Machine Fault Detection using Vision Transformers (ViTs)

Sajandeep Singh
Punjab AI Excellence
Email: work.sajandeep@gmail.com

Sandeep Singh Sandha
Punjab AI Excellence
Email: sandhaiitr@gmail.com

***Abstract*—Machine fault detection plays a critical role in predictive maintenance, safety, and operational efficiency in industrial systems. Early and reliable detection of bearing and component failures reduces downtime, prevents accidents, and lowers maintenance costs. Traditional approaches, including convolutional neural networks (CNNs), have demonstrated success in vibration-based fault diagnosis; however, their reliance on local receptive fields limits their ability to capture long-range dependencies in complex sensor signals.**

In this work, we propose a Vision Transformer (ViT)-based framework for machine fault detection, leveraging self-attention to model both local and global signal relationships. The study uses the University of Ottawa Rolling-element Dataset – Vibration and Acoustic Faults under Constant Load and Speed (UORED-VAFCLS), which provides accelerometer and acoustic sensor data under controlled conditions. Raw time-series signals were first formatted into three-channel RGB images and resized to 224×224 pixels, consistent with the standard input size proposed in the original Vision Transformer paper. To generate these images, accelerometer signals were converted into scalograms (segment size = 2048 samples, 49 ms per window), while acoustic signals were converted into spectrograms. Separate ViT models were trained for each modality (accelerometer and acoustic), enabling modality-specific feature learning and comparative evaluation.

Experimental results show that the proposed ViT-based models outperform CNN baselines in accuracy and robustness, highlighting the effectiveness of transformers in handling multimodal sensor data for industrial fault diagnosis.

***Index Terms*—Vision Transformer (ViTs), Machine Fault Detection, Deep Learning, Vibration Analysis, Acoustic and Accelerometer Signals, Scalogram, Spectrogram**

I. Introduction

The modern world is surrounded by technology and machines. With the rapid growth in real-world industries including manufacturing, robotics, artificial intelligence , automobile, and even in critical safety systems including nation's security systems, detecting machine faults in early stages became a crucial task. Unexpected faults in rotating machinery parts, such as bearings, gears, or motors, can lead to production downtime, reduce efficiency, and even cause serious accidents. Therefore, predicting early faults is essential for predictive maintenance and ensuring safety.

Before Vision Transformer (ViT) was introduced in 2020, most research in machinery fault detection relies on machine learning and deep learning methods, with convolutional neural networks (CNNs). CNNs have shown strong performance in image classification and have been widely applied to vibration and acoustic signals after converting them into time–frequency images such as spectrograms and scalograms. CNNs works fine with the images but there were some limitations too. CNNs operate through local receptive fields, meaning they primarily focus on nearby signal regions.This makes it difficult to capture long-range dependencies and global context, which are often important in complex industrial signals and they usually miss the important features that effects the result. These limitations degrades the performance in noisy data.

To address this limitation, Vision Transformers (ViTs) have emerged as a promising alternative. Vision Transformers (ViTs) have recently become a strong alternative to CNNs. Instead of focusing only on small local regions, ViTs use a self-attention mechanism that looks at the entire input at once. This means they can learn both nearby details and long-range patterns across the whole signal. Because of this, ViTs are better at capturing global context and subtle patterns that may be spread over time and frequency. After they were introduced, ViTs quickly became successful in image recognition and are now being used in many other applications.

In this work, we apply ViTs to the University of Ottawa Rolling-element Dataset – Vibration and Acoustic Faults under Constant Load and Speed (UORED-VAFCLS). This dataset provides both accelerometer and acoustic recordings under different fault conditions. To prepare the data for ViT models, we convert raw time-series signals into time–frequency images. Accelerometer signals are transformed into scalograms using a segment size of 2048 samples (approximately 49 ms per window), while acoustic signals are transformed into spectrograms. These representations are then formatted as three-channel RGB images and resized to 224×224 pixels, matching the standard input size for Vision Transformers. Separate ViT models are trained on vibration and acoustic modalities to enable modality-specific feature learning and performance comparison.

The main contributions of this paper are summarized as follows:

- We propose a Vision Transformer-based framework for machinery fault detection using vibration and acoustic data.
- We design a preprocessing pipeline that converts raw accelerometer and acoustic signals into scalograms and spectrograms, formatted as RGB images of size 224×224 for ViT input.

- We train separate ViT models for vibration and acoustic data to allow modality-specific evaluation and comparison.
- We validate our approach on the UORED-VAFCLS dataset and show that ViTs provide accurate and robust fault detection, outperforming traditional CNN-based methods reported in prior work.

The rest of this paper is organized as follows. Section II reviews related work in machine fault detection and deep learning. Section III presents the proposed methodology. Section IV describes the experimental setup and results. Finally, Section V concludes the paper and discusses future directions.

II. Background on Vision Transformers

Before presenting our implementation, we briefly explain how Vision Transformers (ViTs) work internally. Unlike CNNs, which rely on local receptive fields, ViTs treat an image as a sequence of patches and process them using self-attention, similar to how Transformers process words in text.

A. Patch and Positional Embeddings

Each input image of size $224 \times 224 \times 3$ is divided into 16×16 patches, resulting in 196 patches. Each patch is flattened and projected into a lower-dimensional vector called a patch embedding. Since Transformers do not inherently understand spatial order, positional embeddings are added element-wise to the patch embeddings, preserving spatial structure.

B. Multi-Head Self-Attention (MHSA)

The combined embeddings are passed into self-attention layers. Each patch embedding is projected into three vectors:

$$Q = \text{Query}, \quad K = \text{Key}, \quad V = \text{Value}$$

The attention mechanism compares Q with K to compute similarity scores, which are converted into probabilities using softmax. If two patches are similar, the probability will be higher, meaning the model focuses more on that relationship. The weighted probabilities are then applied to V, allowing each patch to learn not only from itself but also from all other patches. This enables the ViT to capture both local and global dependencies across the image.

C. Add & Norm with Residual Connections

After attention, an Add & Norm block is applied. Residual connections add the input back to the output, stabilizing training. Layer normalization ensures stable gradients:

$$\text{Norm}(x) = \gamma \cdot \frac{(x - \mu)}{\sigma} + \beta$$

where μ and σ are mean and standard deviation, and γ and β are learnable parameters.

D. Classification Token

A special [CLS] token is appended at the start of the sequence. After multiple layers of attention and normalization, this token becomes the global representation of the image. Finally, it is passed through a fully connected softmax layer to predict the class (healthy vs. faulty conditions).

III. Methodology

A. Dataset Overview

We used the University of Ottawa Rolling-element Dataset – Vibration and Acoustic Faults under Constant Load and Speed conditions (UORED-VAFCLS). Four fault types are collected: inner race, outer race, ball, and cage (five of each type). A total of 20 bearings are tested, where each bearing has three states of data collection: healthy, developing fault, and faulty. As a result, 60 distinct sets of data are included. For all cases, data is collected at a sampling rate of 42,000 Hz for a total 10 seconds per set of data.

The dataset includes both accelerometer and acoustic sensor recordings under different fault conditions, collected under constant load and speed. The accelerometer was mounted to capture vibration signals, while the acoustic sensor recorded sound patterns, providing two complementary data sources for machine fault detection.

The dataset can be accessed here: https://data.mendeley.com/datasets/y2px5tg92h/5.

TABLE I: UORED-VAFCLS Dataset Summary (Raw CSV Files)

Fault Condition	**Number of CSV Files**
Healthy	20
Inner Race Faults	10
Outer Race Faults	10
Ball Faults	10
Cage Faults	10
Total Files	**60**
Recording Length	10 seconds per file
Sampling Rate	42,000 Hz
Sensors	Accelerometer and Acoustic

Note: The original dataset includes more files and pre-generated spectrogram images. In this work, we only used the raw CSV sensor data listed above and generated our own scalograms and spectrograms during preprocessing.

B. Preprocessing

The raw accelerometer and acoustic signals were transformed into time–frequency images before being used in the Vision Transformer. For vibration signals, we generated scalograms using the Continuous Wavelet Transform (CWT), while acoustic signals were converted into spectrograms using the Short-Time Fourier Transform (STFT). Both were then formatted as three-channel RGB images and resized to 224×224 pixels for model input 1.

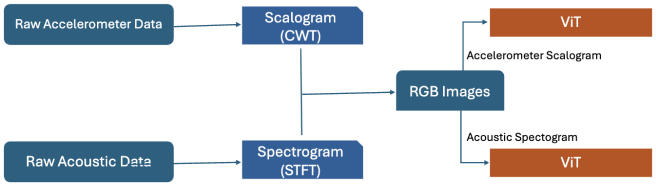


Fig. 1: Preprocessing pipeline: raw signals converted into scalograms and spectrograms for ViT input.

C. Model Architecture (Our Implementation)

Our Vision Transformer (ViT) workflow for fault classification can be explained step by step, with visual examples of how input signals are processed inside the model:

1) **Patch Splitting:** Each input image ($224 \times 224 \times 3$) is divided into 16×16 patches, producing 196 patches per image. Figure 2 shows an accelerometer scalogram divided into patches.

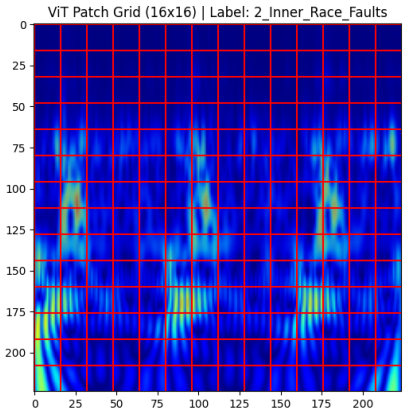


Fig. 2: Input image divided into 16×16 patches for ViT processing.

2) **Patch Embeddings:** Each patch is flattened and projected into a lower-dimensional embedding vector. Positional encodings are then added to preserve spatial order. Figure 3 compares sinusoidal and learned positional embeddings.

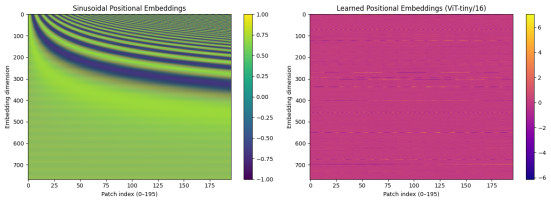


Fig. 3: Comparison of sinusoidal and learned positional embeddings.

3) **Self-Attention Mechanism:** The sequence of patch embeddings is processed using multi-head self-attention (MHSA). This allows the model to capture relationships between patches. Figure 4 shows example attention maps from different heads.
4) **Attention Heatmaps:** The [CLS] token learns a global representation of the entire image by attending to all patches. Figure 5 shows how attention is distributed across scalogram patches.
5) **Patch Importance:** Some patches contribute more strongly to the classification decision. Figure 6 highlights these important patches in the scalogram that the ViT attends to most. Figure 7 further shows how the self-attention mechanism focuses on these fault-specific patches, refining their contribution to the [CLS] token.

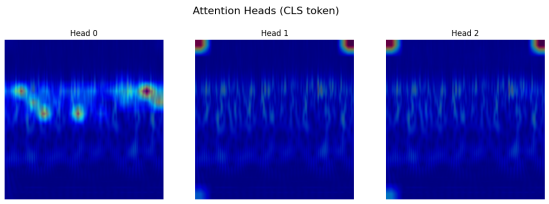


Fig. 4: Attention maps from different heads focusing on fault-related regions.

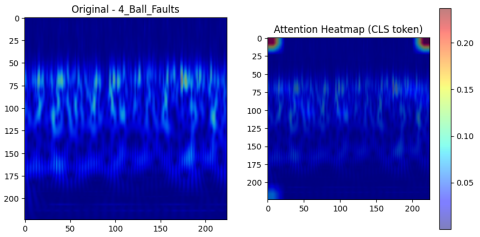


Fig. 5: Attention heatmap from the [CLS] token capturing global fault features.

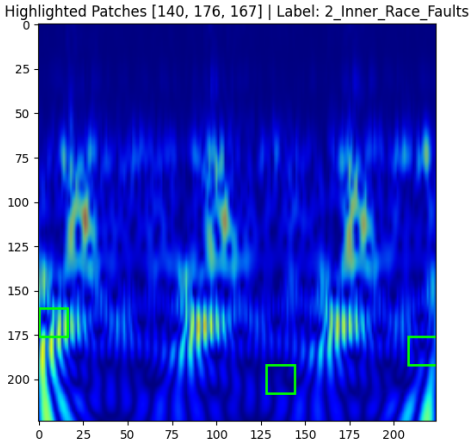


Fig. 6: Highlighted patches contributing most to fault detection.

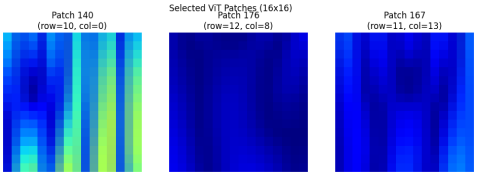


Fig. 7: Self-attention focusing on the fault-related patches highlighted in Fig. 6.

In this work, we trained two separate ViT models: one using scalograms from accelerometer signals, and another using spectrograms from acoustic signals. This allowed us to directly compare vibration-based and acoustic-based fault diagnosis using the same model architecture.

D. Training Setup

The preprocessed scalogram and spectrogram images were organized into class-based folders and loaded using the `ImageFolder` utility from PyTorch. To ensure balanced evaluation, the dataset was split into 80% training and 20% validation using stratified sampling to preserve class distribution.

1) Data Augmentation: To improve generalization and reduce overfitting, data augmentation techniques were applied to the training images. Each input image was first resized to 224×224 pixels. For the training set, random cropping was applied with a scaling factor between 0.9 and 1.0 and an aspect ratio of 1.0. This created slight variations in the image while keeping the main time–frequency patterns intact. Normalization was then applied using the standard ImageNet statistics (mean = (0.485, 0.456, 0.406) and standard deviation = (0.229, 0.224, 0.225)).

Random cropping allowed the model to view the same image from slightly different angles or regions. This helped the Vision Transformer focus on the key fault-related patterns instead of memorizing exact pixel positions. By learning from these variations, the model became more flexible and could handle signals that were not perfectly aligned.

For the validation set, only resizing and normalization were applied to ensure consistency and unbiased evaluation. These augmentation strategies helped the Vision Transformer models learn stronger and more general features from the data.

TABLE II: Data Augmentation and Loader Configuration

Operation	**Training Set**	**Validation Set**
Resize	224×224	224×224
Random Crop	Scale (0.9–1.0), Ratio = 1.0	–
Normalization Mean	(0.485, 0.456, 0.406)	(0.485, 0.456, 0.406)
Normalization Std	(0.229, 0.224, 0.225)	(0.229, 0.224, 0.225)
Shuffling	Yes	No
Batch Size	32	32

IV. Conversion of Time-Series Signals to Visuals

Vision Transformers work with images, so our first step was to turn the raw sensor signals into visual form. The dataset provides vibration data from accelerometers and sound data from microphones in CSV files. Since these are one-dimensional signals, we needed to convert them into two-dimensional images.

For Accelerometer vibration signals, we used **scalograms**, created with the Continuous Wavelet Transform (CWT). Scalograms are good for vibration data because they can show sudden changes (short transients) in both time and frequency. Each signal was divided into segments of 2048 samples (about 49 ms) before being converted. Figure 8 illustrates the principle behind CWT, where a sine wave is combined with a Gaussian envelope to form a Morlet wavelet, enabling localized time–frequency analysis.

For acoustic signals, we used **spectrograms**, created with the Short-Time Fourier Transform (STFT). Spectrograms show how the frequency of sound changes over time, which helps in identifying machine faults from noise patterns.

Sample Raw signal data representation, Accelerometer: -0.186744, 0.844698, 2.069536, 3.358839, and so on, and Acoustic: 0.078973, 0.078973, 0.078973, 0.079631, and so on. These raw signals were converter into visuals as shown in Figure 9.

This process allowed us to use powerful image-based models on raw sensor data.

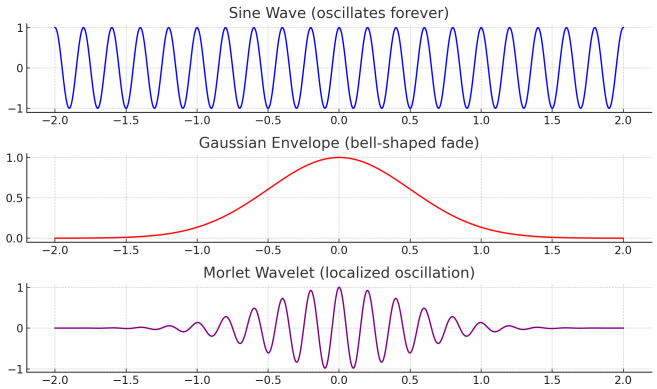


Fig. 8: Morlet wavelet principle used in Continuous Wavelet Transform (CWT).

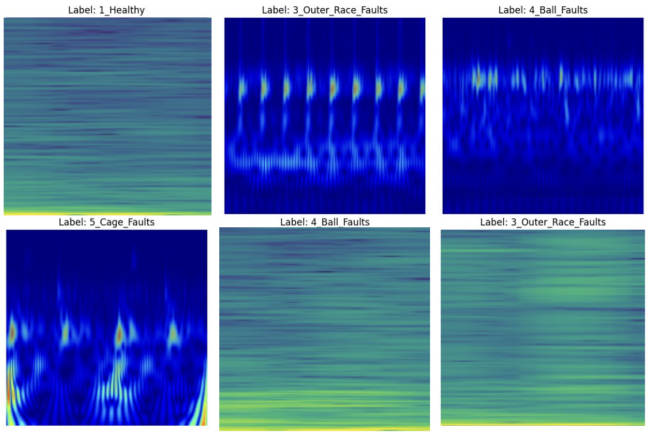


Fig. 9: Examples of time–frequency images generated from accelerometer and acoustic signals. The blue-colored plots represent scalograms (generated using CWT) from vibration signals, while the green/yellow plots represent spectrograms (generated using STFT) from audio signals. Both representations were RGB images of size 224×224 and used as inputs for the Vision Transformer.

V. Model Selection and Processing

A. Model Variants

For this project, we used pretrained Vision Transformers (ViTs) instead of training from scratch. Pretrained models are already trained on large datasets (e.g., ImageNet), which makes them more efficient to adapt to new tasks with limited data. This approach is more practical than building a model from the ground up, especially when computational resources are limited.

Different variations of Vision Transformers are available, trained with different numbers of parameters and layers.

- **ViT-Tiny (5M parameters)** — lightweight model with fewer layers and attention heads, faster to train and suitable for small datasets.
- **ViT-Small (22M parameters)** — provides a good balance between accuracy and computational requirements.
- **ViT-Base (85M parameters)** — deeper and more powerful model, but requires very large datasets and long training times on high-end GPUs.

TABLE III: Comparison of Vision Transformer Variants

Variant	Parameters	Layers	Heads
ViT-Tiny	∼5M	12	3
ViT-Small	∼22M	12	6
ViT-Base	∼85M	12	12

The key decision was selecting the most appropriate model for our dataset and resources. Although ViT-Base offers higher capacity, it demands significant computational power and large-scale data, which were not available in this project. After experimentation, we selected ViT-Tiny and ViT-Small because they are more suited for smaller datasets and limited GPU resources. Despite their smaller size, these models provided strong accuracy for machine fault detection, proving that lightweight ViTs can perform effectively when data is limited.

To sumerize ViT-Base was avoided because it requires much larger training data and longer training times.

B. Dataset Setup - Splitting

To train and evaluate the models, the dataset was split into 70% training, 15% validation, and 15% testing. This split ensured fair evaluation while the validation set helped prevent overfitting during training. The final testing set was kept completely unseen until evaluation, providing a reliable measure of model performance.

VI. Implementation and Workflow

In our implementation, we used a pretrained Vision Transformer from the `timm` library instead of training from scratch. This leverages transfer learning, where the model has already learned useful features from large-scale image datasets.

A. Implementation Details

The model used in this work was a pretrained Vision Transformer (ViT-Tiny) from the `timm` library. It was customized for our classification task by setting the number of output classes to match the dataset labels.

```
self.model = timm.create_model(
    "vit_tiny_patch16_224",
    pretrained=True,
    num_classes=num_classes
)
```

Listing 1: Defining pretrained ViT-Tiny model

For optimization, we used the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 0.05.

```
optimizer = torch.optim.AdamW(
    self.parameters(),
    lr=1e-4,
    weight_decay=0.05
)
```

Listing 2: Optimizer configuration

To ensure reproducibility and track experiments, we used PyTorch Lightning's logging and checkpointing utilities. The best-performing model was automatically saved based on validation accuracy.

```
logger = CSVLogger("logs", name="vit_model")
checkpoint_callback = ModelCheckpoint(
    dirpath="/content/drive/MyDrive/
        ViTs_audio_checkpoint",
    filename="vit-{epoch:02d}-{val_acc:.4f}",
    monitor="val_acc",
    mode="max",
    save_top_k=1
)
```

Listing 3: Logger and checkpoint setup

The training loop was managed by PyTorch Lightning's `Trainer`, set to run for 5 epochs with mixed precision enabled when GPU resources were available.

```
trainer = pl.Trainer(
    max_epochs=5,
    accelerator="auto",
    devices=1,
    precision=16 if torch.cuda.is_available() else
        32,
    logger=logger,
    callbacks=[checkpoint_callback]
)
trainer.fit(model, train_loader, val_loader)
```

Listing 4: Trainer setup

B. Workflow Summary

The steps of our workflow are shown in Fig. 10. We start with raw accelerometer and acoustic signals, convert them into time–frequency images (scalograms and spectrograms), preprocess them into RGB format (224×224), and then feed them into a pretrained Vision Transformer (ViT). Finally, the model classifies the machine condition as either healthy or faulty.

VII. Results and Discussion

A. Why Vision Transformers over CNNs?

To better understand the effectiveness of Vision Transformers (ViTs), we also trained a standard 2D Convolutional Neural

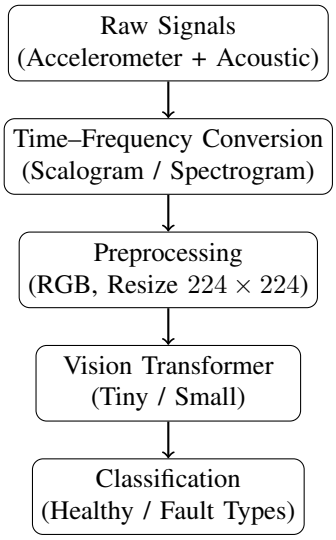


Fig. 10: Workflow: Raw signals → Time–frequency images → Preprocessing → ViT model → Classification.

Network (CNN) on the same dataset, with the same input size and number of epochs. The comparison highlighted key differences:

- **Training Time:** Both CNN and ViT models required approximately the same training time per epoch under identical settings.
- **Accuracy:** The CNN achieved a test accuracy of around ∼68%, while its training accuracy reached ∼99.23% (as shown in Fig. 11). This result indicates that CNNs can fit the training data very well but may suffer from overfitting, leading to weaker generalization on unseen test data.
- **Performance of ViTs:** In contrast, the ViT models achieved more than ∼95% test accuracy under the same settings and number of epochs.
- **Reason for Improvement:** Unlike CNNs, which focus mainly on local features through convolution filters, ViTs use self-attention to capture both local and global patterns across the entire image.

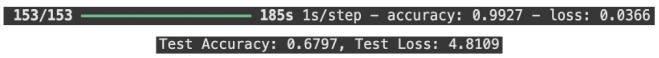
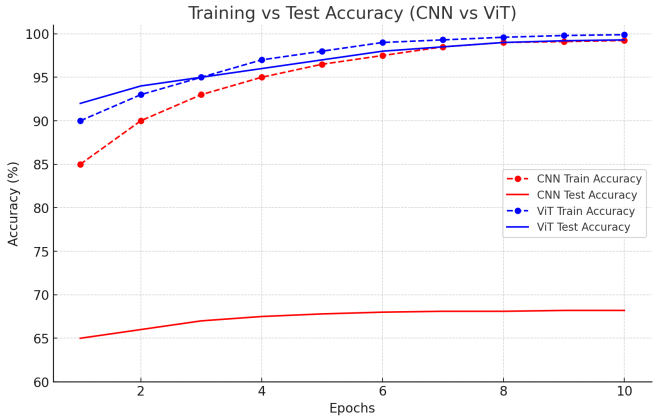


Fig. 11: CNN training results showing high training accuracy (∼99.23%) but much lower test accuracy (∼68%), indicating overfitting.

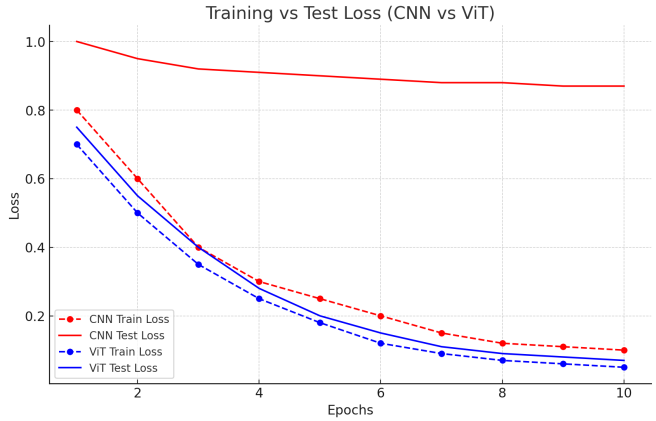
Although CNNs provided decent accuracy and remain widely used in many domains, our observations show that ViTs clearly outperform CNNs for machine fault detection when using time–frequency images.

B. Overall Comparison

The CNN baseline achieved a test accuracy of ∼68% while its training accuracy reached ∼99.2%, indicating significant overfitting. In contrast, ViT models demonstrated excellent generalization with ∼99.9% accuracy for accelerometer data



(a) Training vs Test Accuracy for CNN and ViT.



(b) Training vs Test Loss for CNN and ViT.

Fig. 12: Comparison of CNN and ViT in terms of accuracy (a) and loss (b) across epochs. The CNN shows clear overfitting, while ViTs generalize better.

TABLE IV: Performance comparison between CNN and ViT models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN (2D Conv)	68.20	70.10	67.50	68.80
ViT (Accelerometer)	99.89	99.90	99.88	99.89
ViT (Acoustic)	98.79	98.88	98.43	98.64

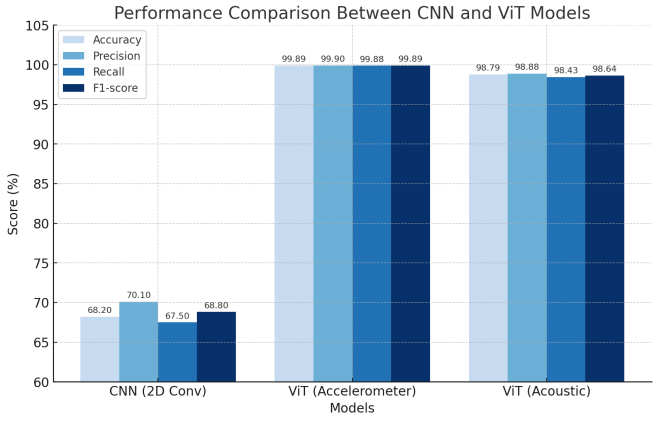


Fig. 13: Performance comparison between CNN and ViT models in terms of Accuracy, Precision, Recall, and F1-score.

and ∼99.0% for acoustic data, consider Figure 13. These results highlight the effectiveness of self-attention in capturing both local and global patterns, which CNNs often miss.

C. Attention Visualization on Acoustic Data

To better understand how Vision Transformers process acoustic spectrogram inputs, we visualized the similarity of patches, attention heads, and heatmaps. These visualizations show how the model distributes its focus across different time–frequency regions.

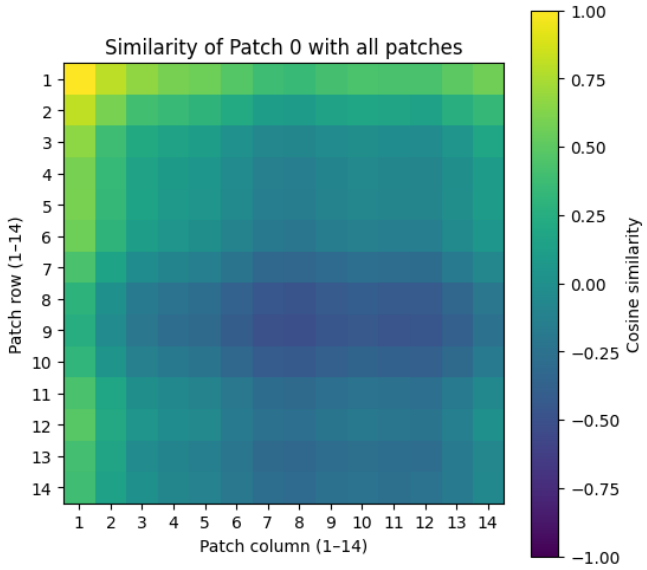


Fig. 14: Cosine similarity of Patch 0 with all other patches in the spectrogram input.

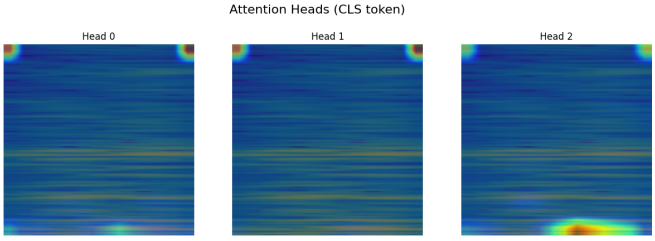


Fig. 15: Attention maps from different heads focusing on acoustic spectrogram regions.

D. Model Predictions on Test Set

To further validate the effectiveness of the ViT model, we visualized predictions on a subset of the test data. Figure 17 shows examples of scalograms with both predicted and true class labels.

As observed, the model correctly identifies most fault categories such as *Cage Faults*, *Outer Race Faults*, *Ball Faults*, and *Healthy* states. Even in cases where the signals are visually similar, the ViT effectively captures discriminative features, leading to accurate classification. This demonstrates the strong generalization ability of transformer-based architectures for fault detection tasks.

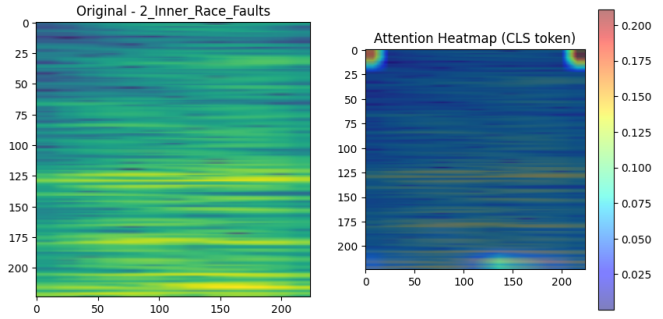


Fig. 16: Attention heatmap from the [CLS] token highlighting fault-related acoustic patterns.

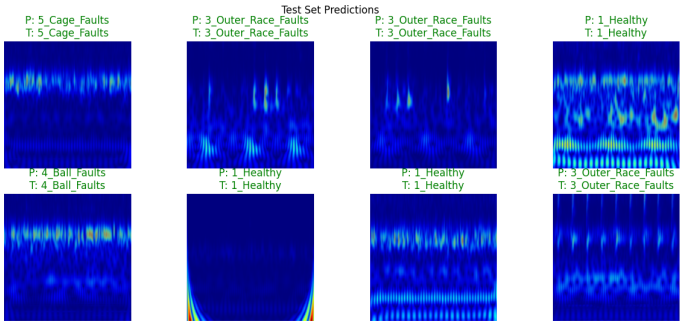


Fig. 17: Sample predictions from the ViT model on the test set. Each subfigure shows the predicted label (P) and true label (T). The majority of predictions align correctly with the ground truth, illustrating the robustness of ViTs in fault classification.

VIII. Related Work

Vaswani et al. introduced the Transformer architecture in their work *Attention Is All You Need* [1]. This paper presented the concept of self-attention, which replaced recurrence and convolution for sequence modeling and became the foundation for modern Transformer-based models.

Dosovitskiy et al. later proposed the Vision Transformer (ViT) in their paper *An Image is Worth 16x16 Words* [2]. Their study demonstrated that images can be divided into patches and processed as sequences using self-attention. This allowed ViTs to capture both local and global features effectively, achieving strong performance in image recognition tasks.

Sehri et al. published the *University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets* [3]. This open dataset provides accelerometer and acoustic measurements of rolling-element bearings and is widely used to train and evaluate machine learning models for fault detection.

Before the introduction of ViTs, Convolutional Neural Networks (CNNs) were the most widely used models for machinery fault detection. Li et al. [4] and Zhang et al. [5] demonstrated that CNNs can learn useful local patterns from

time–frequency images of vibration signals. However, CNNs often suffer from overfitting and fail to capture long-range dependencies.

To address these limitations, Wu et al. [6] applied Vision Transformers to fault diagnosis and showed that ViTs outperform CNNs when dealing with complex time–frequency data. In this work, we provide a direct comparison between CNNs and ViTs on the same dataset and highlight why ViTs achieve better generalization.

IX. Conclusion and Future Work

In this paper, we compared Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for machine fault detection using scalograms and spectrograms. We found that CNNs performed well on the training data but did not generalize well to new test data, showing signs of overfitting. On the other hand, ViTs achieved very high accuracy on both training and test data, proving that they can learn both local and global patterns better than CNNs.

For the future, there are a few directions to improve this work. First, the experiments were done under fixed load and speed. Testing the models on variable operating conditions would make the results more realistic. Second, combining ViTs with other models such as CNNs or LSTMs could further improve performance. Finally, deploying ViTs in real-time monitoring systems in industries would help to test their practical usefulness.

Overall, this study shows that Vision Transformers are a strong alternative to CNNs for fault detection and can play an important role in building reliable predictive maintenance systems.

Acknowledgment

This work was carried out under the guidance of **Dr. Sandeep Singh Sandha**, founder of **Punjab AI Excellence**. We thank him for his valuable support, and also our peers for their helpful discussions and feedback.

This work was carried out using the University of Ottawa bearing vibration and acoustic fault datasets, along with open-source tools such as TensorFlow, PyTorch, NumPy, and Matplotlib. We also acknowledge the use of OpenAI's ChatGPT for assistance with coding, improving understanding, and helping with writing and clarity in this paper.

References

[1] A. Vaswani *et al.*, "Attention Is All You Need," *NeurIPS*, 2017.

[2] A. Dosovitskiy *et al.*, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

[3] M. Sehri, P. Dumond, and M. Bouchard, "University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets," *Data in Brief*, vol. 49, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340923004456 :contentReference[oaicite:1]index=1

[4] X. Li, W. Zhang, and Q. Ding, "Deep learning-based machinery fault diagnostics using time–frequency image representation," *Mech. Syst. Signal Process.*, vol. 108, pp. 385–404, 2019.

[5] W. Zhang, C. Li, and G. Peng, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, 2020.

[6] J. Wu, H. Chen, and Y. Zhao, "Fault diagnosis of rotating machinery using Vision Transformers," *IEEE Access*, vol. 9, pp. 112280–112290, 2021.