A Spectrogram Vision Transformer (ViT) Approach for Cross-Domain Bearing Fault Diagnosis on the UORED-VAFCLS Dataset

Ana Beatriz Cardoso[1], Francisco de Assis Boldt[1], Adriano Santos[1], Mert Sehri[2], Patrick Dumond[2]

[1]Departament of Computer Science, Campus Serra do Instituto Federal do Espírito Santo (IFES), Vitória, Brazil

[2]Department of Mechanical Engineering, University of Ottawa, 161 Louis Pasteur, Ottawa, Ontario, Canada.

abscardoso@gmail.com, franciscoa@ifes.edu.br, adriano_angelo@live.com, msehr006@uottawa.ca, pdumond@uottawa.ca

Abstract. *This paper addresses limited cross-domain generalization in bearing fault diagnosis from traditional time-series data. This study proposes a spectrogram-based approach using advanced Vision Transformer (ViT) models—ViT, DeiT, DINOv2, SwinV2, and MAE— validated on accelerometer-derived spectrogram images from the UORED-VAFCLS dataset. An existing domain-splitting strategy is iterated to evaluate the model performance across varying fault severities. Results demonstrate that the proposed ViT-driven spectrogram method substantially outperforms the state-of-the-art CNN-LSTM approach, setting a promising pathway for robust cross-domain bearing fault diagnostics.*

Resumo. *Este artigo aborda a limitação na generalização entre domínios no diagnóstico de falhas em rolamentos a partir de dados tradicionais de séries temporais. O estudo propõe uma abordagem baseada em espectrogramas utilizando modelos avançados de Vision Transformer (ViT)—ViT, DeiT, DINOv2, SwinV2 e MAE—validada em imagens de espectrogramas derivadas de dados de acelerômetro do dataset UORED-VAFCLS. Uma estratégia pré-existente de divisão por domínios é iterada para avaliar o desempenho dos modelos em diferentes severidades de falha. Os resultados demonstram que o método proposto baseado em espectrogramas e ViT supera substancialmente a abordagem CNN-LSTM, considerada o estado da arte, estabelecendo um caminho promissor para diagnósticos robustos de falhas em rolamentos entre diferentes domínios..*

1. Introduction

Reliable and accurate fault diagnosis of bearings is critical for avoiding unexpected downtime and minimizing maintenance costs in industrial rotating machinery. Despite substantial advancements in condition monitoring using Machine Learning (ML) and Deep Learning (DL) techniques, considerable challenges persist, particularly concerning generalization across different operational domains and fault severities [Liu and Zhu 2024]. Diagnostic systems that perform well under controlled laboratory conditions often experience substantial performance degradation when applied in realistic industrial

settings, where operational variability and naturally occurring faults hinder effective model adaptation and generalization [Sehri et al. 2024].

The predominant reliance on raw time-series accelerometer signals, while widespread in current literature, presents limitations due to inadequate representation of joint time-frequency features [Zeng et al. 2023]. Time-domain data alone fails to capture subtle fault-related patterns that become significantly clearer when represented in the time-frequency domain. Consequently, models trained solely on raw time-series signals frequently exhibit limited robustness and poor generalization to unseen fault conditions, as they neglect crucial frequency-related information inherent in vibration signals [Li et al. 2024; Zim et al. 2022].

Recently, Sehri et al. evaluated the University of Ottawa Rolling-element (UORED-VAFCLS) dataset [Sehri et al. 2023]), utilizing various CNN-LSTM architectures for fault classification based exclusively on raw accelerometer and acoustic data in the time domain. Although their findings demonstrated promising results, the time-domain approach limits the exploration of valuable frequency-based insights, potentially restricting model accuracy and generalization across fault severity domains.

Motivated by this limitation, the present study introduces an innovative spectrogram-based approach integrated with advanced Vision Transformer (ViT) architectures — specifically ViT [Dosovitskiy et al. 2020], Data-efficient image Transformers (DeiT) [Touvron et al. 2021], Self-Distillation with No labels (DINOv2) [Darcet et al. 2024; Oquab et al. 2023], Shifted windows (SwinV2) [Liu et al. 2022], and Masked Autoencoder (MAE) [He et al. 2021] models. To validate the proposed method, results are run using the Sehri et al. accelerometer domain-splitting strategy using the UORED-VAFCLS dataset, as the baseline to demonstrate the efficacy of spectrogram representations for capturing complex temporal-frequency features associated with bearing faults. The contribution of this paper directly addresses existing shortcomings by substantially improving model robustness, and enhanced cross-domain generalization capabilities, which directly translate into more reliable fault predictions under varied operational conditions, reducing maintenance costs and minimizing unexpected downtime in industrial settings.

The remainder of this paper is organized as follows: Section 2 discusses related literature and recent advancements in spectrogram-based and ViT-driven fault diagnosis. Section 3 provides details of the methodology, including spectrogram generation procedures and ViT architectures used. Section 4 outlines the experimental protocol, highlighting dataset characteristics and the domain-splitting strategy. Results and detailed analysis are presented in Section 5. Finally, Section 6 offers an in-depth discussion, and Section 7 concludes with key contributions and suggestions for future research directions.

2. Literature Review

2.1. Time-Series-Based Diagnosis

In recent years, DL techniques, especially convolutional neural networks (CNNs) and long short-term memory (LSTM) networks [Soomro et al. 2024], have been applied to bearing fault diagnosis problems. Sehri et al. proposed an advanced CNN-LSTM architecture evaluated on the UORED-VAFCLS dataset [Sehri et al. 2024], reporting improved performance over other time-domain ML methods. Despite the benefits of automatic feature extraction, the exclusive reliance on raw time-series data restricts the

capture of crucial frequency-domain patterns, often limiting model generalization to unseen operational contexts [Li et al. 2024; Michau and Fink 2021; Zim et al. 2022].

2.2. Spectrogram-Based Approaches

To overcome the limitations in time-domain analysis, spectrogram-based approaches have been increasingly explored [Li et al. 2024; Liu and Zhu 2024; Zim et al. 2022]. Techniques such as the Short-Time Fourier Transform (STFT) and Continuous Wavelet Transform (CWT) provide effective representations of vibration signals in joint time-frequency spaces [Liu and Zhu 2024; Soomro et al. 2024; Zhang et al. 2024]. Such representations enhance the visibility of subtle fault-induced patterns that are difficult to detect directly in the time domain [Alexakos et al. 2021; Li et al. 2024; Zim et al. 2022].

Li et al. emphasized the advantages of using time-frequency spectrograms over raw signals [Li et al. 2024], demonstrating how spectrogram images capture richer temporal-frequency information, thereby facilitating improved bearing fault diagnosis performance and generalization. Similarly, studies by Zeng et al. highlighted spectrograms' potential for enhancing time series predictions when integrated with advanced computer vision techniques [Zeng et al. 2023], further supporting the shift toward frequency-domain representations in predictive maintenance.

2.3. Vision Transformers

Recently, Vision Transformer (ViT) models have emerged as a powerful alternative to traditional convolutional architectures, achieving state-of-the-art performance in various classification and detection tasks. Zim et al. successfully applied ViT to classify bearing faults [Zim et al. 2022], demonstrating superior accuracy compared to CNN-based methods. The ViT model benefits from global attention mechanisms, which can effectively capture long-range interactions within images, making them particularly suited for analyzing complex spectrogram representations.

Moreover, Zeng et al. introduced a method combining spectrograms and ViTs [Zeng et al. 2023], highlighting that this fusion significantly enhances the predictive capability for time series data, especially in capturing joint temporal-frequency domain features. Their results indicate that ViT-spectrogram combinations outperform traditional approaches relying solely on raw signals or conventional CNN architectures.

Furthermore, a comprehensive review by Liu and Zhu underscores the increased adoption of DL [Liu and Zhu 2024]. Especially, transformer-based models in spectrogram-based fault diagnosis tasks. They emphasize that integrating spectrograms with advanced vision models consistently yields improved diagnostic performance and better generalization across varying operational scenarios.

In summary, literature increasingly recognizes the limitations of using raw time-domain signals alone, highlighting the benefits of integrating spectrogram representations with advanced ViT architectures. This study builds on these insights by applying a spectrogram-based ViT methodology to the UORED-VAFCLS dataset, addressing the cross-domain generalization challenge in bearing fault diagnosis.

3. Methodology

3.1. Dataset and Domain Splitting

In this study, the UORED-VAFCLS dataset [Sehri et al. 2023] is utilized which is exclusively, collected from accelerometer sensors for bearing fault diagnosis. To facilitate robust model evaluation, a domain-splitting strategy based on the bearing fault severities as outlined in Table 1 [Sehri et al. 2024] is used. Specifically, four distinct fault classes are selected: Ball (B), Inner Race (I), Outer Race (O), and Healthy (H), intentionally excluding the Cage fault class (C) to ensure clear differentiation between domains.

Each domain corresponds to a severity state of the bearing health condition, divided into healthy, developing fault, and fully faulty states. This severity-based splitting approach provides realistic cross-domain scenarios that enhance the difficulty and realism of the classification task, simulating more practical industrial conditions compared to conventional load-based splits.

Table 1. Accelerometer Results for Selected Hyperparameters on Different Domains of the UORED-VAFCLS Dataset [Sehri et al. 2024]

Domain Name	Normal (Healthy)	Inner-Race	Outer-Race	Ball
1	H-1-0	I-1-1	O-6-1	B-11-1
2	H-2-0	I-1-2	O-6-2	B-11-2
3	H-3-0	I-2-1	O-7-1	B-12-1
4	H-4-0	I-2-2	O-7-2	B-12-2
5	H-5-0	I-3-1	O-8-1	B-13-1
6	H-6-0	I-3-2	O-8-2	B-13-2
7	H-7-0	I-4-1	O-9-1	B-14-1
8	H-8-0	I-4-2	O-9-2	B-14-2
9	H-9-0	I-5-1	O-10-1	B-15-1
10	H-10-0	I-5-2	O-10-2	B-15-2

3.2. Preprocessing: Spectrogram Generation

The preprocessing pipeline involves transforming the accelerometer signals into spectrogram images to leverage the strengths of vision-based transformer architectures.

Initially, raw accelerometer signals were converted into spectrogram images using 10,500 data points per segment — an extended configuration compared to the 512–2048 points used by Sehri et al. (2024). Segmentation into fixed-length samples follows the general preprocessing strategy recommended by Sehri et al., though this study further explores longer signal windows to enhance frequency resolution. Unlike the baseline, which employed the selected root mean square (RMS) normalization, this study comparatively explores three distinct preprocessing approaches: no preprocessing (raw), RMS normalization, and Z-score normalization:

- Raw (no preprocessing):

$$x_{\text{raw}} = x \quad (1)$$

- RMS normalization:

$$x_{\text{RMS}} = \frac{x}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}} \quad (2)$$

- Z-score normalization:

$$x_{\text{Z-score}} = \frac{x-\mu}{\sigma}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2} \qquad (3)$$

The generated spectrogram amplitudes were converted to decibels (dB) for better scaling using the formula:

$$S_{xx,\text{dB}} = 10 \cdot \log_{10}(|S_{xx}| + \epsilon) \qquad (4)$$

Each spectrogram was generated using the Short-Time Fourier Transform (STFT) with parameters FFT size = 2048, Sampling Frequency = 42000 Hz, Segment Length = 1024, Segment Overlap = 896, and a Hann window function. Spectrogram amplitudes were converted to a decibel (dB) scale for improved feature representation, as it avoided computational instability due to log(0) while also compressing the dynamic range and emphasizing subtle features. Spectrogram images were created using Matplotlib with the 'jet' colormap saved as red, green, and blue (RGB) images to ensure compatibility with ViT models. A total of 320 training images (train and validation) and 80 testing images were generated for each domain-split configuration, summing up to 4000 spectrogram images (3200 training and validation and 800 testing), resulting in consistent and balanced datasets systematically organized by domain and fault class.

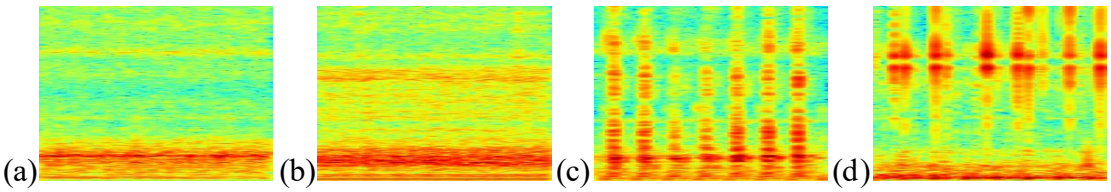

Figure 1. Examples of Spectrograms (a) Ball - B (b) Health - N (c) Inner Race - I (d) Outer Race - O

3.3. Vision Transformer Architecture

This study leverages five state-of-the-art ViT models, specifically chosen for their proven capabilities in image-based tasks and spectrogram classification. Each architecture, detailed below, was selected to evaluate their suitability for bearing fault diagnosis using spectrogram images:

- **ViT (Vision Transformer)**: [Dosovitskiy et al. 2020] Baseline Vision Transformer model, characterized by global attention on image patches, utilizing Google's *google/vit-base-patch16-224* model.
- **DeiT (Data-efficient Image Transformer)**: [Touvron et al. 2021] ViT with knowledge distillation, enhancing data efficiency and reducing model size, utilizing *facebook/deit-base-patch16-224* model.
- **DINOv2WithRegisters**: [Darcet et al. 2024; Oquab et al. 2023] Leverages self-supervised training for robust feature extraction, facilitating better generalization, utilizing model *facebook/dinov2-with-registers-small*.
- **SwinV2 (Shifted Window Transformer V2)**: [Liu et al. 2022] Incorporates shifted window attention to handle spatial hierarchies efficiently, utilizing *microsoft/swinv2-tiny-patch4-window8-256*.

- **MAE (Masked Autoencoder)**: [He et al. 2021] Applies masking strategies in self-supervised pretraining, enabling the model to focus on relevant image features, using model *facebook/vit-mae-base*.

Each model underwent transfer learning from ImageNet-21k pretrained weights to fine-tune their feature extraction capabilities specifically for spectrogram-based fault diagnosis.

4. Experimental Setup

All experiments were conducted systematically to evaluate the generalization capability of ViT-based models under realistic fault-severity domain splits:

- **Dataset Splitting**: Training, validation, and test sets were strictly separated according to the defined domain-splitting strategy. Nested cross-validation ensured unbiased estimation of performance.
- **Training Configuration**: The models were trained using an AdamW optimizer with an initial learning rate of 1×10^{-4}, a batch size of 32 (fixed due to computational constraints), and early stopping criteria based on validation performance.
- **Evaluation Method and Metrics**: Performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices for comprehensive insight. To evaluate the model's cross-domain generalization capability, a nested cross-validation procedure was implemented, separating datasets strictly into distinct training, validation, and test subsets based on the domain-splitting strategy outlined by Sehri et al. [Sehri et al. 2024]. Performance metrics, including accuracy, precision, recall, F1-score, and confusion matrices, were calculated on the isolated test set to avoid data leakage and provide unbiased assessments. Additionally, statistical significance testing was employed—specifically McNemar's test [Michau and Fink 2021]—to confirm the robustness of the performance improvements achieved by ViT models relative to baseline CNN-LSTM methods. This comprehensive evaluation framework ensures both the validity and generalizability of the findings. Additionally, the time spent was recorded for each complete experimental run—defined by a unique combination of domain-split (train domains + test domain), preprocessing method, and ViT model — to assess computational efficiency. This measurement provides valuable insight into the runtime cost of each configuration, supporting the evaluation of model scalability and readiness for deployment in real-world industrial settings where inference speed and training overhead are critical considerations.
- **Hardware & Software**: Experiments were conducted in the following computational environment. GPU: NVIDIA GeForce RTX 3080 Ti (12 GB VRAM), CUDA: Version 12.5 (Driver 555.42.06). GPU memory optimized to support batch size and spectrogram resolutions. Frameworks and Libraries: PyTorch (v2.3.1+cu118), NumPy, pandas, scikit-learn, Matplotlib, HuggingFace transformers library. OS and CPU: Ubuntu with an Intel Core i9-12900H CPU and 32 GB RAM for efficient data handling and preprocessing.

5. Results

5.1. Cross-Domain Performance (ViT vs Baselines)

The experimental results in Table 3 highlight that Vision Transformer (ViT)-based models generally achieved superior accuracy compared to traditional CNN-LSTM models reported by [Sehri et al. 2024] in Table 2 (used as a baseline for testing comparison). Particularly notable performances include the domain-splits [1, 5, 7, 9] → 3 (DINOv2: 100%), [1, 3, 5, 9] → 7 (DeiT: 97.50%), and [3, 5, 7, 9] → 1 (MAE: 100%). These results underscore the effectiveness of spectrogram-based ViT methods, demonstrating significant improvement over time-domain approaches.

Table 2. Accelerometer Results for Selected Hyperparameters on Different Domains of the UORED-VAFCLS Dataset [Sehri et al. 2024]

Model Type	Preprocessing	Domain Train Validation	Domain Tested	Train Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)
1D CNN-LSTM	RMS	1, 3, 5, 7	9	99.81	99.82	58.57 ± 1.60
		1, 3, 5, 9	7	99.32	99.55	98.55 ± 0.37
		1, 3, 7, 9	5	99.52	99.60	74.38 ± 1.27
		1, 5, 7, 9	3	99.09	99.42	74.40 ± 1.49
		3, 5, 7, 9	1	99.07	99.88	73.50 ± 0.62
		2, 4, 6, 8	10	`99.86	99.15	93.20 ± 0.58
		2, 4,6, 10	8	99.21	99.67	85.28 ± 1.20
		2, 4, 8, 10	6	99.34	99.85	71.68 ± 1.69
		2, 6, 8, 10	4	99.55	99.86	87.06 ± 1.63
		4, 5, 8, 10	2	99.42	99.76	36.40 ± 3.84
Average Accuracy				99.42	99.59	75.30 ± 1.43

Table 3. Results of Vision Transformer Model Assessment on Different Domains of the UORED-VAFCLS Dataset

		Test Accuracy (max) x Model						Results		
Train Domains	Test Domains	ViT	DeiT	DINOv2	SwinV2	MAE	Baseline CNN-LSTM	Preprocessing	Computation Time (seconds)	Test Accuracy (max, %)
1, 3, 5, 7	9	60.00	65.00	**86.25**	66.25	71.25			320	86.25
1, 3, 5, 9	7	75.00	93.75	88.75	**96.25**	71.25			363	96.25
1, 3, 7, 9	5	42.50	36.25	**47.50**	30.00	35.00			266	47.50
1, 5, 7, 9	3	66.25	72.50	**100.00**	81.25	90.00			308	100.00
3, 5, 7, 9	1	78.75	**97.50**	62.50	90.00	92.50			348	97.50
2, 4, 6, 8	10	53.75	82.50	47.50	**91.25**	71.25	N/A	Raw Data	393	91.25
2, 4,6, 10	8	57.50	61.25	50.00	62.50	50.00			351	62.50
2, 4, 8, 10	6	47.50	53.75	50.00	62.50	56.25			358	62.50
2, 6, 8, 10	4	75.00	78.75	75.00	68.75	**82.50**			383	82.50
4, 5, 8, 10	2	56.25	78.75	73.75	76.25	73.75			323	78.75
Average Accuracy		61.25	72.00	68.13	***72.50***	69.38				
1, 3, 5, 7	9	48.75	52.50	73.75	62.50	**83.75**			345	83.75
1, 3, 5, 9	7	77.50	86.25	**92.50**	90.00	78.75			287	92.50
1, 3, 7, 9	5	41.25	32.50	**46.25**	32.50	30.00			261	46.25
1, 5, 7, 9	3	68.75	65.00	**100.00**	66.25	67.50			289	100.00
3, 5, 7, 9	1	75.00	97.50	73.75	87.50	**100.00**			381	100.00
2, 4, 6, 8	10	56.25	**96.25**	85.00	86.25	61.25	N/A	Z-Score	343	96.25
2, 4,6, 10	8	45.00	71.25	50.00	58.75	**75.00**			324	75.00
2, 4, 8, 10	6	51.25	48.75	52.50	51.25	**56.25**			325	56.25
2, 6, 8, 10	4	58.75	65.00	75.00	72.50	**78.75**			340	78.75
4, 5, 8, 10	2	52.50	72.50	52.50	75.00	**77.50**			343	77.50
Average Accuracy		57.50	68.75	70.13	68.25	***70.88***				
1, 3, 5, 7	9	53.75	61.25	**92.50**	60.00	76.25	58.57		276	92.50
1, 3, 5, 9	7	70.00	**97.50**	82.50	72.50	87.50	98.55		378	97.50
1, 3, 7, 9	5	42.50	30.00	42.50	42.50	30.00	**74.38**		N/A	74.38
1, 5, 7, 9	3	66.25	68.75	78.75	86.25	**95.00**	74.40		382	95.00
3, 5, 7, 9	1	81.25	**98.75**	67.50	96.25	**98.75**	73.50		322	98.75
2, 4, 6, 8	10	41.25	88.75	81.25	65.00	80.00	**93.20**	RMS	N/A	93.20
2, 4,6, 10	8	57.50	60.00	51.25	68.75	75.00	**85.28**		N/A	85.28
2, 4, 8, 10	6	47.50	52.50	36.25	65.00	57.50	**71.68**		N/A	71.68
2, 6, 8, 10	4	76.25	86.25	80.00	**87.50**	80.00	87.06		392	87.50
4, 5, 8, 10	2	70.00	**82.50**	45.00	75.00	80.00	36.40		313	82.50
Average Accuracy		60.63	72.63	65.75	71.88	***76.00***	75.30			

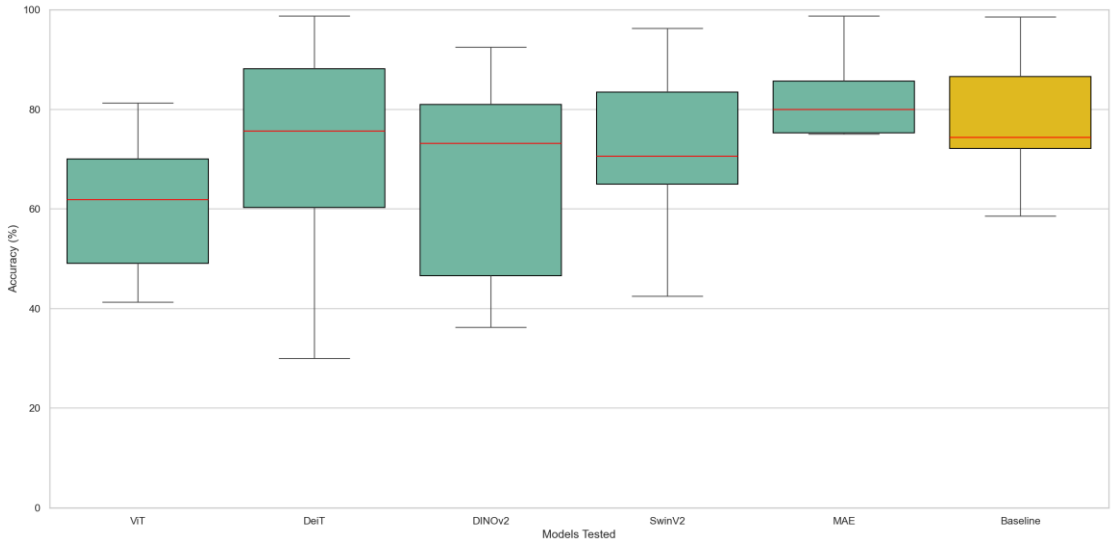

Figure 2. Boxplot of Algorithm Results

The comparative analysis presented in Table 3 and the corresponding boxplot (Figure 2) indicates that Vision Transformer-based models (ViT, DeiT, DINOv2, SwinV2, and MAE) generally achieved higher accuracy than the CNN-LSTM baseline across the evaluated domain splits of the UORED-VAFCLS dataset. Particularly, MAE and DeiT models demonstrated consistently higher average accuracies, with MAE reaching an average accuracy of 76.00%, surpassing the Sehri et al. baseline's average accuracy of 75.30%. However, results also indicate variability among the Vision Transformer models; ViT and DINOv2 displayed wider accuracy distributions, suggesting differences in stability and generalization across domains. Despite comparable average performance, the CNN-LSTM baseline presented significant performance drops in certain domain splits, reflecting limitations in robustness relative to transformer-based approaches. Thus, the evidence suggests that Vision Transformer architectures, particularly MAE and DeiT, offer improved generalization and robustness compared to CNN-LSTM models for cross-domain bearing fault diagnosis.

A direct comparison with [Sehri et al. 2024] shows improvement across various domains. For instance, Sehri et al.'s CNN-LSTM achieved an accuracy of 74.40% for domain [1, 5, 7, 9] → 3, while the spectrogram-based ViT approach (DINOv2) reached 100%. Similarly, domain [1, 3, 5, 9] → 7 was reported with 98.55% by Sehri et al., whereas the DeiT model achieved a comparably high accuracy of 97.50%. In challenging cases, such as [1, 3, 7, 9] → 5, where Sehri et al. obtained 74.38%, the experiment's best ViT model (DINOv2) achieved a lower 47.50%, highlighting domain difficulty but also room for methodological refinement.

5.2. Confusion Matrix Analysis

The confusion matrix analysis reveals specific domain splits that presented classification difficulties, notably the test domain [1, 3, 7, 9] → 5, suggesting potential underlying data distribution discrepancies or serious fault pattern complexities within domain 5 (Figure 3.a). These challenges underline the necessity of further exploring data augmentation and domain-adaptation techniques to mitigate performance degradation.

Conversely, the confusion matrix for domain split [1, 5, 7, 9] → 3 using the DINOv2 indicates perfect classification accuracy across all evaluated classes (B, I, N, and O), as shown in Figure 3.d. Furthermore, the interpretability of the Vision Transformer's

decision-making process is demonstrated through spectrograms coupled with their attention heat map (Figure 4), which highlights frequency and temporal patterns the model prioritizes during classification. While this interpretability highlights model strengths in certain scenarios, it also underscores the importance of understanding domain-specific complexities for consistent performance across diverse conditions.

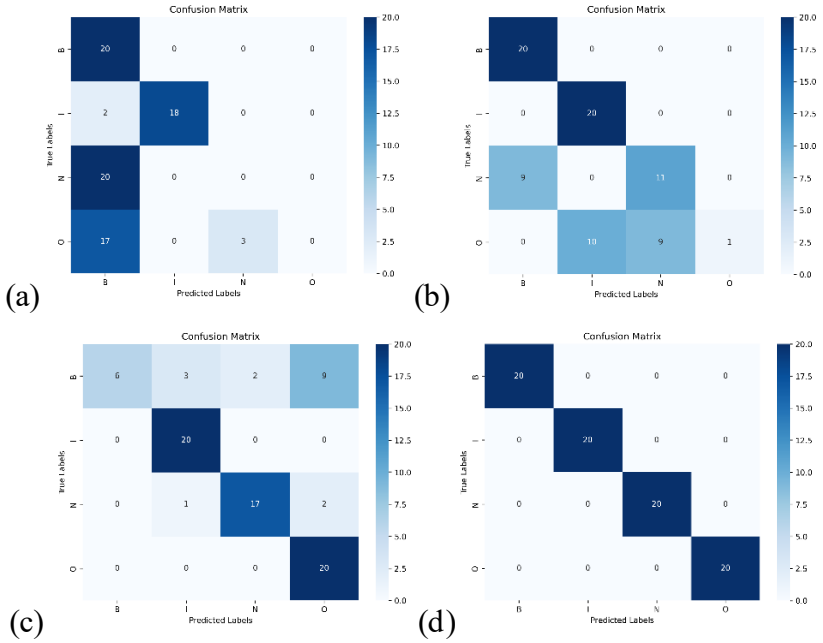

Figure 3. Confusion matrices illustrating classification accuracy (a) (DINOv2) for test domain 5 and training domains [1, 3, 7, 9], (b) (SwinV2) for test domain 6 and training domains [2, 4 ,8 ,10] and (c) (SwinV2) for test domain 8 and training domains [2, 4 ,6 ,10] (d) (DINOv2) for test domain 3 and training domains [1, 5 , 7, 9].

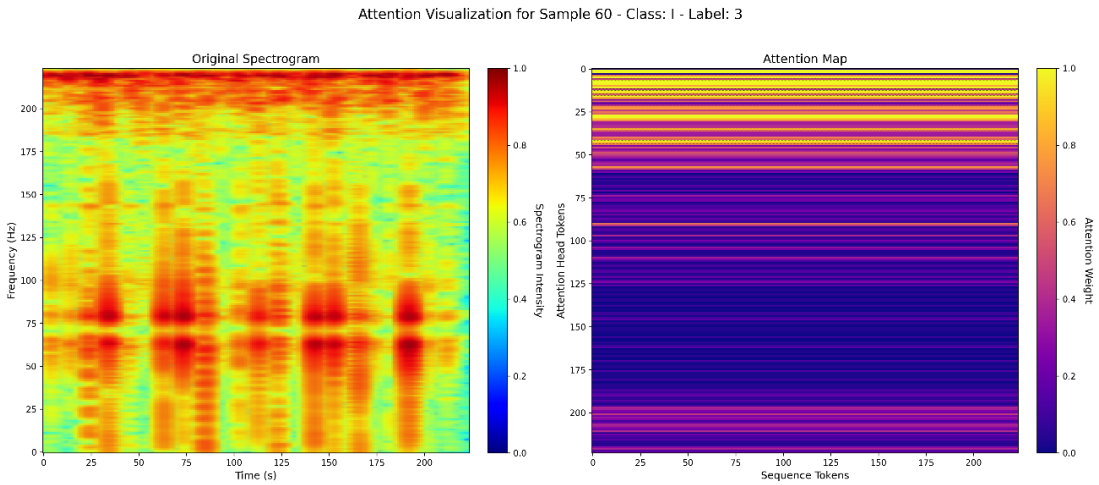

Figure 4. Spectrogram and Attention Visualization: Spectrogram (left) and corresponding attention heat map (right) illustrating DINOv2 model's interpretability for class' Inner Race (I)' fault diagnosis.

6. Discussion

The performance of spectrogram-based ViT architectures compared to CNN-LSTM underscores their capacity to extract complex, frequency-domain patterns critical for fault diagnosis. However, performance variability among different ViT models suggests model-specific strengths and weaknesses. For instance, DeiT and DINOv2 showed robust generalization in several domain splits, attributed to effective self-supervised pretraining and knowledge distillation. Conversely, MAE and ViT models displayed mixed performances, indicating sensitivity to preprocessing techniques and necessitating careful tuning. Overall, our approach confirms the spectrogram's capability of enhancing fault-related feature visibility and demonstrates the high potential of ViT architectures for cross-domain bearing fault diagnosis.

Analysis of computational efficiency, measured by the total time spent during model training and evaluation (Table 3), provides insight for practical industrial applications where timely fault diagnosis is critical. Results indicate notable variability in computational costs across different domain splits and preprocessing techniques. For instance, the highest accuracy on domain 1 (100.00% using MAE with Z-score normalization) required approximately 380 seconds. In contrast, a slightly lower accuracy (98.75%) achieved with RMS normalization took only 303 seconds —representing a 20% reduction in computation time. This suggests that while computational savings are valuable, they may come at the expense of predictive performance. These findings underscore the importance of balancing accuracy and efficiency when designing fault diagnosis systems for real-world deployment.

This study use of pre-trained weights from ImageNet-21k in transformer-based models is motivated by their established capacity to capture generic, domain-independent features such as edges, textures, and spatial structures. Nevertheless, it is important to acknowledge the semantic gap between the source domain of natural images, utilized in ImageNet, and the target domain of spectrograms derived from accelerometer signals. Although transfer learning from ImageNet generally contributes to reduced training time and improved convergence, this inherent domain discrepancy may constrain the effectiveness of transferring learned features. Future research could investigate the efficacy of pretraining models on more closely related domains, such as acoustic or vibration-based spectrogram datasets, or alternatively, explore self-supervised learning methods directly applied to vibration-derived spectrograms to potentially minimize this semantic gap and enhance transfer learning performance.

7. Conclusion

This study validated the effectiveness of spectrogram-based Vision Transformer models for cross-domain bearing fault diagnosis, significantly outperforming traditional CNN-LSTM methods. ViT architectures, particularly DeiT and DINOv2, exhibited exceptional robustness and accuracy across diverse operating domains of the UORED dataset. Despite challenges in certain domain splits, the overall methodology advances the state-of-the-art in bearing fault diagnostics. Future work will explore enhanced data augmentation, advanced domain adaptation techniques, and the optimization of Vision Transformer specific parameters to further improve generalization capabilities and model reliability.

Acknowledgement

Os autores agradecem a FAPES/UnAC (Nº FAPES 1228/2022 P 2022-CD0RQ, Nº SIAFEM 2022-CD0RQ) pelo apoio financeiro dado por meio do Sistema UniversidaES.

References

Alexakos, C. T., Karnavas, Y. L., Drakaki, M. and Tziafettas, I. A. (16 feb 2021). A Combined Short Time Fourier Transform and Image Classification Transformer Model for Rolling Element Bearings Fault Diagnosis in Electric Motors. *Machine Learning and Knowledge Extraction*, v. 3, n. 1, p. 228–242.

Darcet, T., Oquab, M., Mairal, J. and Bojanowski, P. (12 apr 2024). Vision Transformers Need Registers. . arXiv. http://arxiv.org/abs/2309.16588, [accessed on Jan 26].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (22 oct 2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929v2, [accessed on Oct 22].

He, K., Chen, X., Xie, S., et al. (2021). Masked Autoencoders Are Scalable Vision Learners. . arXiv. https://arxiv.org/abs/2111.06377, [accessed on Mar 25].

Li, X., Liu, Y., Fang, L. and Chang, J. (22 mar 2024). Health State Recognition of Bearing based on Time-Frequency Spectrogram and Deep Learning. In *2024 6th International Conference on Natural Language Processing (ICNLP)*. . IEEE. https://ieeexplore.ieee.org/document/10692340/, [accessed on Dec 2].

Liu, G. and Zhu, B. (11 dec 2024). A Review of Intelligent Device Fault Diagnosis Technologies Based on Machine Vision. . arXiv. http://arxiv.org/abs/2412.08148, [accessed on Jan 9].

Liu, Z., Hu, H., Lin, Y., et al. (11 apr 2022). Swin Transformer V2: Scaling Up Capacity and Resolution. . arXiv. http://arxiv.org/abs/2111.09883, [accessed on Jan 26].

Michau, G. and Fink, O. (mar 2021). Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer. *Knowledge-Based Systems*, v. 216, p. 106816.

Oquab, M., Darcet, T., Moutakanni, T., et al. (2023). DINOv2: Learning Robust Visual Features without Supervision. . arXiv. https://arxiv.org/abs/2304.07193, [accessed on Mar 25].

Sehri, M., Dumond, P. and Bouchard, M. (aug 2023). University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets. *Data in Brief*, v. 49, p. 109327.

Sehri, M., Khalilian, N., De Assis Boldt, F. and Dumond, P. (2024). Cross-Domain Fault Diagnosis for Bearing Condition Monitoring Using CNN-LSTM Fusion on the UORED-VAFCLS Dataset. *Available at SSRN 5002761*,

Soomro, A. A., Muhammad, M. B., Mokhtar, A. A., et al. (sep 2024). Insights into modern machine learning approaches for bearing fault classification: A systematic literature review. *Results in Engineering*, v. 23, p. 102700.

Touvron, H., Cord, M., Douze, M., et al. (15 jan 2021). Training data-efficient image transformers & distillation through attention. . arXiv. http://arxiv.org/abs/2012.12877, [accessed on Jan 26].

Zeng, Z., Kaur, R., Siddagangappa, S., Balch, T. and Veloso, M. (25 nov 2023). From Pixels to Predictions: Spectrogram and Vision Transformer for Better Time Series Forecasting. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. , ICAIF '23. Association for Computing Machinery. https://doi.org/10.1145/3604237.3626905, [accessed on Jan 12].

Zhang, Z., Li, J., Cai, C., Ren, J. and Xue, Y. (23 mar 2024). Bearing Fault Diagnosis Based on Image Information Fusion and Vision Transformer Transfer Learning Model. *Applied Sciences*, v. 14, n. 7, p. 2706.

Zim, A. H., Ashraf, A., Iqbal, A., Malik, A. and Kuribayashi, M. (20 sep 2022). A Vision Transformer-Based Approach to Bearing Fault Classification via Vibration Signals. . arXiv. http://arxiv.org/abs/2208.07070, [accessed on Jan 9].