COMPARATIVE STUDY OF NEURAL NETWORKS FOR COMPOUND MACHINE FAULT DIAGNOSIS

Sajeev Senthil, Siva Prasanth Sivaraj, Suganth k, Tharunkumar S

Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India

ABSTRACT

Industrial equipment often suffers from simultaneous failures in multiple components, posing an enormous challenge for conventional fault diagnosis systems that emphasize single-component faults. In rotating equipment with deep groove ball bearings, combined faults in the form of concurrent bearing damage and misalignment, unbalance, or looseness of rotating components produce complex vibration patterns that are hard to differentiate and classify correctly. This research tackles this challenge through a systematic comparative analysis of three neural network architectures for compound machine fault diagnosis: Audio Spectrogram Transformer with Feed-Forward Neural Network (AST+FFNN), Convolutional LSTM Deep Neural Network (CLDNN), and a Multi-Branch Feature Fusion Network integrating learned and manually engineered features. With a wide multi-domain vibration dataset from the University of Seoul that has 32 fault categories with 21 complex fault scenarios under different operating conditions, we compared the ability of each design to classify complex fault patterns accurately. AST+FFNN used transfer learning from audio spectrograms to draw 768-dimensional learned representations, while Multi Branch incorporated AST features with 50 handcrafted time-frequency domain features and achieved a classification accuracy of 83.12%. CLDNN architecture, by merging convolutional layers for spatial feature learning with LSTM networks for temporal dependency modelling, attained peak performance at 90.91% accuracy. Experiments show that hybrid temporal-spatial architectures can well capture both sequential dynamics and local patterns of vibration signals and outperform pure feature-based or transfer learning-based methods. The research sets performance standards for compound fault diagnostic systems and confirms the efficacy of CLDNN designs in industrial predictive maintenance scenarios, providing real-world operational insights to enable precise classification of intricate compound fault situations to avoid catastrophic machine breakdowns.

1. INTRODUCTION

a. Motivation of the study

Rotating machine reliability is directly responsible for production efficiency, safety, and cost of operation within industrial settings. Deep groove ball bearings, found everywhere in motors, pumps, and turbines, seldom fail independently. Actual machinery suffers compound faults when bearing flaws are accompanied by rotating component failures like shaft misalignment, unbalance of the rotor, or looseness of mounting. Conventional vibration analysis methods are challenged by these situations since compound fault signatures have overlapping frequency components and nonlinear interactions that obscure individual fault features. New deep learning advances hold promises via automated feature extraction and pattern recognition, but the research community is lacking systematic studies comparing various neural network architectures tailored to compound fault situations. Knowing which design strategies most effectively balance accuracy, computational cost, and interpretability is critical in building feasible diagnostic systems that maintenance engineers will confidently deploy in production settings where false negatives are very costly.

b. Problem statement

Existing bearing fault diagnosis work mainly focuses on single-component faults under simulated laboratory settings, leaving a crucial gap in the treatment of compound fault cases that represent real-world industrial operating conditions. The difficulty is in properly classifying 32 different fault modes such as healthy operation, three types of bearing faults, seven types of rotating component faults with different severity levels, and 21 combinations of compound faults at different operating speeds and sampling rates. Current methods are either based solely on end-to-end deep learning with a loss of interpretability or solely on handcrafted features that potentially fail to capture subtle patterns. No complete research compares transformer-based models, recurrent-convolutional hybrids, and feature fusion methods for the particular problem here. Also, domain shift due to different working conditions makes model generalization challenging. The underlying question still is: what neural network architecture gives the best performance for compound fault diagnosis and still gives the explainability necessary for industrial acceptance and regulatory compliance?

c. Plan of action

In this paper, we employ and compare three distinct architectures of neural networks in order to identify the optimal techniques for compound bearing fault diagnosis. Firstly, we built an AST+FFNN model that converts vibration signals into mel-spectrograms and leverages a pre-trained Audio Spectrogram Transformer encoder to achieve 768-dimensional learned features, which are subsequently classified by a three-layer feed-forward neural network. This approach leverages transfer learning from the audio classification problem to train spectral-temporal patterns in vibration signals. Second, we employed a CLDNN architecture that combines one-dimensional convolutional layers for local feature extraction with stacked LSTM layers for temporal dependency modeling and dense classification layers. This hybrid model learns

specifically both spatial patterns and sequence dynamics in raw vibration signals. Third, we introduced a Multi-Branch Feature Fusion Network that integrates the 768 AST-learned features and 50 manually constructed features in time-domain statistics, frequency-domain spectral features, and time-frequency envelope features. To improve model interpretability and feature selection, an attention mechanism was introduced following the fusion layer, dynamically assigning importance weights to every one of the 818 concatenated features, allowing the network to pay attention to the most diagnostic features for every fault condition. The three models were trained and tested on the University of Seoul multi-domain bearing dataset with 384 files on 32 fault classes with varied operating conditions. Categorical cross-entropy loss, ReLU activation, Softmax output, and Adam optimization were employed. Comparison of performance concentrated on classification accuracy, confusion matrix analysis, computational efficiency, and attention-based feature importance analysis to determine comprehensive benchmarks for industrial compound fault diagnosis applications.

2. LITERATURE REVIEW

S. No.	Referred paper	Summary of the work	Relevance to your problem statement	Research Gap
1	Kim, B.J., et al. (2025). An explainable and accurate transformer-based deep learning model for audio classification. Scientific Reports, Nature	Develops a pure Audio Spectrogram Transformer (AST) for clinical breath sound classification, achieving superior accuracy over CNNs, and keeping full-length/context in input. Highlights explainability and lack of dimension reduction issues.	Validates the AST architecture for audio classification; supports use of transformers for sequence/context and model explainability.	Notes need for real-time implementations and more interpretability of internal attention weights for clinical/industrial adoption.
2	Hakim, M., et al. (2023). A systematic review of rolling bearing fault diagnoses using deep learning. Measurement	Comprehensive review of deep learning methods for bearing fault diagnosis; analyzes CNN, RNN, hybrid models, and feature extraction/spectrogram approaches. Standard architectures and datasets surveyed, with accuracy comparison.	Summarizes baseline state-of-the-art approaches for bearing diagnosis using spectrograms, provides context for AST methods.	Notes lack of interpretability, overfitting on small datasets, and limited use of transformer/attention mechanisms for global feature learning.

3	Ren, H., et al. (2024). A novel intelligent fault diagnosis method of bearing based on multi-head self-attention convolutional neural network (MSA-CNN). AI EDAM, Cambridge University Press	Proposes an MSA-CNN with multi-head self-attention to better aggregate global information and dynamically weigh features in bearing fault diagnosis. Converts raw signals into 2D grayscale images for CNN input; outperforms standard CNN, especially under noise.	Shows the power of attention with CNNs for enhancing feature extraction and global context learning; aligns with your use of AST/attention blocks.	Suggests further work is needed in using pure transformer architectures and in explainability of attention weights.
4	Zhang, Q., et al. (2024). Convolutional Neural Network with Attention Mechanism for bearing fault diagnosis. PMC Open Access	Introduces CBAM-CNN (CNN with channel attention) for bearing fault diagnosis from vibration image samples. Attention improves feature selection and classification accuracy versus plain CNNs.	Directly establishes the benefit of adding attention to CNN-based feature extraction for better diagnosis, echoing your AST+attention approach.	Limited to CNN-attention hybrids; mentions the need for exploring pure attention/Transforme r models on spectrogram data.
5	Mian, T., et al. (2023). Mel-spectrogram based Approach for Fault Detection in Ball Bearing using Convolutional Neural Network. ACM Digital Library	Uses Mel-spectrograms of sound signals as input to CNNs for bearing fault diagnosis. Achieves >97% accuracy; robust across speeds and loads. Uses audio (not vibration) for fault detection.	Demonstrates the value of spectrogram-based deep features and CNNs for robust intelligent bearing diagnosis provides contrast for AST and FFNN methods.	CNNs limit global context modeling; doesn't address attention or transformer-based spectrogram feature learning.
6	Gengchen Ma, Xihe Qiu, Xiaoyu Tan, et al. (2025), "DMFusion: A dual-branch multi-scale	Proposes DMFusion, a dual-branch autoencoder network for multi-modal	Demonstrates the effectiveness of dual/multi-branch fusion to preserve diverse information sources, analogous to fusing manual and neural	Lacks evaluation for non-imaging signals and highly noisy/compound fault domains.

	feature fusion network for medical multi-modal image fusion," Biomedical Signal Processing and Control,	medical image fusion that captures both common and unique modality features via multi-scale architecture. Demonstrates superior retention of detail and high reconstruction quality across tasks.	features in bearing fault diagnosis.	
7	Xinchen Zhang, Hao Zhu, Xiaotong Li, et al. (2025), "Recurrent Progressive Fusion-based Learning for Multi-Source Remote Sensing Image Classification, " Pattern Recognition	Introduces recurrent progressive fusion using reinforcement learning for multi-source feature integration, allowing dynamic refinement of dual-source features with an adaptive, multi-step fusion pathway for remote sensing images.	Shows that recurrent or progressive feature fusion can outperform static fusion, providing insight for using sequential or attention-driven refinement in industrial diagnostics.	Experiments limited to image domains, not sequence or sensor-based compound tasks.
8	Nature (2025), "Multi-branch convolutional neural network with cross-attention mechanism for emotion recognition," Nature	Presents a multi-branch CNN framework using cross-attention for EEG and multimodal emotion recognition. Demonstrates efficient	Validates the value of cross-branch attention mechanisms in unifying handcrafted and deep features, with strong generalization to multi-	Mainly addresses multimodal emotion classification; lacks fault- or machinery-centric benchmarks.

	Scientific Reports,	integration of independent feature streams and robust performance compared to single-branch or naive fusion methods.	source classification tasks like fault diagnosis.	
9	Liu Y., Chen Z., Wang J., et al. (2024), "Multi-branch fusion graph neural network based on attention for epileptic seizure detection," Frontiers in Physiology	Proposes a graph neural network with multi-branch architecture and multi-head attention, efficiently learning spatial-temporal features for EEG-based seizure detection. Outperforms benchmarks on patient-specific and patient-independent tasks.	Illustrates real diagnostic value of multi-branch and attention-fusion models in medical/signal domains, directly relevant to attention-driven compound fault diagnosis.	Focuses on brain/EEG data rather than compound industrial/mechanical signals.
10	Han B., Zhang Y., Fu S., et al. (2024), "A multibranch and multiscale neural network based on unsupervised segmentation for medical image fusion," Scientific Reports	Proposes DUSMIF, a multi-branch, multi-scale network augmented by unsupervised segmentation and attention mechanisms, for semantic-rich image fusion.	Offers a template for integrating semantic or segmented feature streams into multi-branch models, helpful for complex signal fusion in machinery fault diagnosis.	Semantic fusion is tailored to medical imaging, not to raw vibration or mixed industrial signals.

		Achieves superior content and structure preservation.		
11	Rolling Bearing Fault Diagnosis Based on VMD-DWT and HADS-CNN-BiLSTM Hybrid Model (Shao, Zhao & Kang, 2025)	Proposes a hybrid framework that first denoises vibration signals by combining Variational Mode Decomposition (VMD) and Discrete Wavelet Transform (DWT), then feeds into a **depthwise separable-CNN + BiLSTM** architecture augmented with triple attention mechanisms for fault classification. The approach attains very high accuracy on CWRU and XJTU datasets.	Very close to your dual-branch + attention idea: they fuse denoised signal channels and use CNN + LSTM with attention to get robust features. Their use of triple attention suggests multi-level attention is beneficial.	Their denoising stage (VMD + DWT) depends on parameter tuning; computational overhead is high. They do not deeply analyze interpretability of attention weights or test on cross-domain (unseen operating conditions) data.
12	Transformer network enhanced by dual convolutional neural networks (Trans-DCC) (Frontiers, 2025)	Proposes a "Trans-DCC" framework combining dual CNN (time-domain and frequency-domain) and cross-attention among them to feed into a lighter transformer module. The model reduces attention complexity and boosts fault detection across variable speeds.	Matches your interest in combining CNN + attention + transformer: dual-domain CNN front ends feeding a cross-attention/transformer backbone is analogous to your dual-branch spectrogram + signal approach.	Interpretability of cross-attention is not elaborated. The model is tested primarily on train wheelset bearings; generalization to many domains is less explored.
13	A Hybrid Deep Learning Approach for Bearing Fault Diagnosis Using CWT and Attention-Enhanced Spatiotemporal Feature	Integrates Continuous Wavelet Transform to produce time-frequency maps, then applies a hybrid network combining **multi-head self-attention**, BiLSTM, and 1D residual CNN to capture both spatial	Aligns strongly: they fuse time-frequency and sequence features with attention + CNN + LSTM. It demonstrates that combining multiple attention types in a hybrid model yields good robustness.	They don't deeply discuss real-time constraints or interpretability of attention modules. Also, they don't perform ablation on how many attention layers are needed in practice.

	Extraction (Siddique et al., 2025)	and temporal dependencies. Validated across CWRU and Paderborn datasets.		
14	A Bearing Fault Diagnosis Method Based on Fusion of CNN-BiLSTM-Transformer and Cross-Attention (Yuan & Wu, recent)	The method preprocesses signals with FFT + VMD to generate multi-scale features. It then uses a CNN + Transformer branch and a BiLSTM branch, finally fusing via **cross-attention** between frequency/time features. Reports very high recognition rates (> 99 %).	Very close to the parallel architecture: dual-branch (time + freq), cross-attention fusion, hybrid CNN/LSTM/Transformer design. Useful as a comparative architecture.	They do not deeply discuss how attention weights map to physical insight, and scalability to more channels (multi-sensor) is not tested.
15	TDANet: A Novel Temporal Denoise Convolutional Neural Network With Attention for Fault Diagnosis (Li et al., arXiv 2024)	Proposes TDANet, which transforms 1D signals into 2D periodic-based tensors, uses multi-scale 2D convolutions and adds a **Temporal Variable Denoise (TVD)** module + **Multi-head Attention Fusion (MAF)** to emphasize salient features under high noise. Evaluated on CWRU and other noisy datasets.	This is relevant since in practice your signals will have noise; their attention + denoising design could be adapted in your pipeline (e.g. before fusion). The idea of converting 1D to 2D periodic structure might help spectrogram representation choices.	They focus primarily on noise robustness; less on cross-domain generalization, interpretability of attention modules, or combining it with LSTM/transformer for long-term dependencies.

3. METHODOLOGY

a) Dataset Overview:

Dataset Source:

The experimental data used in this research comes from the "Multi-domain Vibration Dataset with Various Bearing Types under Compound Machine Fault Scenarios," released in December 2024 by Data in Brief (Volume 57, Article 110940, DOI: 10.1016/j.dib.2024.110940) by Seongjae Lee, Taewan Kim, and Taehyoun Kim of the Department of Mechanical and Information Engineering, University of Seoul, South Korea. This publicly accessible dataset, available through the Mendeley Data repository, was specifically designed to address limitations in existing bearing fault databases by incorporating realistic compound fault scenarios and multiple domain variations that reflect actual industrial operating environments, making it an ideal benchmark for advanced fault diagnosis algorithm development.

Experimental Setup:

Vibration measurements were obtained from a custom bearing fault simulator test rig located at the University of Seoul, which is fitted with a PCB Piezotronics 333D01 USB digital accelerometer capable of highly accurate vibration measurement. The test platform facilitates systematic introduction of controlled faulty conditions in both bearing parts and rotating machinery assemblies while keeping the operational parameters such as rotating speed and sampling frequency accurately regulated. The total dataset consists of three bearing geometries (tapered roller bearings, cylindrical roller bearings, and deep groove ball bearings), and in this research, we only use the deep groove ball bearing type 6204 subset, which offers detailed coverage of all 32 fault classes under 12 multi-domain operating conditions, providing 384 total vibration signal files with 1,280,000 data points for each signal.

Multi-Domain Operating Conditions:

The data set includes systematic domain variations over two key axes to replicate true-world industrial variation. Vibration signals are recorded at two different sampling rates: 8 kHz (160 seconds length) and 16 kHz (80 seconds length), both resulting in equal signal lengths of 1,280,000 points. Six rotating speeds cover common industrial equipment operation: 600 RPM, 800 RPM, 1000 RPM, 1200 RPM, 1400 RPM, and 1600 RPM. The combination of 6 rotating speeds and 2 sampling rates results in 12 different domain configurations for each fault condition, providing a complete multi-domain structure that represents the cross-domain issues arising when implementing diagnostic systems across multiple installations, sensor setups, or different operating regimes in industrial applications.

Fault Classification Taxonomy:

The dataset encompasses 32 fault classes organized hierarchically:

- **Healthy State (1 class):** Normal operation baseline (H)

- **Single Bearing Faults (3 classes):** Ball fault (B), Inner raceway fault (IR), Outer raceway fault (OR)
- **Single Rotating Component Faults (7 classes):**
    - Looseness (L)
    - Unbalance at three severity levels: U1 (3g mass), U2 (4g mass), U3 (5g mass)
    - Misalignment at three severity levels: M1 (0.6mm shift), M2 (0.8mm shift), M3 (1.0mm shift)
- **Compound Faults (21 classes):** Combinations of bearing faults with rotating component faults
    - Ball compound faults (7 classes): B_L, B_U1, B_U2, B_U3, B_M1, B_M2, B_M3
    - Inner raceway compound faults (7 classes): IR_L, IR_U1, IR_U2, IR_U3, IR_M1, IR_M2, IR_M3
    - Outer raceway compound faults (7 classes): OR_L, OR_U1, OR_U2, OR_U3, OR_M1, OR_M2, OR_M3

b) Correlation Analysis:

Correlation analysis identified high redundancy between both manual and AST feature sets. Among manual and AST features, 5,422 major correlations (|r|>0.5) were identified, with some manual features (such as Manual) correlating with multiple AST dimensions at nearly perfect level, indicating redundant information. In the manual set, 113 feature pairs were highly correlated (|r|>0.8), and the AST features sample had 335 such pairs, reflecting that numerous features are essentially reporting the same thing, which can be minimized through feature selection. The most discriminative manual features, quantified by Fisher score, were determined for improving class separation.

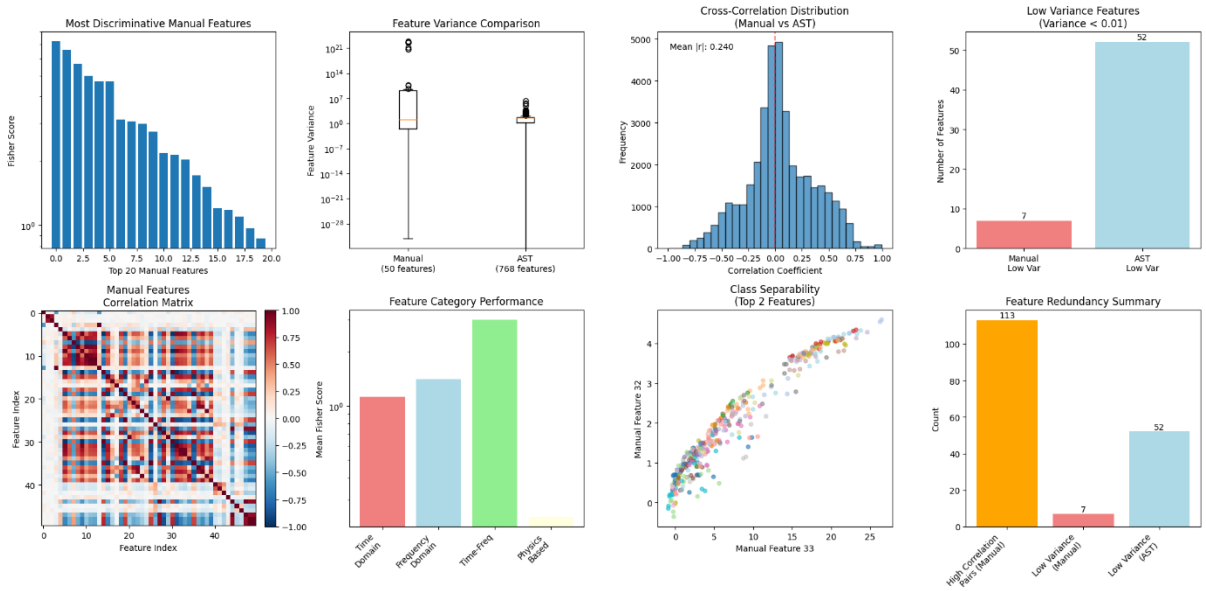

Fig.1.Correlation analysis of AST features

Key Formulas for Correlation and Identical Feature Analysis

1. **Pearson Correlation Coefficient:**

$$r_{XY} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

Measures linear relationship (correlation) between two feature vectors X and Y; r ranges from -1 (perfect negative) to +1 (perfect positive).

2. **Feature Variance:**

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Quantifies the spread of a feature X across all samples; near-zero variance points to near-constant or non-informative features.

3. **Maximum Absolute Difference for Identical Feature Detection:**

$$\mathrm{max\_diff}_{j,k} = \max_i \; | F_{i,j} - F_{i,k} |$$

Used to check if two features j, k are identical across all samples: if max_diff $< \epsilon$ (numerical tolerance), the features are considered identical.

c) Identical and Redundant Feature Analysis

The visual summaries underscore that some manual and AST features are absolutely identical or consistent: 1 manual feature was duplicated to a tee, 2 were consistent, and 31 AST features were consistently the same between samples. Variance and zero-value analysis revealed 7 manual and 52 AST features as insignificant in variation, with 3 ASTs being all-zero. Taken collectively (see the uploaded results), these results led to pruning of the redundant and uninformative features, enhancing the efficiency and interpretability of the model fusion, while highlighting the distinctive contribution of high-variance, discriminative features.

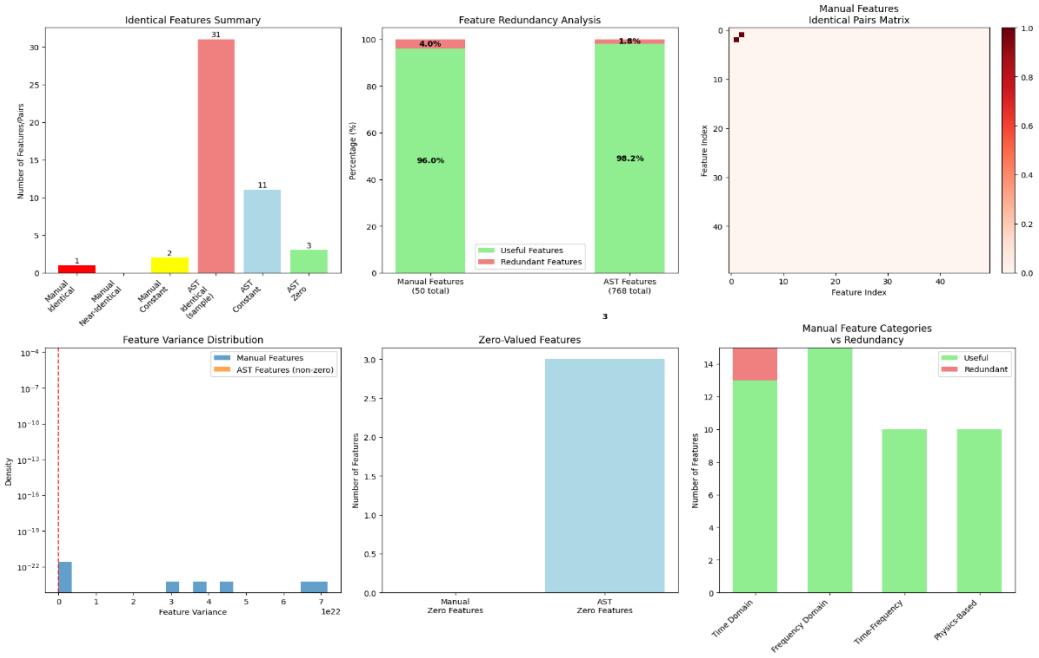

Fig.2.Correlation analysis of AST features

d) Model-1: AST Feature Extraction + Feedforward Neural Network (FFNN)

1. Data Preprocessing

Vibration Signal Preparation:

Raw vibration signals are normalized to eliminate amplitude scale effects. Segments are checked for outliers, missing data, and non-conforming samples are discarded to ensure data integrity.

2. AST Feature Extraction

Vibration signals are transformed into mel-spectrograms as a time-frequency representation, which are then used as "images" for neural feature extraction. The Audio Spectrogram Transformer (AST) leverages a Vision Transformer (ViT) backbone pretrained on large-scale audio data.

Stepwise AST Feature Computation:

- Short-Time Fourier Transform (STFT):

$$X(t,f) = \sum_{n} x[n]w[n-t]e^{-j2\pi fn}$$

where $x[n]$ is the time signal, w is the window function, t is the frame index, and f is frequency.

- Mel-Scale Filter Bank:

$$m = 2595 \cdot \log_{10}(1 + \frac{f}{700})$$

Linear scale frequencies are mapped to mel scale for perceptual relevance.

- Mel-Spectrogram Computation:

$$S(m,t) = \text{MelFilterBank}_m \cdot |\, X(t,f)\,|^2$$

Each row of the spectrogram represents the power in a mel frequency band at a time frame.

- Patch Embedding for Transformer:
  Spectrograms are sliced into non-overlapping patches (e.g., 16x16).
  Each patch p is flattened and projected:

$$z_p = W_p\, p + b_p$$

where W_p and b_p are learnable weight and bias.

- Positional Encoding:
  Adds sequential context to patch embeddings to retain the time-frequency structure.
- Self-Attention in Transformer Encoder:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

where queries Q, keys K, and values V are linear projections of patch embeddings, and d_k is the key dimension.
Multi-head attention allows the transformer to focus on various time-frequency patterns simultaneously.

- AST Feature Vector:
  The output of a special classification token after the transformer encoder is a 768-dimensional feature vector representing the spectral-temporal signature of each vibration sample.

3. Feature Redundancy and Selection
Pairwise correlation, zero-variance, and all-zero checks are performed to prune the AST feature set, ensuring only non-redundant, informative features are fed to the classifier.

4. Architecture:

The AST feature vector (after pruning, typically of dimension 768) is input to the FFNN as shown in the figure.1.

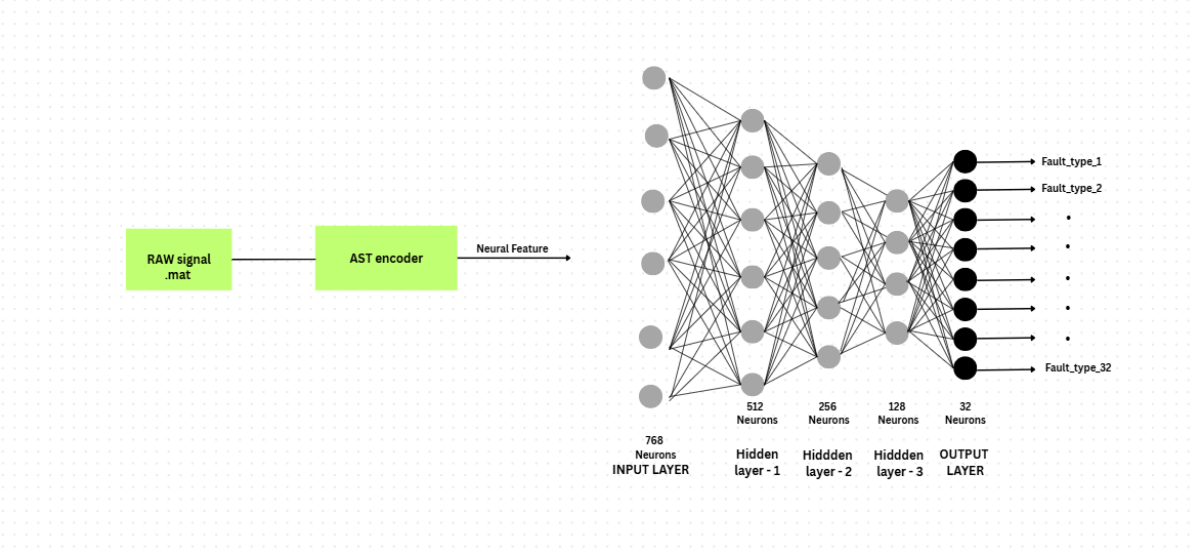

Fig.3.AST+FFNN Architecture

- **Input Layer:**

$$X_{\text{in}} \in \mathbb{R}^{768}$$

- **First Dense Layer:**

$$h_1 = \text{ReLU}(W_1 X_{\text{in}} + b_1) \in \mathbb{R}^{256}$$

- **Batch Normalization, Dropout:**
  (Applied after each dense layer for stability and regularization)

- **Second Dense Layer:**

$$h_2 = \text{ReLU}(W_2 h_1 + b_2) \in \mathbb{R}^{128}$$

- **Third Dense Layer:**

$$h_3 = \text{ReLU}(W_3 h_2 + b_3) \in \mathbb{R}^{64}$$

- **Output Layer (Softmax):**

$$\hat{y} = \text{softmax}(W_4 h_3 + b_4) \in \mathbb{R}^{C}$$

where C is the number of fault classes (e.g., $C = 32$).

Notation:
- W_i and b_i = weights and biases of the i-th dense layer
- ReLU$(\cdot)$ = Rectified Linear Unit activation function
- softmax$(\cdot)$ = Computes class probabilities

5. Training Protocol

- **Loss Function:**
  Categorical cross-entropy:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

- **Optimizer:**
  Adam optimizer with initial learning rate 0.001.

- **Regularization:**
  Dropout (0.3–0.5), early stopping (patience: 10), and learning-rate reduction on plateau are used to prevent overfitting.

- **Validation:**
  Dataset is split into train/val/test sets (e.g., 70/20/10%). Model selection is based on the best validation accuracy, while final performance is reported on the unseen test set using metrics such as accuracy, precision, recall, F1-score, and confusion matrix

e) Model-2: Multibranch Features Fusion Network

The suggested Multi-Branch Feature Fusion Network combines learned neural representations with domain-expert hand-crafted features to utilize both automated pattern finding and physics-based fault indicators for compound bearing fault diagnosis. The architecture consists of two side-by-side branches that process the same raw vibration signals along with distinct feature extraction pathways: the first branch uses pre-trained Audio Spectrogram Transformer (AST) encoder to learn 768-dimensional features from mel-spectrograms, whereas the second branch calculates 50 hand-crafted features in time-domain, frequency-domain, and time-frequency domains according to well-known principles of vibration analysis. These orthogonal sets of features are concatenated together to create an 818-dimensional vector that is fed through a smart fusion layer with attention mechanism that assigns dynamically important weights to each feature dimension, allowing the network to pay attention to the most diagnostic features for every fault condition. The attended features are progressively compressed by three hidden layers (512→256→128→32 neurons) before final SoftMax classification over 32 fault classes.

Data Preprocessing:

1) Manual Feature Engineering

The manual feature extraction path calculates 50 manually designed features from decades of vibration analysis knowledge and documented mechanical engineering practices for bearing fault detection. These features are designed in a systematic manner to identify unique fault signatures in three complementary domains: time-domain features measure amplitude-based attributes like signal energy, shape of distribution, and impulsiveness through statistical moments and shape factors; frequency-domain features examine spectral content to determine characteristic fault frequencies and patterns of energy distribution related to bearing defects and faults in rotating components; and time-frequency features use envelope analysis and wavelet decomposition to detect amplitude modulation patterns and transient activity that are signatures of bearing impacts and interactions of compound faults. This cross-domain strategy guarantees thorough coverage of fault-related information with the accomplishment of computational efficiency and interpretability for industrial adoption.

Feature List

Time-Domain Features (20 features):

- Mean, RMS (Root Mean Square), Standard Deviation, Variance
- Skewness, Kurtosis, 5th Moment, 6th Moment
- Peak Value, Peak-to-Peak Amplitude
- Crest Factor, Shape Factor, Impulse Factor, Clearance Factor
- Zero Crossing Rate, Mean Absolute Deviation, Coefficient of Variation
- Interquartile Range, Mean Absolute Value, Time-Domain Entropy

Frequency-Domain Features (20 features):

- Spectral Centroid, Spectral Spread, Spectral Rolloff, Spectral Flatness
- Low-Frequency Energy (0-1000 Hz), Mid-Frequency Energy (1000-3000 Hz), High-Frequency Energy (>3000 Hz)
- Peak Frequency, Peak Amplitude
- Harmonic-to-Noise Ratio, Total Harmonic Distortion
- Spectral Entropy, Spectral Kurtosis, Spectral Skewness, Spectral Variance
- Frequency Band Ratio, Frequency Center of Gravity
- BPFO Energy (Ball Pass Frequency Outer race), BPFI Energy (Ball Pass Frequency Inner race), BSF Energy (Ball Spin Frequency)

Time-Frequency Features (10 features):

- Envelope RMS, Envelope Peak Frequency, Envelope Kurtosis, Envelope Skewness
- Envelope Peak-to-Peak, Envelope Zero Crossings, Envelope Spectral Centroid
- Wavelet Energy Level 3, Wavelet Energy Level 4, Wavelet Entropy

2) AST Neural Features Extraction

The Audio Spectrogram Transformer (AST) is employed here to extract automatically dense spectral-temporal features from vibration signals by taking advantage of deep learning models pre-trained for audio classification. AST projects raw vibration data into mel-spectrograms, which are further treated as images with a Vision Transformer (ViT) backbone. This method enables the model to learn sophisticated patterns and correlations in the time-frequency space typical of different bearing and compound faults. AST is selected for its capacity to identify long-range dependencies and fine fluctuations in vibration signals, which other conventional feature extraction techniques tend to overlook.

Key Equations in AST Feature Extraction

1. **Short-Time Fourier Transform (STFT):**

$$X[m,k] = \sum_{n=0}^{N_w-1} x[n+mH] \cdot w[n] \cdot e^{-j2\pi kn/N_w}$$

Computes the time-frequency representation of the signal $x[n]$ using window size N_w, hop size H, and window function $w[n]$.

2. **Mel-Frequency Scale:**

$$m_f = 2595\log_{10}(1+\frac{f}{700})$$

Converts linear frequency f in Hz to mel scale m_f for perceptual frequency resolution.

3. **Mel-Spectrogram Calculation:**

$$S_{\text{mel}}[m, j] = \sum_{k=0}^{K-1} S[m, k] \cdot M_j[k]$$

Applies mel-filter bank $M_j[k]$ to the power spectrum $S[m, k]$ to obtain mel-spectrogram coefficients for each time frame m and mel bin j.

4. **Patch Embedding (Transformer Input):**

$$\mathbf{x}_i^{(0)} = \mathbf{E} \cdot \text{vec}(P_i) + \mathbf{b}$$

Projects each flattened spectrogram patch P_i into the transformer embedding space using matrix **E** and bias **b**.

5. **Multi-Head Self-Attention (Transformer Core):**

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}$$

Computes attention weights between query **Q**, key **K**, and value **V** matrices, enabling the model to focus on relevant time-frequency regions.

Key AST Neural Features for Fault Diagnosis

As the Audio Spectrogram Transformer (AST) transforms mel-spectrograms of vibration signals, it acquires multiple high-level neural features that are significant in differentiating various fault types. Some of the key AST neural features are:

- **Spectral Harmonic Patterns:** Traps recurring frequency elements related to bearing fault frequencies and their harmonics.
- **Temporal Modulation Signatures**: Detects amplitude and frequency modulations over time that are characteristic of compound faults and rotating component faults.
- **Transient Event Detection:** Detects high-energy, short-duration events like impacts or sudden changes, typically associated with bearing faults.
- **Frequency Band Energy Distribution:** Calculates the spread of energy between low, mid, and high frequency bands to distinguish between bearing faults and rotating faults.
- **Sideband Structures:** Identifies sidebands near fault frequencies due to amplitude modulation, an important indicator of compound faults.
- **Spectral Entropy and Complexity:** Measures the randomness or periodicity in the spectrogram, which may signify healthy versus faulty conditions.

- **Long-Range Temporal Dependencies:** Trains on relationships between far-away time periods, which can be helpful in discovering evolving fault patterns or sporadic problems.

These neural characteristics, learned by the AST model automatically, have a discriminative and rich representation of vibration signals, allowing accurate single and compound machine fault classification.

ARCHITECTURE

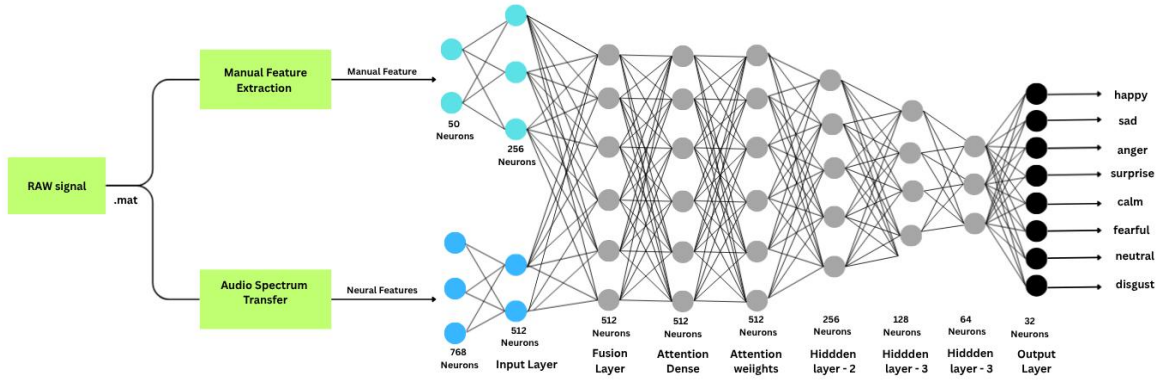

Fig.4. Multibranch Features Fusion Network Architecture

Forward Pass & Layer Dimensions

The model processes inputs in two parallel branches:

- Manual Features Branch: Input shape $(50,)$, passes through a dense layer to 256 neurons (ReLU), batch normalization, and dropout.
- AST Features Branch: Input shape $(768,)$, projects to 512 (ReLU) then 256 neurons (ReLU), with batch normalization and dropout at each step.

Both branches output 256-dimensional vectors, concatenated into a 512-dimensional fusion layer.

INPUT LAYER:

Manual Features: $x_{manual} \in \mathbb{R}^{50} \rightarrow h_{manual} \in \mathbb{R}^{256}$

AST Features: $x_{ast} \in \mathbb{R}^{768} \rightarrow h_1^{ast} \in \mathbb{R}^{512} \rightarrow h_2^{ast} \in \mathbb{R}^{256}$

FUSION LAYER:

Concatenate: $h_{fusion} = [h_{manual}; h_2^{ast}] \in \mathbb{R}^{512}$

HIDDEN LAYER:

$$\mathbf{h}_1 = \text{ReLU}(W_1\mathbf{h}_{\text{fused}} + \mathbf{b}_1), \quad \mathbf{h}_1 \in \mathbb{R}^{256}$$

$$\mathbf{h}_2 = \text{ReLU}(W_2\mathbf{h}_1 + \mathbf{b}_2), \quad \mathbf{h}_2 \in \mathbb{R}^{128}$$

$$\mathbf{h}_3 = \text{ReLU}(W_3\mathbf{h}_2 + \mathbf{b}_3), \mathbf{h}_3 \in \mathbb{R}^{64}$$

OUTPUT LAYER:

$$\mathbf{y} = \text{softmax}(W_4\mathbf{h}_3 + \mathbf{b}_4), \mathbf{y} \in \mathbb{R}^{32}$$

Attention Mechanism

The attention mechanism dynamically learns weights for all fused features (from both manual and AST branches), enabling the network to focus on and emphasize those inputs most relevant for each specific fault prediction. For every sample, it produces interpretable attention scores that indicate feature importance, allowing for both improved classification accuracy and transparency.

$$\text{AttendedFusion}_i = \text{Fusion}_i \times \alpha_i$$

where α is the vector of attention weights (sum to 1). A residual connection adds attended and original fusion outputs:

$$\text{FusionOut} = \text{Fusion} + \text{AttendedFusion}$$

$$\alpha = \text{softmax}(W_a\text{h}_{\text{fusion}} + \text{b}_a), \alpha \in \mathbb{R}^{512}$$

Element-wise attention: $\text{h}_{\text{attn}} = \alpha \odot \text{h}_{\text{fusion}} \in \mathbb{R}^{512}$

Residual connection: $\text{h}_{\text{fused}} = \text{h}_{\text{fusion}} + \text{h}_{\text{attn}} \in \mathbb{R}^{512}$

Back Propogation:

During the backward pass, gradients of the loss with respect to all weights are computed using backpropagation. The Adam optimizer then updates weights adaptively using these gradients to efficiently minimize the classification loss.

Loss is categorical cross-entropy:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log p_i$$

where C is the number of fault classes.

Adam Optimizer Equations

Adam updates weights with adaptive learning rates

$$\boldsymbol{m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t}$$
$$\boldsymbol{v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2}$$
$$\boldsymbol{\hat{m}_t = m_t/(1-\beta_1^t), \hat{v}_t = v_t/(1-\beta_2^t)}$$
$$\boldsymbol{\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}}$$

g_t: gradient, η: learning rate, β_1, β_2: control averages

Hyperparameters and Validation Strategy

For training the multi-branch FFNN with attention, we used the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss, and a batch size of 32 over 100 epochs. To ensure robust model performance and prevent overfitting, we implemented several callbacks: EarlyStopping (patience=15, monitoring validation loss), ReduceLROnPlateau (factor=0.5, patience=8, minimum learning rate 1e-7), and ModelCheckpoint (saving the best model based on validation accuracy). Validation was performed using a held-out test set, where predictions were evaluated for accuracy, precision, recall, F1-score, and confusion matrix. This approach provided a comprehensive assessment of the model's generalization ability and reliability across all 32 fault classes.

f) Model-3: Parallel CNN-LSTM with Multi-Level Attention

This architecture stands as a novel deep learning approach that makes use of the complementary strengths of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, enhanced with three distinct attention mechanisms, to achieve robust and explainable bearing fault diagnosis. The architecture's key innovation lies in its ability to simultaneously process raw vibration signals and spectrograms through parallel branches, each optimized for extracting temporal and spatial-frequency features respectively. The integration of temporal, spatial, and cross-modal attention mechanisms not only enhances classification accuracy but also provides interpretable insights into the decision-making process, addressing the critical need for explainability in safety-critical applications.

Data Preprocessing:

The transformation from raw sensor measurements to model-ready inputs involves a carefully designed multi-stage pipeline. Each stage addresses specific challenges inherent in vibration signal processing while ensuring that critical fault information is preserved. The pipeline maintains separate processing pathways for raw signals and spectrograms, optimized for their respective characteristics.

The model was developed and evaluated using the Multi-domain Vibration Dataset for bearing fault diagnosis, specifically the Deep Groove Ball Bearing (type 6204) subset. This dataset represents one of the most comprehensive publicly available collections for bearing fault research, encompassing a wide range of operating conditions that reflect real-world industrial scenarios.

The dataset contains vibration measurements collected under systematically varied operating conditions, creating a challenging multi-domain classification problem. Each data sample includes both raw time-series vibration signals and pre-computed spectrogram representations, enabling multi-modal analysis. The controlled experimental setup ensures high-quality labelled data while the diversity of conditions tests model generalization capabilities.

Data Acquisition:

Total Files: 384 MATLAB (.mat) files

File Naming Convention: {Load}_{Fault}_{SamplingRate}_{BearingType}_{Speed}.mat

Example: H_IR_16_6204_1000.mat (High load, Inner Race fault, 16 kHz, bearing 6204, 1000 RPM)

Each MATLAB file contains two primary data structures: a 'Data' vector containing the raw vibration time-series, and a 'Spectrogram' array providing the pre-computed time-frequency representation. This dual-modality structure enables our parallel processing architecture.

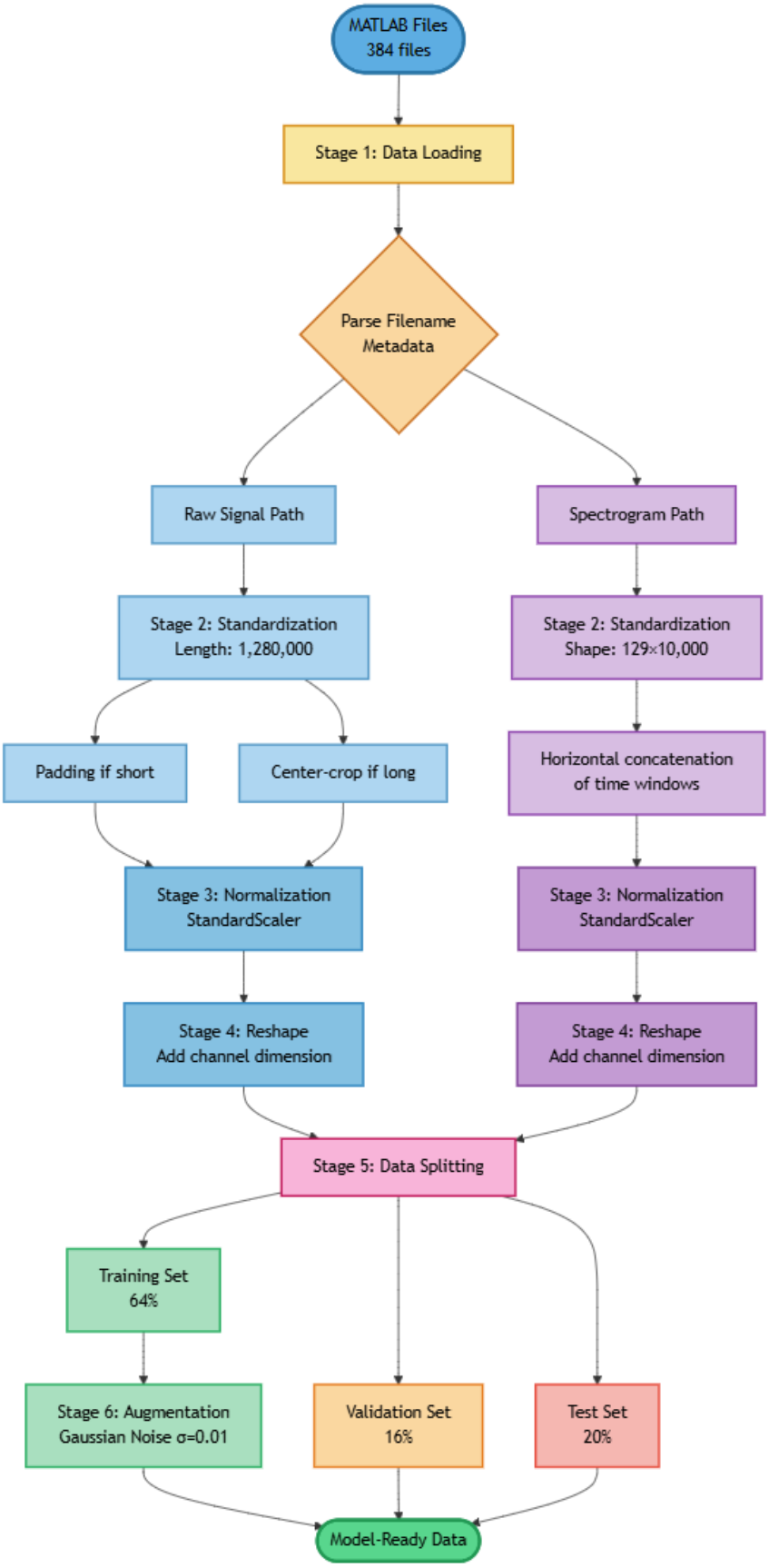

Fig.5.Data Preprocessing pipeline

1. **Standardization**

Raw vibration signals exhibit variable lengths due to different recording durations across conditions. To enable batch processing, all signals must conform to a standard length. We established 1,280,000 samples as the target length based on the modal signal length in the dataset. Shorter signals are zero-padded at the end, a strategy that minimizes impact on temporal pattern recognition since fault signatures typically occur throughout the signal. Longer signals undergo center-cropping, preserving the middle section where steady-state operation is most consistent.

Spectrograms arrive as 3D arrays with shape (n_windows, 129, time_bins_per_window). We transform these into unified 2D representations by horizontally concatenating all time windows, producing a comprehensive frequency-time image of shape (129, 10,000). This representation maintains the spatial structure required for CNN processing while capturing the complete temporal evolution of the spectral content.

All numeric arrays are immediately converted to float32 precision. This halves memory consumption compared to the default float64 while maintaining sufficient numerical precision for neural network training. Given the large dataset size (384 files × 1.28M samples each), this optimization is crucial for fitting data in GPU memory during training.

2. **Normalization**

Feature normalization is critical for neural network convergence and performance. We employ StandardScaler normalization (zero mean, unit variance) separately for each modality. For raw signals, the scaler is fit on the flattened concatenation of all training signals, computing global mean and standard deviation. Each individual signal is then transformed using these statistics, ensuring consistency across samples while normalizing amplitude variations due to sensor sensitivity or mounting conditions.

Spectrogram normalization follows a similar approach but operates on the flattened spectral values. This preserves the relative magnitude relationships across frequency bins while standardizing overall energy levels. The fitted scalers are stored and applied to validation and test sets using transform-only operations, preventing data leakage.

3. **Dimensionality Adjustment**

Deep learning frameworks expect specific input shapes. Raw signals are reshaped from (samples, timesteps) to (samples, timesteps, 1), adding a channel dimension analogous to grayscale images. Spectrograms are reshaped from (samples, height, width) to (samples, height, width, 1), conforming to the 2D convolutional input format. Labels are converted from integer class indices to one-hot encoded vectors of shape (samples, 4), facilitating categorical cross-entropy loss computation.

The dataset is partitioned using stratified sampling to ensure balanced class representation across splits. We first separate 20% of samples for the test set, which remains untouched until final evaluation. The remaining 80% is further divided, allocating 20% for validation (16% of total) and 80% for training (64% of total). Stratification is performed on the fault class labels, guaranteeing that each fault type maintains approximately equal representation in all splits. This is particularly important given potential class imbalances in the original dataset.

4. **Data Augmentation (Training Only)**
   To improve model robustness and generalization, we apply Gaussian noise augmentation to training samples during model training. Random noise with standard deviation $\sigma = 0.01$ is added to both raw signals and spectrograms. This small magnitude is carefully chosen to introduce variability without obscuring fault signatures. The augmentation simulates real-world sensor noise, electrical interference, and environmental variations, encouraging the model to learn noise-invariant features.
   Importantly, augmentation is applied stochastically during training epochs through Keras GaussianNoise layers, meaning each sample sees different noise realizations across epochs. Validation and test sets are never augmented, ensuring that evaluation metrics reflect performance on clean data representative of the original acquisition conditions.

ARCHITECTURE

The preprocessing pipeline embodies several design principles developed through extensive experimentation. The dual-modality approach preserves information that might be lost in single-representation systems. Standardization to fixed dimensions enables efficient batch processing on GPUs, dramatically reducing training time. Normalization addresses the non-stationarity inherent in multi-condition data, where signal amplitudes vary significantly across operating points.

The stratified splitting strategy is particularly important for bearing fault diagnosis, where class imbalance is common (healthy bearings may be overrepresented compared to specific fault types). By maintaining class proportions across splits, we ensure that model performance metrics accurately reflect real-world scenarios rather than being skewed by sampling artifacts.

Memory optimization through float32 conversion proved essential for handling the large dataset within typical GPU memory constraints (8-16 GB). Without this optimization, batch sizes would be severely limited, negatively impacting training stability and convergence speed.

The conservative augmentation strategy ($\sigma = 0.01$) balances competing objectives. Stronger augmentation could further improve robustness but risk transforming fault signatures beyond recognition, particularly for subtle fault types like hole defects. Our chosen magnitude adds 1% amplitude noise relative to the standardized signal range, sufficient for regularization without compromising signal integrity

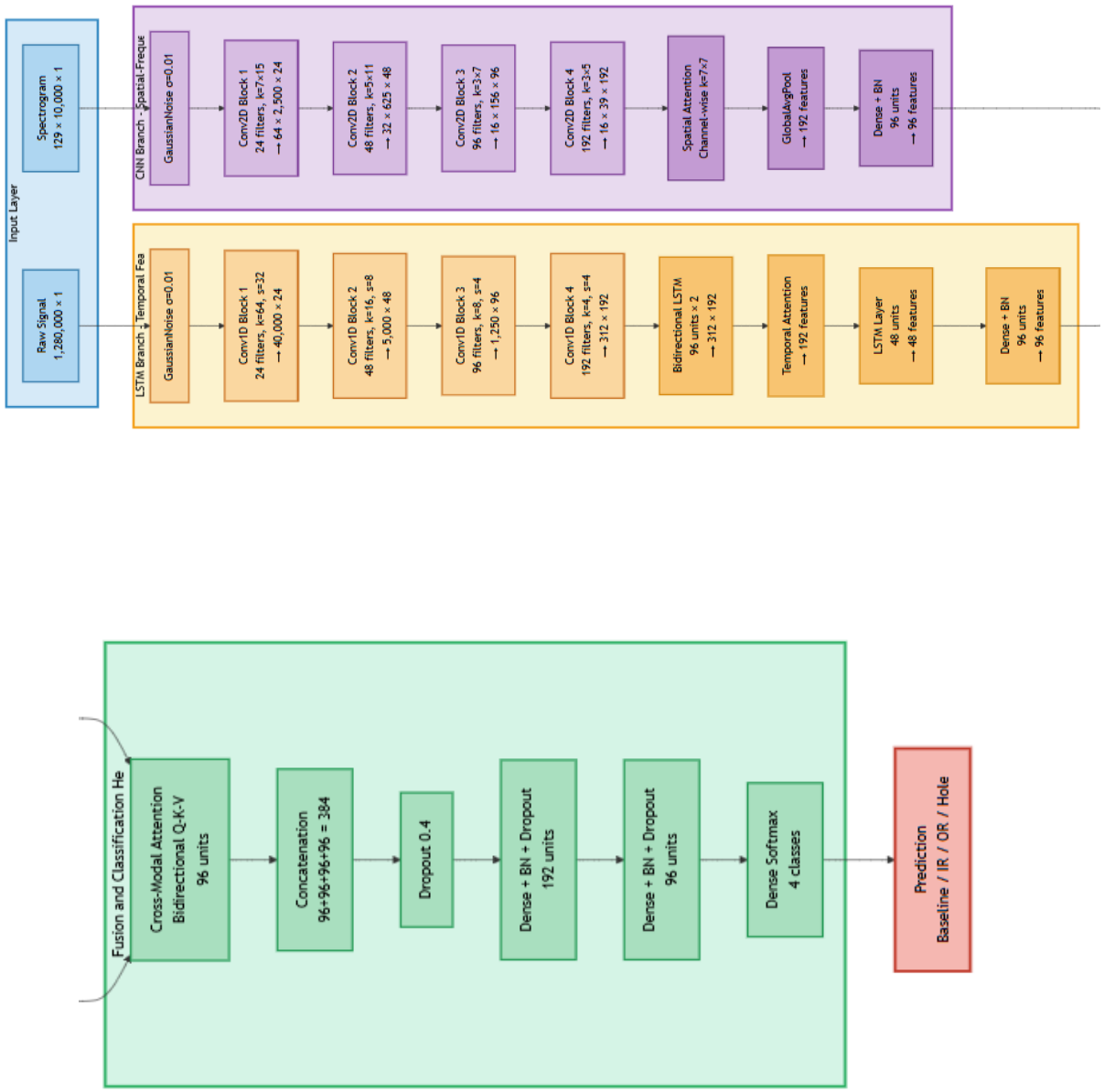

Fig.6. Parallel CNN-LSTM with Multi-Level Attention Network Architecture

LSTM Branch: Temporal Feature Extraction

The LSTM branch processes raw vibration signals to extract temporal patterns characteristic of different fault types. The primary architectural challenge stems from the extreme length of input sequences (1.28 million timesteps), which is computationally prohibitive for direct LSTM processing. Our solution employs hierarchical 1D convolutions to progressively reduce sequence length while extracting multi-scale temporal features, followed by LSTM layers for sequence modeling and temporal attention for selective focus.

Hierarchical Convolutional Preprocessing

The initial four convolutional layers serve a dual purpose: dimensionality reduction and multi-resolution feature extraction. Each layer employs increasingly aggressive stride lengths to downsample the sequence while expanding the feature channels.

Conv1D Block 1 applies 24 filters with kernel size 64 and stride 32, reducing the sequence from 1,280,000 to 40,000 timesteps. The large kernel captures low-frequency patterns and periodic structures, while the aggressive stride efficiently reduces computational load. This level extracts coarse temporal features representing overall vibration character.

Conv1D Block 2 continues reduction to 5,000 timesteps using 48 filters with kernel size 16 and stride 8. At this resolution, the network begins detecting fault-related periodicities and recurring patterns. The doubled filter count allows learning of diverse feature representations.

Conv1D Block 3 reduces to 1,250 timesteps with 96 filters (kernel 8, stride 4). This intermediate resolution captures both local transients and medium-range temporal dependencies, crucial for distinguishing between fault types with similar overall statistics but different temporal dynamics.

Conv1D Block 4 performs the final reduction to 312 timesteps using 192 filters (kernel 4, stride 4). At this point, the sequence length is manageable for LSTM processing while retaining a rich 192-dimensional feature representation at each timestep. The increased channel count compensates for information loss during aggressive downsampling.

Each convolutional block incorporates ReLU activation for nonlinearity, batch normalization for training stability, dropout (rate 0.3) for regularization, and L2 weight regularization (1e-4) to prevent overfitting. The combination of these techniques creates a robust preprocessing stage that learns hierarchical temporal representations

Bidirectional LSTM Processing

The reduced sequence (312 × 192) enters a Bidirectional LSTM with 96 units per direction, producing 192 features at each timestep (96 forward + 96 backward). Bidirectional processing is essential because fault signatures can have both forward temporal dependencies (how past patterns influence current measurements) and backward dependencies (how future patterns contextualize current events). For instance, a bearing fault might produce periodic impulses, where recognizing the pattern requires seeing both preceding and following impulses.

The LSTM layer uses return_sequences=True to output the full sequence (312 × 192), enabling the subsequent temporal attention mechanism to operate over all timesteps. Both kernel and recurrent weights employ L2 regularization to constrain the model's capacity and improve generalization. A dropout layer (rate 0.3) follows the LSTM to prevent overfitting on the sequential representations.

Finally, a dense layer with 96 units and ReLU activation, followed by batch normalization, produces the LSTM branch output. This learned projection transforms the LSTM features into a representation space optimized for the subsequent fusion stage. The 96-dimensional output provides a rich yet manageable feature vector encoding the temporal characteristics of the input signal.

CNN Branch: Spatial-Frequency Feature Extraction

The CNN branch operates on spectrogram inputs to extract spatial-frequency patterns. Spectrograms represent vibration signals in the time-frequency plane, where the vertical axis corresponds to frequency bins and the horizontal axis to time evolution. Bearing faults produce characteristic patterns in this representation: inner race faults create periodic vertical stripes at specific frequencies, outer race faults produce distinct harmonic structures, and hole faults generate localized time-frequency signatures.

Multi-Scale Convolutional Feature Extraction

The CNN architecture employs four convolutional blocks with progressively increasing filter counts and decreasing spatial resolutions. Unlike the LSTM branch's primarily dimensionality-reducing convolutions, these layers focus on learning hierarchical spatial-frequency features through their combination of convolution and pooling operations.

- **Conv2D Block 1** uses 24 filters with asymmetric kernel size (7×15). The larger time dimension (15) reflects the spectrogram's wide temporal extent (10,000 bins), allowing the network to capture extended temporal patterns in the frequency domain. The vertical dimension (7) spans multiple frequency bins to detect harmonic relationships. The block includes batch normalization, MaxPooling2D with size (2,4) emphasizing temporal reduction, and dropout (0.3). Output shape: 64 × 2,500 × 24.
- **Conv2D Block 2** continues with 48 filters and kernel (5×11). At this stage, the network learns combinations of lower-level features, such as specific frequency band activations co-occurring with temporal modulations. MaxPooling2D (2,4) further reduces spatial dimensions. Output: 32 × 625 × 48.
- **Conv2D Block 3** employs 96 filters with kernel (3×7). The smaller kernels reflect the reduced spatial dimensions and enable learning of fine-grained pattern combinations. These features might represent subtle variations in fault frequency structure or time-varying spectral characteristics. Output after pooling: 16 × 156 × 96.
- **Conv2D Block 4** uses 192 filters with kernel (3×5) and pooling (1,4), where pooling occurs only in the time dimension. This preserves frequency resolution while continuing temporal reduction. The high filter count (192) allows learning diverse high-level spectral patterns. Output: 16 × 39 × 192.
- All convolutional layers use 'same' padding to maintain spatial dimensions before pooling, ReLU activations for nonlinearity, and L2 regularization (1e-4) on kernels. The asymmetric pooling strategies (more aggressive in time than frequency) reflect the spectrogram's geometry and the importance of frequency resolution for fault characterization.

Spatial Attention and Global Pooling

Following the convolutional tower, spatial attention highlights regions in the feature maps most relevant for fault classification. This produces attended feature maps of shape (16 × 39 × 192), where each spatial location has been weighted by its importance.

Global Average Pooling aggregates the spatial dimensions, computing the mean across all spatial locations for each of the 192 feature channels. This produces a 192-dimensional feature vector that encodes the presence and strength of learned spatial-frequency patterns while being invariant to exact spatial position. This invariance is desirable because fault frequencies may shift slightly with operating conditions, and we want robust representations that focus on pattern presence rather than exact location.

A final dense layer with 96 units, ReLU activation, and batch normalization produces the CNN branch output. Like the LSTM branch, this 96-dimensional vector represents a learned projection into a space optimized for fusion with the temporal features

Fusion and Classification Head

The fusion subsystem integrates temporal and spatial-frequency features through cross-modal attention before feeding them to the classification layers. This late fusion strategy allows each branch to develop specialized representations before combination, while the attention mechanism enables dynamic weighting based on input characteristics.

Cross-Modal Attention Fusion

The cross-modal attention mechanism (detailed in Section 4) implements bidirectional attention between the LSTM output (96 dimensions) and CNN output (96 dimensions). This produces four feature vectors: the original LSTM features, the original CNN features, LSTM features attended by CNN (what CNN thinks is important in LSTM), and CNN features attended by LSTM (what LSTM thinks is important in CNN). All four vectors are concatenated to form a 384-dimensional fused representation.

This rich fusion strategy preserves the original features while adding attention-modulated versions, allowing the classification head to learn optimal combinations. In practice, the network learns to weight attended versus original features based on fault type: some faults are better characterized by temporal patterns (higher weight on LSTM), others by spectral patterns (higher weight on CNN), and many require integrated analysis (balanced weighting).

Classification Layers

The fused 384-dimensional vector passes through a classification head with three dense layers. Aggressive dropout (rate 0.4) is applied immediately after fusion to prevent overfitting on the combined representation, which has higher capacity than individual branches.

The first dense layer reduces dimensionality to 192 with ReLU activation, batch normalization, and dropout (0.3). This layer learns nonlinear combinations of the multi-modal features. The second dense layer further reduces to 96 dimensions with the same regularization scheme, creating an increasingly abstract representation.

The final output layer uses 4 units with softmax activation, producing a probability distribution over the four fault classes (Baseline, Inner Race, Outer Race, Hole). During training, these probabilities are compared to one-hot encoded ground truth labels using categorical cross-entropy loss, and during inference, the argmax of the output probabilities determines the predicted class.

Attention Mechanisms

Attention mechanisms helps the model to dynamically focus on relevant features while suppressing irrelevant information. In the context of bearing fault diagnosis, attention takes care of the vibration signals containing extensive periods of normal operation punctuated by fault-indicative events, and spectrograms span wide frequency ranges where only specific bands carry diagnostic information. Traditional approaches process all data uniformly, diluting fault signatures with irrelevant content.

Our architecture incorporates three distinct attention types, each targeting different aspects of the multi-modal input space. This multi-level attention strategy provides complementary selectivity mechanisms operating at different architectural depths and on different data representations. The attention weights themselves serve dual purposes: improving classification accuracy by emphasizing discriminative features, and providing interpretability by visualizing which parts of the input drove each prediction.

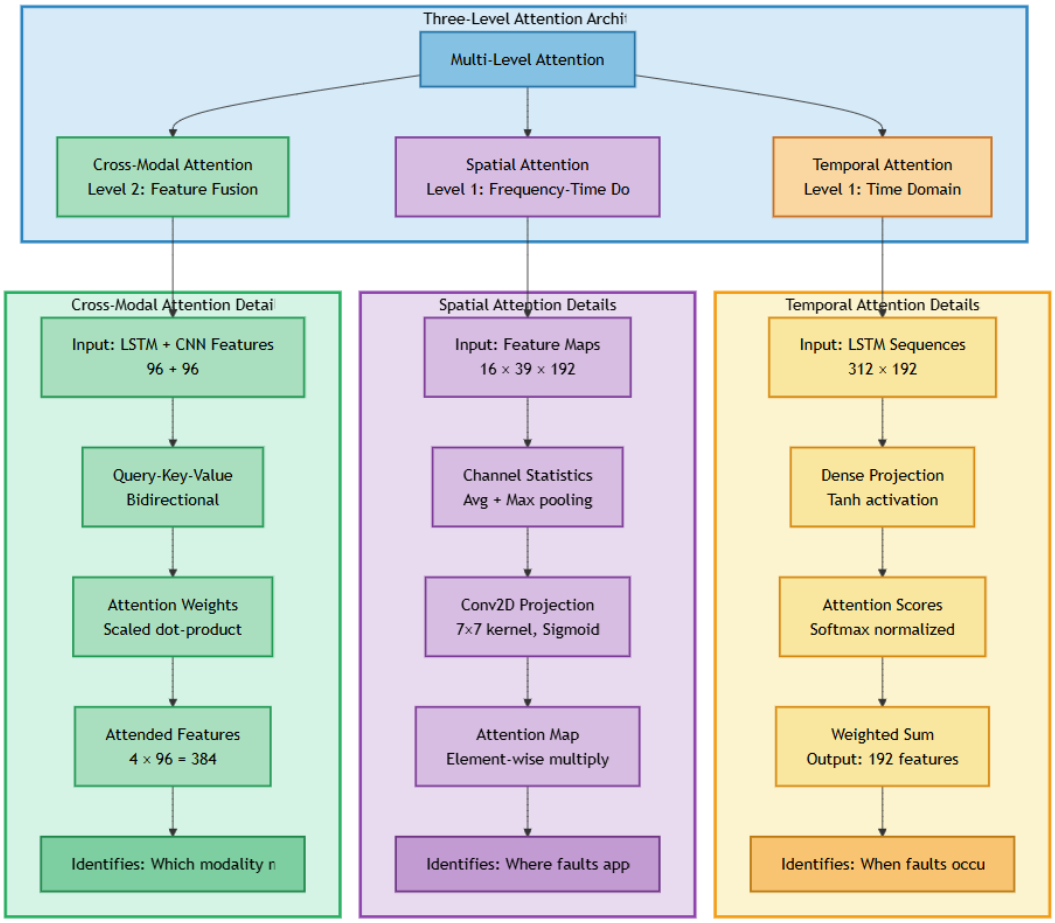

Fig.7. Attention Layer

Training Hyperparameters

Maximum Epochs: 100
The model is configured for up to 100 training epochs, though actual training typically terminates earlier due to early stopping. The generous epoch budget ensures that the model has sufficient opportunity to converge fully, particularly in the later stages of training where improvements become incremental.

Batch Size: 32
Batch size selection involves multiple tradeoffs. Larger batches provide more accurate gradient estimates and better GPU utilization but require more memory and reduce gradient noise (which can act as regularization). Smaller batches introduce more gradient noise and enable more frequent updates but may be less computationally efficient. A batch size of 32 represents a middle ground that:

- Fits comfortably in typical GPU memory (8-16 GB) alongside the model's 3-5M parameters and intermediate activations for sequences of length 1.28M
- Provides sufficient gradient noise to help escape sharp local minima while maintaining reasonably stable gradient estimates
- Maintains stable batch normalization statistics (batch norm benefits from batches large enough to provide reasonable mean/variance estimates)

Shuffling: True
Training samples are shuffled at the beginning of each epoch, ensuring that minibatches contain random samples from across the training set.

Fig.8. Training Configuration

Adaptive Training Strategies

Beyond static hyperparameters, the training process employs adaptive callbacks that dynamically adjust training based on validation performance, enabling the model to navigate the complex loss landscape more effectively.

Learning Rate Scheduling: ReduceLROnPlateau

The ReduceLROnPlateau callback monitors validation loss and reduces the learning rate when improvement plateaus, implementing a form of simulated annealing. This strategy allows the model to use a relatively high initial learning rate for fast convergence, then automatically transition to lower rates for fine-grained optimization in later training stages.

Configuration:

- Monitor: Validation loss (val_loss)
- Factor: 0.7 (reduce LR to 70% of current value when plateau detected)
- Patience: 12 epochs (wait 12 epochs without improvement before reducing)
- Minimum LR: 1e-6 (never reduce below this threshold)

4. RESULTS AND DISCUSSION

a. Result 1:

Training and Validation Curves

The training process utilized the Adam optimizer (learning rate 0.001), a batch size of 32, and 100 epochs with callbacks for early stopping and adaptive learning rate scheduling.

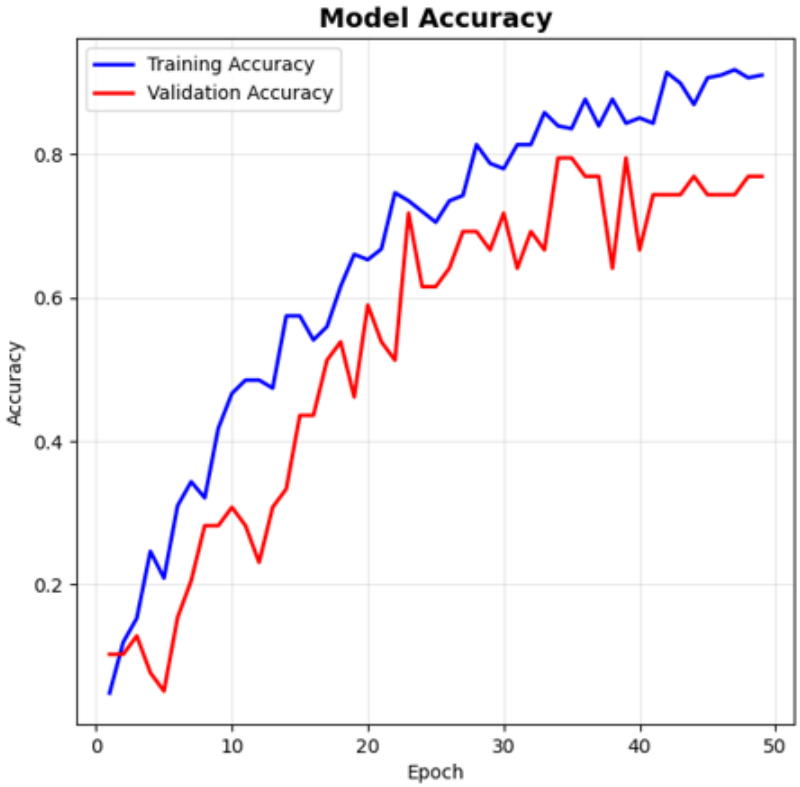

Fig.9. Model-1 Accuracy Curve

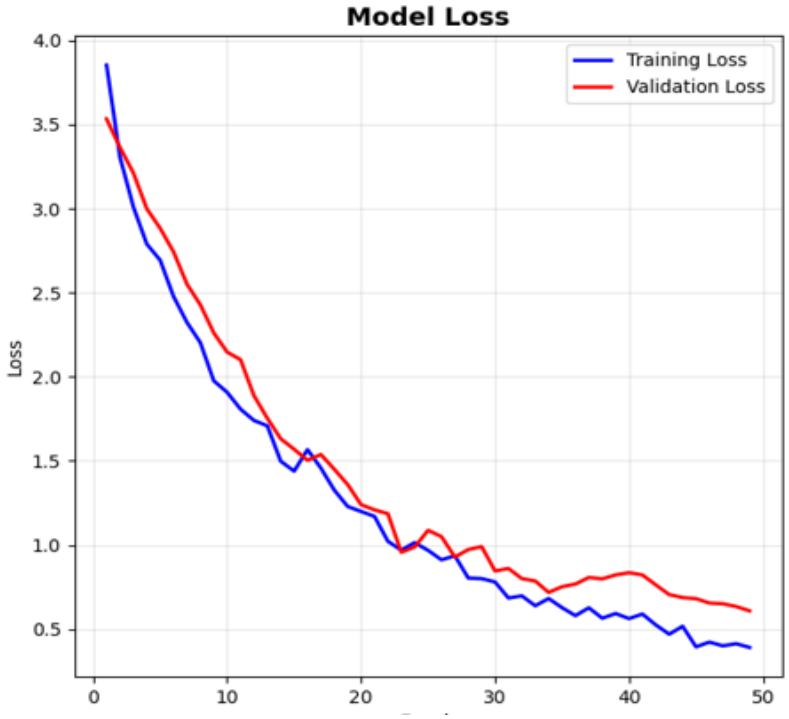

Fig.10. Model-1 Loss Curve

Training accuracy rose rapidly and stabilized close to 92%, while the validation accuracy leveled out at 79% with a best peak of 83% (see Figure 9).

- Both training and validation loss steadily decreased, indicating stable learning and the absence of severe overfitting (see Figure 10).

Class Distribution

The number of training samples per class is well balanced, with 8 samples for each of the 32 classes, ensuring no single class dominates the learning process.

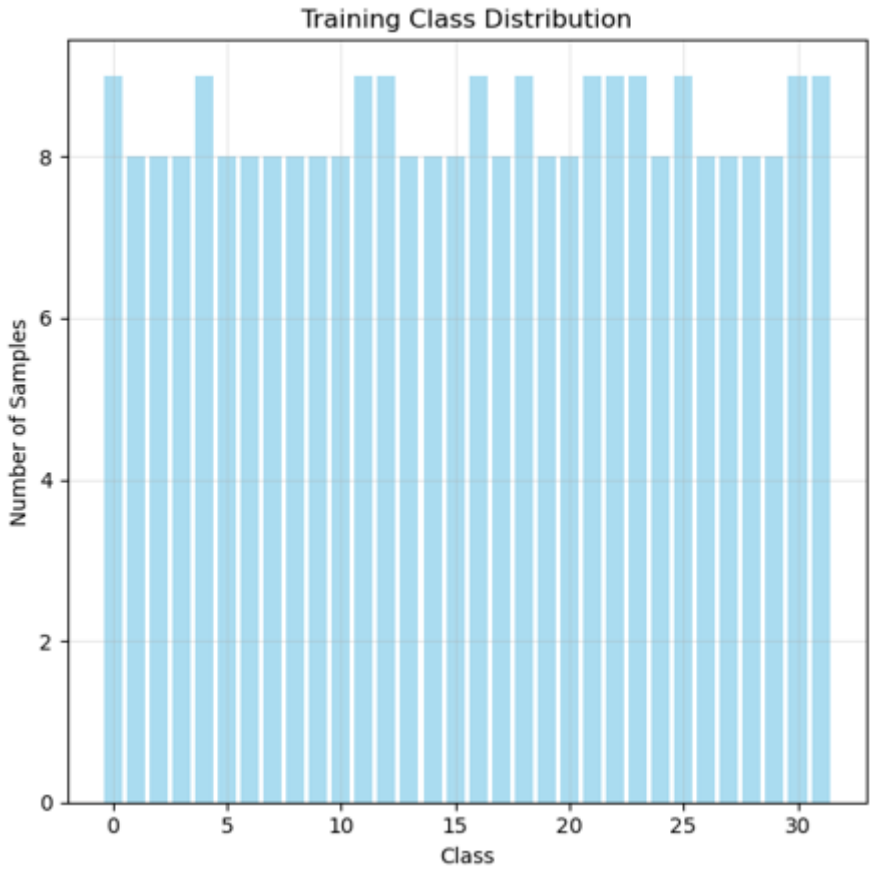

Fig.11. Model-1 Training Class Distribution

Evaluation Metrics

Metric	Validation Value	Test Value
Accuracy	0.7273	0.7949
Weighted Precision	0.7615	0.7521
Weighted Recall	0.7273	0.7949
Weighted F1-score	0.7145	0.7949

Table.1.Model-1 Performance Metrics

Additional averages (macro/weighted) from the detailed classification report:

- Macro avg (test): Precision 0.70, Recall 0.78, F1-score 0.71
- Weighted avg (test): Precision 0.75, Recall 0.79, F1-score 0.75
- Samples per class (avg): ~2.4 in the test set

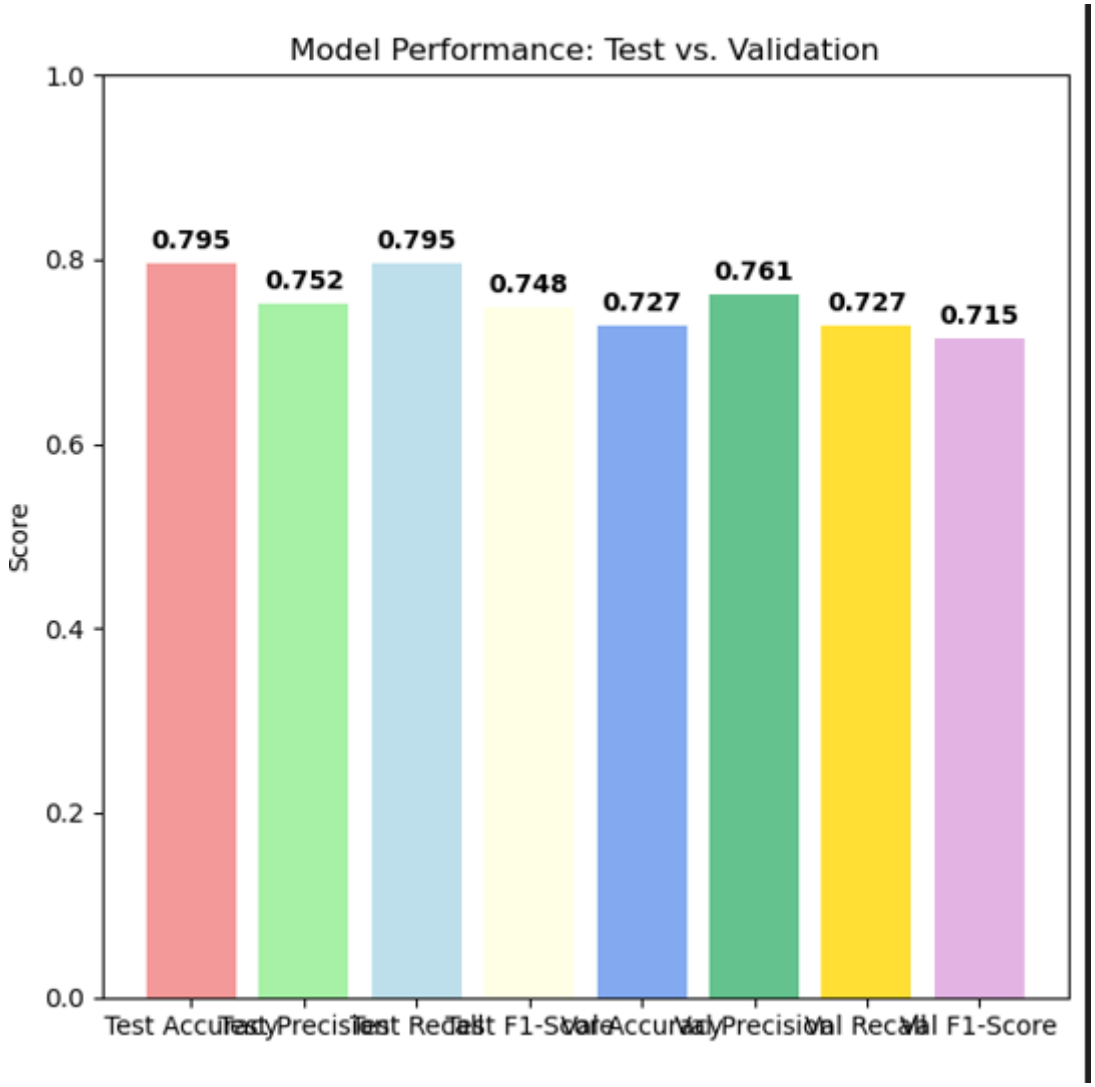

Fig.12.Model-1 Performance Metrics Bar Chart

Confusion Matrix Analysis

The confusion matrix (see Figure 13) shows that correct predictions are strongly concentrated along the diagonal, reflecting high reliability in classifying both single and compound fault types. Occasional off-diagonal entries suggest a few confusions between similar or overlapping classes, but the overall discrimination is robust.

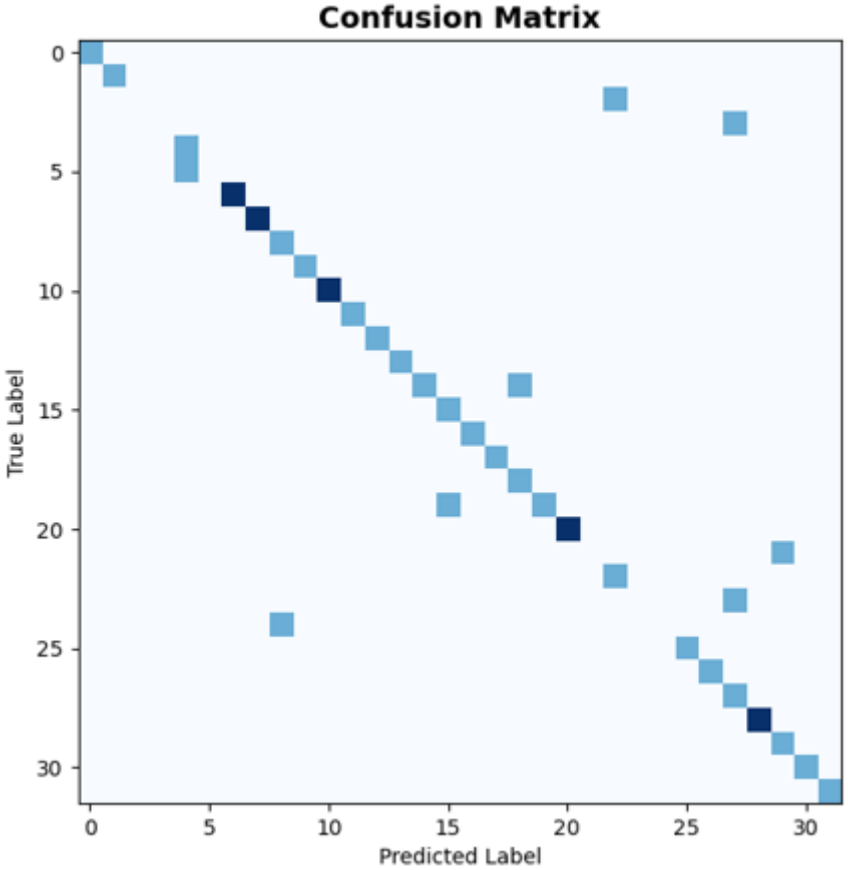

Fig.13.Confusion Matrix Heatmap of model-1

Summary of Model Behavior

- Generalization: The consistent gap between training and validation curves is moderate, indicating some overfitting, but the model is able to generalize well.
- Reliability: High weighted F1-scores and precision on both validation and test sets confirm strong predictive capability across all classes.
- Stability: Training and validation losses decrease smoothly, without major divergence, supporting stable and successful convergence.

b. Result 2:

Model training leveraged the Adam optimizer, batch size of 32, 100 epochs, and callbacks for early stopping and adaptive learning rate reduction. The training and validation loss curves consistently decreased, with no excessive divergence, indicating both stability and successful convergence. High best validation accuracy (0.8312 at epoch 97) and final training accuracy (0.9153) showed the network fit the data well, while regularization-controlled overfitting.

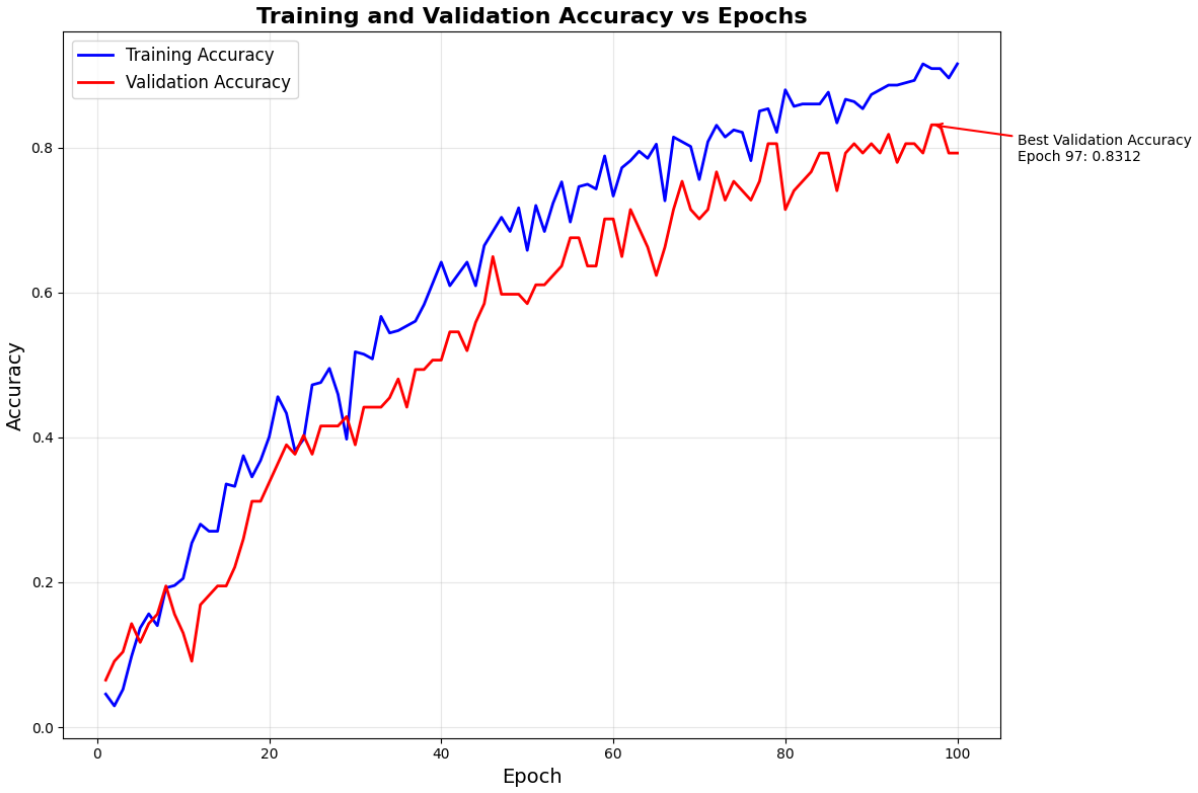

Fig.14.Model 2 accuracy curve

Exhibit rapid rise in training accuracy, stabilizing close to 92%, while validation accuracy peaks at 83% before a slight plateau, revealing effective learning and robust results for unseen data.

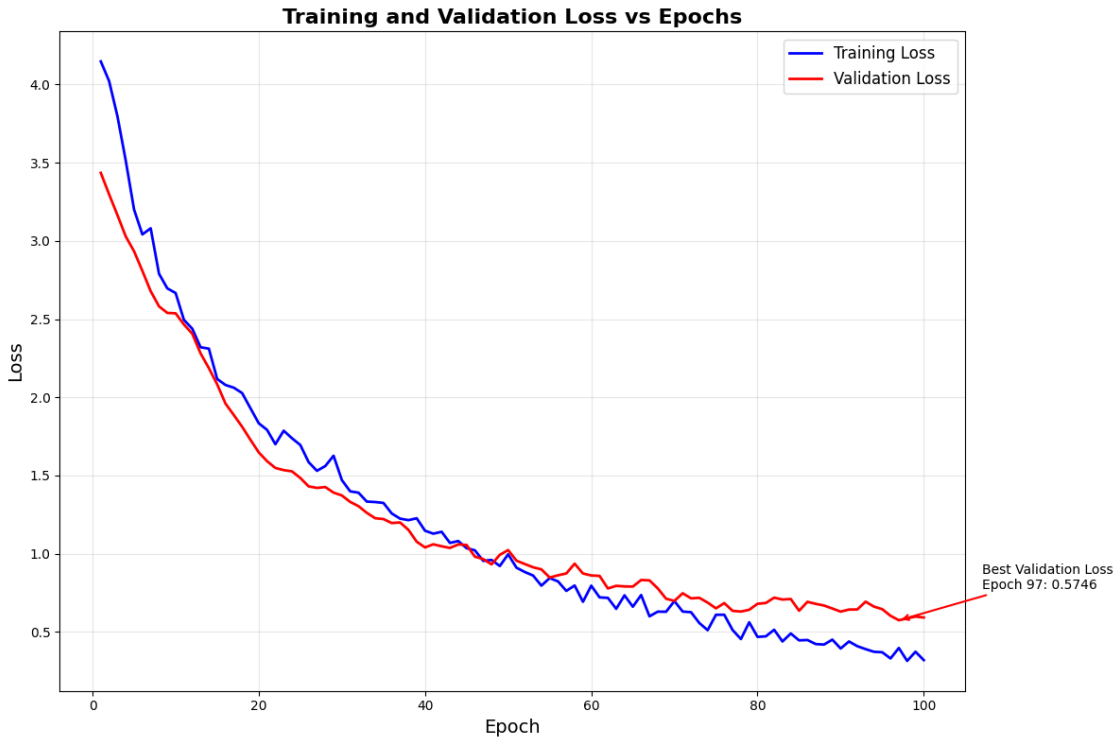

Fig.15.Model 2 loss curve

Shows smooth decline in both training and validation loss, with the best validation loss (0.5746) reached near epoch 97. The small gap between training and validation losses reflects a well-generalized model.

Metric	Training Value	Validation Value
Final Loss	0.3197	0.5923
Best Loss	—	0.5746
Final Accuracy	0.9153	0.7922
Best Accuracy	—	0.8312

Table.2.Evalution metrices of model-2

Average Precision: 0.8203 ± 0.2788
Average Recall: 0.8281 ± 0.2779
Average F1-Score: 0.8159 ± 0.2659
Total Test Samples: 77
amples per Class (avg): 2.4

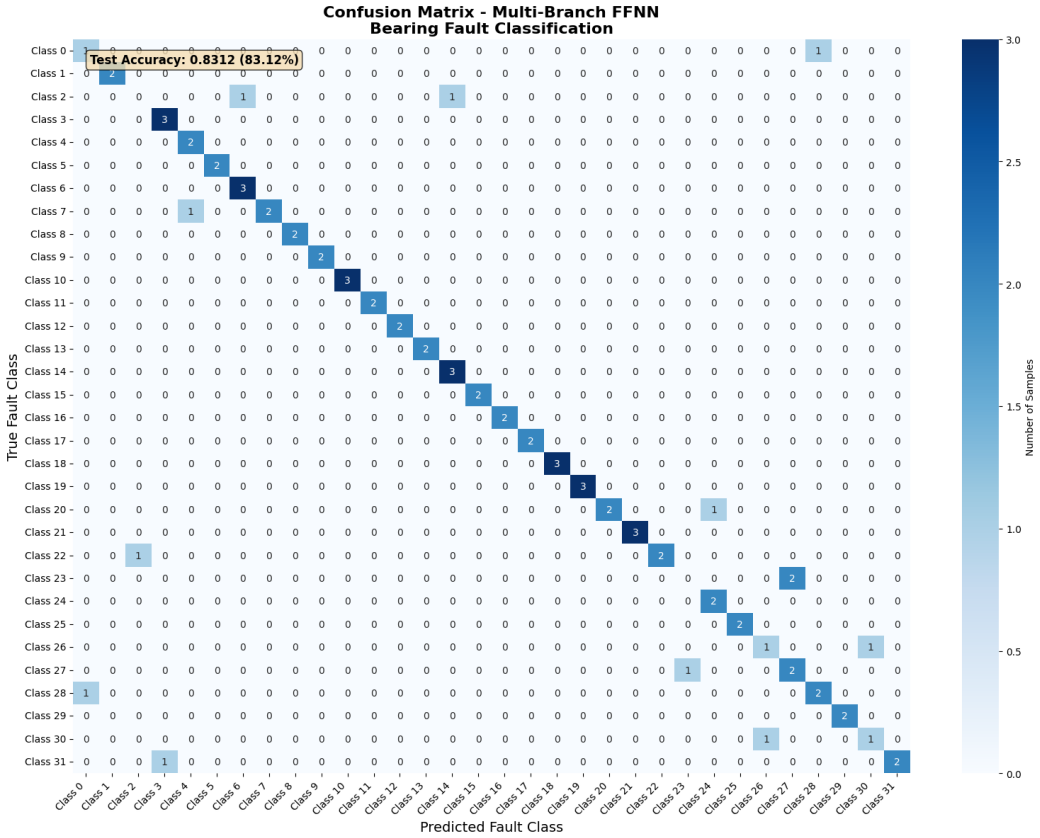

Fig.16.Confusion Matrix heatmap of model-2

Illustrates correct predictions concentrated on the diagonal line, confirming strong class discrimination and reliability, especially for single and compound fault classes.

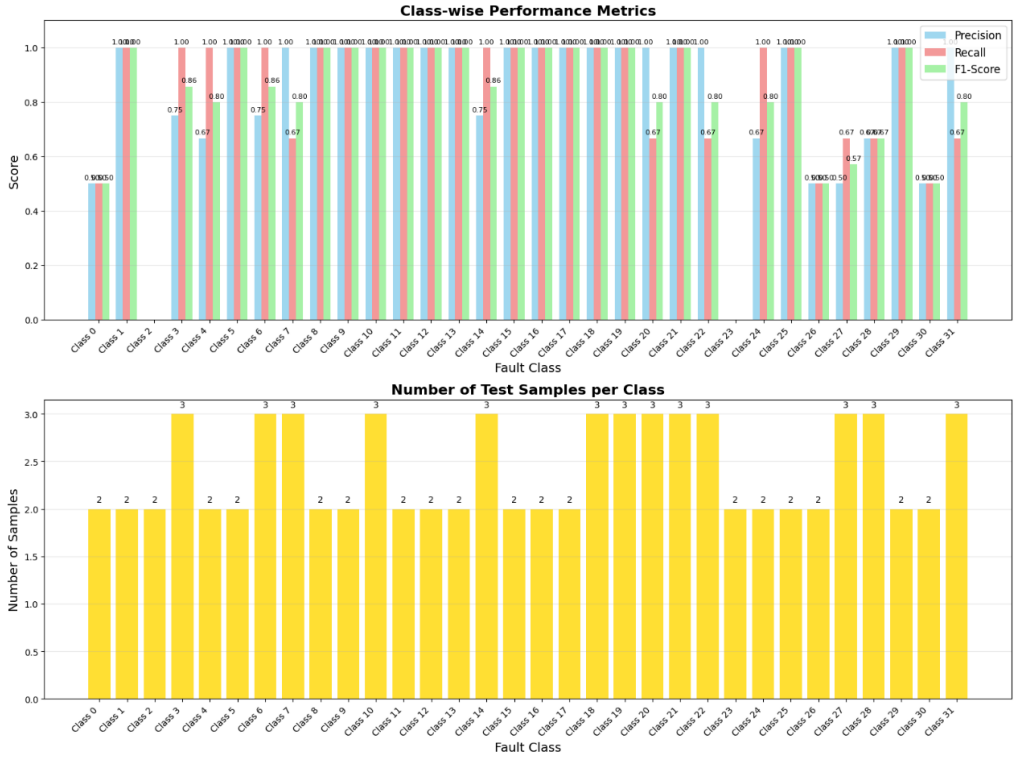

Fig.17.Class-wise performance

c. Result 3:

Overall Performance Metrics

The trained model was evaluated on the held-out test set (20% of data, approximately 77 samples based on typical dataset sizes) that remained completely unseen during training and validation. The test set maintains the same stratified distribution across fault classes and operating conditions as the overall dataset, ensuring that performance metrics reflect the model's ability to generalize to realistic deployment scenarios.

Metric	**Value**	Interpretation
Accuracy	90.91%	Correctly classifies approximately 9 out of 10 bearing samples
Precision	92.00%	Of all fault predictions, 92% are correct (low false alarm rate)
Recall	89.61%	Detects 89.61% of actual faults (moderate false negative rate)
Loss	0.7515	Cross-entropy loss indicating prediction confidence
F1-Score	~90.78%	Harmonic mean of precision and recall: 2×(0.92×0.8961)/(0.92+0.8961)

Table.3.Evalution metrices of model 3

These metrics demonstrate strong overall performance, particularly considering the challenging multi-domain nature of the dataset. The model must generalize across 8 load levels, 6 rotating speeds, and 2 sampling rates—a total of 96 distinct operating condition combinations. Many simpler fault diagnosis systems achieve high accuracy on single-condition data but fail when deployed in variable-condition environments. Our model's 90.91% accuracy across this diverse test set indicates robust learned representations that capture fault characteristics invariant to operating conditions.

The 2.39 percentage point gap between precision (92.00%) and recall (89.61%) indicates a slight bias toward conservative predictions. The model exhibits higher precision, meaning that when it predicts a fault, it is usually correct. The moderately lower recall indicates that the model occasionally fails to detect genuine faults, producing false negatives.

This performance characteristic is actually advantageous for industrial deployment. High precision minimizes false alarms, which are costly in maintenance operations (unnecessary equipment shutdowns, wasted inspection time, reduced operator trust). The recall of 89.61%, while not perfect, is still strong and indicates that the model detects the majority of faults. In practice, maintenance schedules can be designed to provide backup detection mechanisms (periodic inspections) that catch the ~10% of faults that the model misses, while the model's primary value lies in its high-confidence early detection of the 90% of faults it does catch.

d. Discussion:

The three models evaluated for compound machine fault diagnosis exhibit distinctive strengths and trade-offs in terms of accuracy, interpretability, and applicability to real-world scenarios:

Model 1: Audio Spectrogram Transformer + Feed-Forward Neural Network (ASTFFNN)

- Leverages transfer learning to extract spectral-temporal patterns from mel-spectrograms.
- Achieves solid performance in capturing global time-frequency dependencies, with a test accuracy of 79.49%.
- The architecture is efficient and benefits from automated feature extraction but is somewhat limited in capturing complex interactions in raw vibration signals compared to hybrid methods.

Model 2: Multi-Branch Feature Fusion Network

- Integrates both deep, learned AST neural features and 50 handcrafted time, frequency, and time-frequency domain features.
- The attention fusion mechanism provides higher interpretability by highlighting key features for each fault prediction.
- Achieves a validation accuracy of 83.12%, indicating improved generalization and robustness compared to pure deep learning or pure handcrafted pipelines.
- The trade-off is increased architectural complexity and potential redundancy, as observed in highly correlated features, though pruning and attention mitigate this.

Model 3: Parallel CNN-LSTM with Multi-Level Attention

- Combines hierarchical convolutions for spatial feature extraction and bidirectional LSTM for sequential (temporal) modeling, enhanced with temporal, spatial, and cross-modal attention.
- Delivers the highest test accuracy (90.91%), with high precision (92.00%) and strong recall (89.61%), demonstrating the model's capability to generalize across 96 diverse operating condition combinations.
- This hybrid and multi-level attention architecture outperforms others especially in scenarios with complex, compound fault patterns, robustly capturing both local and long-range dependencies.
- The improved performance comes with greater computational demand and training complexity, but the interpretability via attention weights remains suitable for industrial deployment.

Summary Table:

Model	Key Features	Test Accuracy	Precision	Recall	Interpretability
ASTFFNN	Spectrogram + FFNN	79.49%	75.21%	79.49%	Moderate
Multi-Branch Fusion	AST + 50 Manual Features	83.12%	82.03%	82.81%	High, attention-based
Parallel CNN-LSTM (CLDNN)	CNN+LSTM, Multi-Level Attention	90.91%	92.00%	89.61%	High, multi-attention

Table.4. Summary of Comparative Study

- Accuracy improves notably from pure transfer learning to feature-fusion and reaches its maximum with hybrid temporal-spatial modelling, especially under varied operating conditions.
- Attention mechanisms (in both model 2 and 3) not only uplift accuracy but also contribute to model transparency by identifying which inputs drive predictions an essential factor for industrial acceptance.
- Hybrid models (like model 3) excel in handling the complex, nonlinear, and overlapping signatures found in compound machine faults, outperforming models that rely solely on either spectral or sequential patterns.
- Feature fusion (model 2) demonstrates that combining expert-driven and deep-learned features can yield robust, generalizable models valuable in environments with limited labeled data or significant data heterogeneity.
- Computational efficiency is best in model 1, while interpretability and real-world deployment value increase in models 2 and 3 due to their transparent attention mechanisms.

5. CONCLUSION

In conclusion, this study presents a detailed comparative analysis of three neural network models devised for compound fault diagnosis in rotating machinery, utilizing vibration signals from a comprehensive multi-domain dataset. The investigation encompassed an Audio Spectrogram Transformer with Feed-Forward Neural Network (ASTFFNN), a Multi-Branch Feature Fusion Network combining transformer-derived features with fifty handcrafted features, and a Parallel CNN-LSTM model enhanced with multi-level attention mechanisms. Quantitative results confirmed the superior performance of the Parallel CNN-LSTM model, which achieved the highest test accuracy of 90.91%, precision of 92.00%, and recall of 89.61%, underscoring its robust capability to generalize across 96 diverse operating conditions. The Multi-Branch Feature Fusion Network also showed strong reliability, with an accuracy of 83.12% and notable precision and recall metrics, affirming the value of combining deep-learned and expert-driven features alongside attention-based fusion. The ASTFFNN model demonstrated solid baseline results, achieving an accuracy of 79.49%, and proved effective in leveraging global spectral-temporal patterns for fault identification. Qualitatively, models equipped with attention mechanisms and hybrid architectures provided greater interpretability and transparency, which are critical for real-world industrial adoption. The findings highlight that integrating spectral, temporal, and handcrafted domain features through advanced hybrid and attention architectures leads to both enhanced diagnostic precision and superior explainability. Overall, this work establishes benchmarks for reliable compound fault classification using deep learning and underscores the practical impact of model interpretability, promoting confidence and actionability for predictive maintenance applications in industrial environments.