLU, enabling detailed, class-discriminative localization without modifying the network architecture [35]. In the study by Harikumar et al., LIME enhances interpretability of CNN-based fetal ultrasound image classification by generating explanations through input perturbations and analyzing output changes [36]. Study by Lerma et al. introduces GC-MRCNN, a method combining Grad-CAM++ and Mask R-CNN for object detection and localization in images. Grad-CAM++ generates high-resolution heatmaps for precise object localization, while Mask R-CNN performs detection and segmentation. The backbone network, ResNet 101, extracts features efficiently. The training process uses learning rate annealing and ROI pooling. GC-MRCNN's key contribution is its ability to recognize multi-scale objects in the COCO dataset, offering improved texture-based decisions for easier localization. By integrating Grad-CAM++ with Mask R-CNN, the method achieves superior performance compared to existing techniques, providing enhanced visual explanations and accurate object localization [37]. In this study, we have incorporated the Grad-CAM technique to generate heatmaps to validate our work. In this study, Grad-CAM highlights key CQT spectrogram regions influencing predictions by computing gradients of the target class relative to convolutional activations, aiding interpretability. Feature maps reveal hierarchical feature extraction, with lower layers capturing edges and deeper layers identifying fault-related patterns. Figure **??** illustrates the visualization of feature maps at layer 20, providing insights into the regions of the input image that are most influential for the model's predictions in the Grad-CAM analysis.

4 Experimental Results

In this study, gear fault data collected under 10 distinct loading conditions is utilized, employing a "leave-one-out" approach to construct tenfold cross-validation sets by reserving data from one loading condition for testing while using the rest for training and validation. The experiments are conducted on Google Colab with a T4 GPU (15 GB GPU memory), 12.7 GB system RAM, and 112.6 GB disk space. To optimize performance, the study uses a batch size of 16 to balance computational efficiency using the Adam optimizer. Training spans 100 epochs with a fixed stopping criterion and leverages Sparse Categorical Cross entropy as the loss function for binary classification.

The CNN architecture incorporates key components, including Modified Inception, ConvNeXt Block, Receptive Field Attention, SE Block, Grouped Convolutions, and Depthwise Token Mixers, with a dropout rate of 0.5 to mitigate overfitting and ensure optimal performance and generalization. We evaluated a custom CNN model's real-time feasibility and interpretability for gearbox fault detection using inference time analysis, Grad-CAM visualization, and feature map exploration. Inference time was measured over 100 iterations to ensure consistency, while comparing simpler architectures assessed complexity trade-offs. Batch processing and model conversion to TensorFlow Lite is proposed to optimize deployment on edge devices. These techniques ensured efficient evaluation of the model's performance and decision-making process, balancing accuracy with real-time applicability.

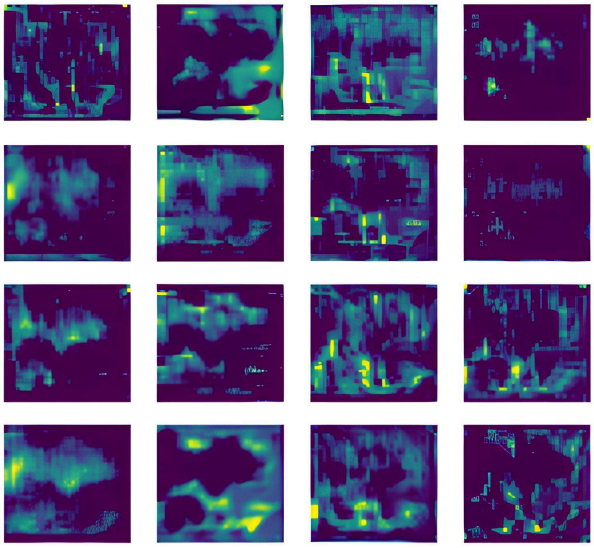

Fig. 7 Feature Map Visualization at Layer 20 for Grad-CAM Analysis

4.1 Model Training and Testing

We conducted our experiments using a dataset of 4,460 spectrograms, evenly distributed between healthy samples (2,240) and broken tooth CQT spectrograms (2,220). The dataset underwent a three-way split, allocating 70% to training, 15% to validation, and 15% to testing. This resulted in 3,776 images for the combined training and validation processes, and 688 images for testing purposes. Classification performance was enhanced through multiple training iterations and hyperparameter adjustments.

The hyperparameter tuning strategy employed in our study is grid search, involving multiple training iterations and hyperparameter adjustments to enhance convergence speed and stability. A dropout rate of 0.5 is selected to mitigate overfitting, considering the dataset size of 4,460 spectrograms, which necessitates dropout. Additionally, the sensitivity of the Adam optimizer to the learning rate further necessitates the use of dropout. The hyperparameters utilized for this study are provided in Table 2.

4.2 Classification performance

To evaluate the classification performance of the gearbox fault detection model, several standard metrics are employed, each offering unique insights into the model's effectiveness. In the study, Neupane et al. focus on fault detection and diagnosis accuracy using metrics like classification accuracy, precision, recall, F-score, and AUC, employing methods such as Traditional Machine Learning, Deep Learning, and signal processing techniques, with statistical preprocessing, feature extraction (FFT, WT, EMD), and performance evaluation, emphasizing the role of explainable AI in building trust in the models [38].

To represent the classification results of the CNN model proposed in this work, we used a confusion matrix. The model correctly classified 308 out of 325 images from the healthy class, while misclassifying 17 CQT spectrogram images as belonging to the broken class. Additionally, out of a total of 331 images from the broken class, 322 were correctly identified as broken, whereas 9 were misclassified as healthy. Figure 8 shows the confusion matrix for gearbox diagnosis, illustrating the classification results of the CNN model proposed in this study. In the study, the model achieved the highest accuracy of 0.97 on CWRU dataset with detailed fault classification, where performance varied across fault types, with inner race and outer race faults often achieving perfect scores, particularly at larger fault sizes, while ball defects showed more variability. Time series segments, containing 2048 points each and covering 0.04 seconds at a 48kHz sampling frequency, were converted into CQT spectrograms.

The models were evaluated on their ability to distinguish the fault types based on precision, recall, and F1-scores across a test set, where inner race and certain outer race faults were consistently identified with high reliability across models, particularly as fault sizes increased, while ball defects showed more variation in classification success. Accuracy measures the proportion of correctly classified instances, reflecting the model's overall reliability and is a fundamental performance indicator. Precision determines the proportion of true positive predictions among all positive predictions, making it crucial for applications where false positives are costly. Recall or Sensitivity

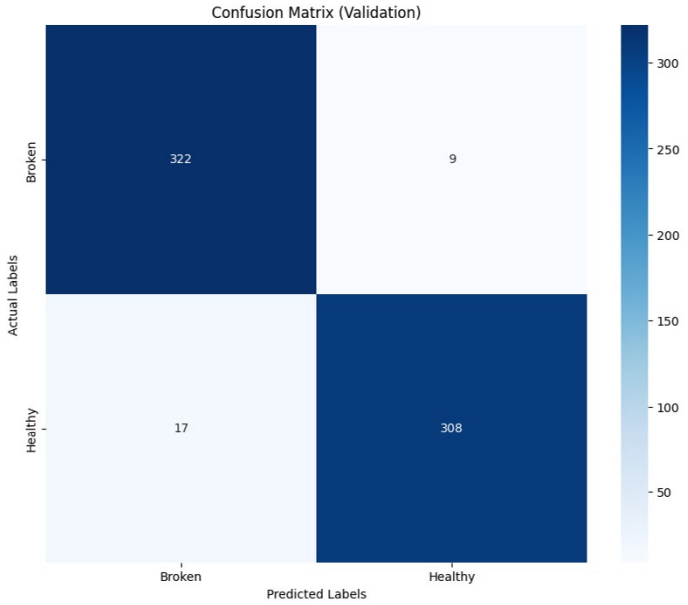

Fig. 8 Confusion Matrix for gearbox diagnosis

assesses the model's ability to correctly identify all actual positive instances, vital for early fault detection to prevent failures. The F1 Score provides a balanced metric by harmonizing precision and recall, particularly useful in imbalanced datasets. Cohen's Kappa Coefficient accounts for chance agreement, offering a robust reliability measure. The Matthews Correlation Coefficient (MCC) evaluates the correlation between actual and predicted classifications, making it suitable for imbalanced datasets. Specificity highlights the model's ability to correctly identify negatives, reducing unnecessary interventions. Negative Predictive Value (NPV) measures the proportion of correct negative predictions, ensuring reliability in identifying nonfaulty cases. To minimize misclassification of faults, the False Negative Rate (FNR) and False Positive Rate (FPR) are essential, as they measure the rates of missed and incorrectly flagged faults, respectively. The False Discovery Rate (FDR) evaluates false positives among predicted positives, ensuring the reliability of positive classifications. Additionally, the Receiver Operating Characteristic (ROC) curve visualizes the trade-off between true positive and false positive rates, with its Area Under the Curve (AUC) quantifying the model's ability to distinguish between classes. Micro and macro-averaged ROC curves summarize performance across multiple classes, providing a holistic evaluation. Figure 9 presents the Receiver Operating Characteristic (ROC) curve for gearbox diagnosis, illustrating the trade-off between true positive and false positive rates, along with the AUC.

The mathematical formulations of these metrics are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

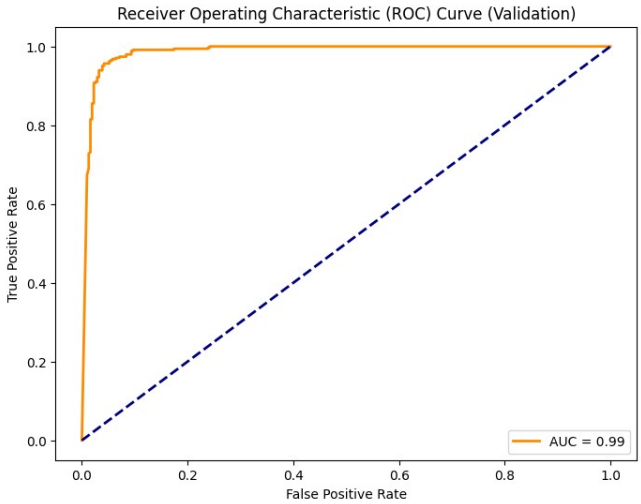

Fig. 9 Receiver Operating Characteristic (ROC) Curve for gearbox diagnosis

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{F1 Score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

$$\text{NPV} = \frac{TN}{TN + FN} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

Where - TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.

4.3 Comparison of the proposed framework with state of-the-art techniques

The proposed method is compared with recent advancements in deep learning and machine learning, as presented in Table 4. It outperforms earlier techniques in fault detection, achieving an impressive accuracy of 96.04%. A key advantage of the proposed method is its simplicity, as it eliminates the need for feature extraction, unlike some techniques in Table 4 that rely on manual feature map extraction alongside deep learning, adding complexity to the process.

Moreover, the proposed system requires minimal training data, enhancing its efficiency and practicality for real-world applications. In Figure 10, the relationship

between different performance metrics and threshold values for Gearbox fault diagnosis is compared.

We compute key metrics including accuracy, precision, recall, F1 score, Matthews correlation coefficient (MCC), specificity, and false positive rate (FPR) for various threshold values ranging from 0 to 1.

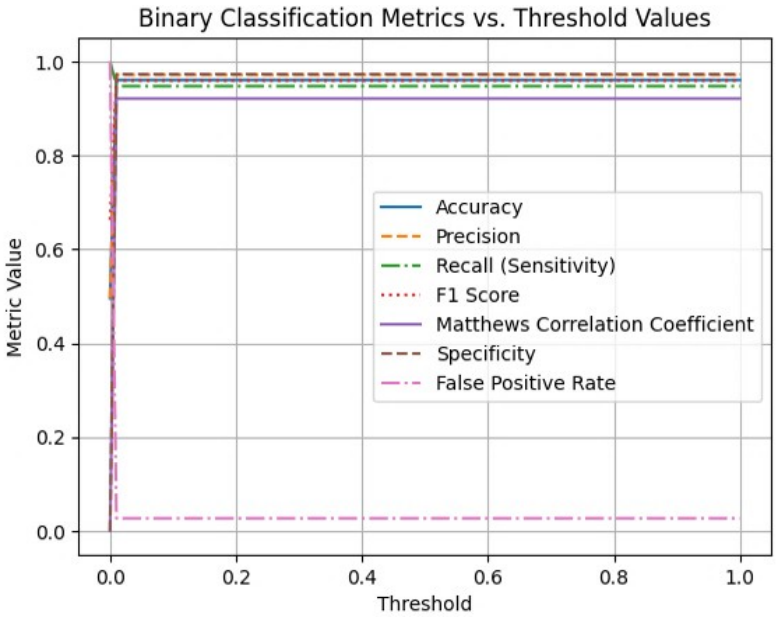

Fig. 10 Comparison of classification metrics vs. threshold values for gearbox diagnosis

By adjusting the decision threshold, we analyze how each metric changes and assist in selecting the optimal threshold for diagnosing faults in a gearbox. This approach provides insights into the trade-offs between different performance measures.

5 Conclusions

This study introduces a fault diagnosis method using a Fusion Convolutional Network to analyze bearing vibration signals. The approach transforms the vibration into Constant-Q Transform (CQT) spectrograms, offering adaptable time-frequency resolution, efficient harmonic analysis, and pitch invariance, making it valuable for fault detection and analysis. Innovative CNN techniques, including the Modified Inception ConvNeXt Block, Receptive Field Attention (RFA), and Squeeze-Aggregated Excitation (SAX) Block, are used in this study to enhance feature extraction, computational efficiency, and multi-scale context modeling, improving performance in fault classification.

The proposed lightweight deep learning model, combined with simplified preprocessing techniques, achieves impressive accuracy (96.03%), precision (97.16%), and

specificity (97.28%) in distinguishing between healthy and broken gearbox conditions. The proposed deep learning model excels in detecting subtle patterns and fault indicators in mechanical systems, achieving remarkable accuracy in diagnosing and predicting gearbox faults, showcasing its potential to revolutionize predictive maintenance in industrial machinery. The results outperform existing methods, showing potential for real-time industrial monitoring, with future improvements focused on developing lightweight models and expanding datasets to include multiple sensor types for early fault detection.

Compliance with Ethical Standards

Disclosure of Potential Conflicts of Interest

The authors declare that they have no known financial interests, personal relationships, or competing non-financial interests that could have appeared to influence the work reported in this paper.

Funding

The authors received no specific funding for this research from any governmental, commercial, or not-for-profit funding agency.

Informed Consent

The data used in this study were obtained from a publicly available, de-identified database. Therefore, no additional informed consent was required for this study.

Data Availability

The dataset analyzed during the current study is publicly available at: `https://data.openei.org/submissions/623`

Code Availability

The source code developed for this study is publicly available at https://github.com/tirimula/EGFNET.

Author Contributions

All authors contributed to the conception and design of the study. Material preparation, data preprocessing, model development, experimentation, and manuscript writing were performed by the authors. All authors read and approved the final manuscript.

References

[1] Z. Xie, J. Chen, Y. Feng, K. Zhang, and Z. Zhou, "End to end multi-task learning with attention for multi-objective fault diagnosis under small sample," *Journal of Manufacturing Systems*, vol. 62, pp. 301–316, 2022.

[2] H. Wu, M. J. Triebe, and J. W. Sutherland, "A transformer-based approach for novel fault detection and fault classification/diagnosis in manufacturing: a rotary system application," *Journal of Manufacturing Systems*, vol. 67, pp. 439–452, 2023.

[3] G. Jiang, C. Jia, S. Nie, X. Wu, Q. He, and P. Xie, "Multiview enhanced fault diagnosis for wind turbine gearbox bearings with fusion of vibration and current signals," *Measurement*, vol. 196, p. 111159, 2022.

[4] S. Faulstich, B. Hahn, and P. J. Tavner, "Wind turbine downtime and its importance for offshore deployment," *Wind Energy*, 2010. https://doi.org/10.1002/we.421

[5] A. Bliznyuk, I. Dadon, R. Klein, and J. Bortman, "Gear diagnostics – Fault type characteristics," Laboratory for Health, Department of Mechanical Engineering, Ben-Gurion University of the Negev.

[6] Z. Chen, C. Li, and R.-V. Sanchez, "Gearbox fault identification and classification with convolutional neural networks," *Shock and Vibration*, vol. 2015, Article 390134, 2015. https://doi.org/10.1155/2015/390134

[7] Y. E. Karabacak and N. G. Özmen, "Common spatial pattern-based feature extraction and worm gear fault detection through vibration and acoustic measurements," *Karadeniz Technical University, Department of Mechanical Engineering.* https://doi.org/10.1016/j.measurement.2021.110366

[8] S. Kumar, D. Goyal, R. K. Dang, S. S. Dhami, and B. S. Pabla, "Condition-based maintenance of bearings and gears for fault detection – A review," *Materials Today: Proceedings*, vol. 5, pp. 6128–6137, 2018.

[9] D. Neupane and J. Seok, "Bearing fault detection and diagnosis using Case Western Reserve University dataset with deep learning approaches: A review," *IEEE Access*, 2020.

[10] T.-D. Nguyen, H.-C. Nguyen, D.-H. Pham, and P.-D. Nguyen, "A distinguished deep learning method for gear fault classification using time-frequency representation," *Discover Applied Sciences*, vol. 6, no. 7, p. 340, 2024. https://doi.org/10.1007/s42452-024-06033-7

[11] J. Kim, J. Kim, and H. Kim, "A study on gear defect detection via frequency analysis based on dnn," *Machines*, vol. 10, no. 8, p. 659, 2022.

[12] S. Patil and K. Wani, "Gear fault detection using noise analysis and machine learning algorithm with yamnet pretrained network," *Materials Today: Proceedings*, vol. 72, pp. 1322–1327, 2023.

[13] K. Ravikumar, A. Yadav, H. Kumar, K. Gangadharan, and A. Narasimhadhan, "Gearbox fault diagnosis based on multi-scale deep residual learning and stacked lstm model," *Measurement*, vol. 186, p. 110099, 2021.

[14] P. Hu, C. Zhao, J. Huang, and T. Song, "Intelligent and small samples gear fault detection based on wavelet analysis and improved cnn," *Processes*, vol. 11, no. 10, p. 2969, 2023.

[15] I. Ahmed, M. Ahmad, A. Chehri, and G. Jeon, "A smart-anomaly-detection system for industrial machines based on feature autoencoder and deep learning," *Micromachines*, vol. 14, no. 1, p. 154, 2023.

[16] P. Dutta, K. K. Podder, M. S. I. Sumon, M. E. H. Chowdhury, A. Khandakar, et al., "Gearfaultnet: Novel network for automatic and early detection of gearbox faults," *IEEE Access*, vol. 12, pp. 188755–188765, 2024.

[17] K. Cheng, L. Cheng, T. Mao, and M. Wu, "Fault diagnosis of gearbox based on cnn and lstm with attention mechanism," *Highlights in Science, Engineering and Technology*, vol. 93, pp. 285–294, 2024.

[18] Y. Yao, S. Zhang, S. Yang, and G. Gui, "Learning attention representation with a multi-scale cnn for gear fault diagnosis under different working conditions," *Sensors*, vol. 20, no. 4, p. 1233, 2020.

[19] R. Magar, L. Ghule, J. Li, Y. Zhao, and A. B. Farimani, "Faultnet: A deep convolutional neural network for bearing fault classification," *IEEE Access*, vol. 9, pp. 25189–25193, 2021.

[20] Y. Bai, W. Cheng, W. Wen, Y. Liu, and M. A. Lopez, "Application of time-frequency analysis in rotating machinery fault diagnosis," *Shock and Vibration*, vol. 2023, pp. 1–16, 2023.

[21] P. Hu, C. Zhao, J. Huang, and T. Song, "Intelligent and small samples gear fault detection based on wavelet analysis and improved cnn," *Processes*, vol. 11, no. 10, p. 2969, 2023.

[22] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2698–2701, 1992.

[23] L. Bahatti, M. Zazoui, O. Bouattane, and A. Rebbani, "Short-term sinusoidal modeling of an oriental music signal by using CQT transform," *Journal of Signal and Information Processing*, vol. 4, pp. 51–56, 2013.

[24] M. Nergiz, "Classification of precancerous colorectal lesions via ConvNeXt on histopathological images," *Balkan Journal of Electrical & Computer Engineering*, vol. 11, no. 2, 2023.

[25] W. Yu, P. Zhou, S. Yan, and X. Wang, "InceptionNeXt: When Inception meets ConvNeXt," *arXiv preprint arXiv:2303.16900v2*, 2024.

[26] O. El Gannour, S. Hamida, B. Cherradi, M. Al-Sarem, A. Raihani, F. Saeed, and M. Hadwan, "Concatenation of pre-trained convolutional neural networks for enhanced COVID-19 screening using transfer learning technique," *Electronics*, vol. 11, p. 103, 2022.

[27] M. Tan, X. Yuan, B. Liang, and S. Han, "DRFnet: Dynamic receptive field network for object detection and image recognition," *Frontiers in Neurorobotics*, vol. 16, p. 1100697, 2023.

[28] L. Zhou, H. Cai, J. Gu, Z. Li, Y. Liu, X. Chen, Y. Qiao, and C. Dong, "Efficient image super-resolution using vast-receptive-field attention," *arXiv preprint arXiv:2210.05960*, 2022.

[29] Y. Yao, S. Zhang, S. Yang, and G. Gui, "Learning attention representation with a multi-scale CNN for gear fault diagnosis under different working conditions," *Sensors*, vol. 20, no. 1233, 2020.

[30] N. Mahendran, "Variations of squeeze and excitation networks," *arXiv preprint arXiv:2304.06502*, 2023.

[31] N. Mahendran, "Squeeze aggregated excitation network," *arXiv preprint arXiv:2308.13343*, 2023.

[32] J. Lo, S. Nithiyanantham, J. Cardinell, D. Young, S. Cho, A. Kirubarajan, M. W. Wagner, R. Azma, S. Miller, and M. Seed, et al., "Cross attention squeeze excitation network (CASE-Net) for whole body fetal MRI segmentation," *Sensors*, vol. 21, p. 4490, 2021.

[33] A. N. Abdullah and T. Aydin, "Ninformer: A network in network transformer with token mixing as a gating function generator," *arXiv preprint arXiv:2403.02411v5*, 2024.

[34] Z. Lu, L. Kang, and J. Huang, "Depthwise convolution with channel mixer: Rethinking mlp in metaformer for faster and more accurate vehicle detection," in *Lecture Notes in Computer Science (LNCS)*, vol. 14263, 2023.

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[36] A. Harikumar, S. Surendran, and S. Gargi, "Explainable AI in deep learning based classification of fetal ultrasound image planes," *Procedia Computer Science*, vol. 233, pp. 1023–1033, 2024.

[37] M. Lerma and M. Lucas, "Grad-CAM++ is equivalent to Grad-CAM with positive gradients," in *24th Irish Machine Vision and Image Processing Conference*, pp. 113–120, 2022.

[38] D. Neupane, M. R. Bouadjenek, R. Dazeley, and S. Aryal, "Data-driven machinery fault detection: A comprehensive review," *arXiv preprint arXiv:2405.18843*, 2024.

Table 1 Recent Literature on Approaches to the Automatic Diagnosis of Gearbox Faults

Sno.	Author(s)	Year	Feature Extraction Technique/Method	Accuracy (%)	Remarks/Limitations
1	Dutta et al. [16]	2024	GearFaultNet (1D CNN with z-score and min-max normalization)	94.04	Lacks broader fault detection capabilities.
2	Ahmed et al. [15]	2023	6-layer Autoencoder-based deep learning model with denoising techniques	91	Limited by dataset reliance, sensor quality preprocessing needs.
3	Hu et al. [14]	2023	Wavelet decomposition, Markov transition fields, and VGG-based CNN with data augmentation	88.33	Sensitivity to wavelet decomposition levels
4	Kim et al. [11]	2022	Frequency band sums with DNN for real-time diagnosis	95.5	Limited to frequency band reliance.
5	Patil et al. [12]	2022	Transfer learning with YAMNet and Mel Spectrograms	95	Small dataset; focuses on a single fault typ
6	Ravikumar et al. [13]	2021	Multi-scale CNN with residual learning and stacked LSTM model (MDRL-SLSTM)	94.33	High computational demands make these m costly and resource-intensive.
7	Yao et al. [18]	2020	Multi-scale CNN with attention mechanism for gear fault detection	93.3	Potential overfitting and performance varia in noisy environments.

Table 2 Training Parameters

Parameter	Value/Method
Batch Size	16
Number of Epochs	100
Learning Rate	0.001
Optimizer	Adam
Loss Function	Sparse Categorical Cross Entropy
Dropout Rate	0.5
Activation Function (Output)	Softmax

Table 3 Detailed Performance Metrics of the Model

Sl No	Metric	Value
1	Accuracy	96.04%
2	Precision	96.06%
3	Recall (Sensitivity)	96.04%
4	F1 Score	96.04%
5	Kappa Coefficient	92.07%
6	Matthews Correlation Coefficient	92.10%
7	Specificity	97.28%
8	Negative Predictive Value	94.99%
9	False Negative Rate	5.23%
10	False Discovery Rate	2.84%
11	False Positive Rate	2.72%

Table 4 Performance comparison of the proposed method with literature

Literature	Feature Extraction	Classifier	Accuracy	Precision	Sensitivity	F1-score	Specificity
Dutta et al., 2024 [16]	z-score and min-max normalization	GearFaultNet (1D CNN)	94.06	94.11	94.06	94.05	94.01
Ahmed et al., 2023 [15]	Denoising techniques	6-layer Autoencoder	91	98	83	–	–
Ravikumar et al., 2021 [13]	Multi-scale features	MDRL-SLSTM	94.33	93.11	92.06	95.05	92.01
Proposed Method	Modified Inception ConvNeXt with Receptive Field Attention	Fusion CNN	**96.03**	**97.16**	**97.28**	94.76	–