Real-Time Vibration Fault Detection in Rotating Machines Using Transformers to Minimize Production Losses in Industry 5.0: VIBT

Toukap Nono Fernand Joseph[1]*, Tokoue Ngatcha Dianorré[1] , Offole Florence [2] , Nyatte Steyve[3] and Mouzong Pemi Marcelin[4]

[1] Laboratory of Mechatronics, Energiatronics and Sustainable Mobility (LaMEMD), National Higher polytechnic School of Douala, University of Douala.
[2] Laboratory of Mechanics and Materials (LaMM), National Higher polytechnic School of Douala, University of Douala.
[3] Laboratory of technology and applied Sciences, University of Douala, Cameroun
[4] Higher Technical Teacher's Training College, Depaartment of Renewable energy, University of Buea, Cameroun.

*Corresponding author's email: toukap_nono@enspd-udo.cm

Abstract. Quickly identifying anomalies in rotating machinery is crucial for safety and profitability in contemporary industry (Industry 5.0). Unidentified failures can cause costly malfunctions and production interruptions. This research proposes an innovative strategy based on Transformer for the analysis of multidimensional vibration events (VIBT), with a view to early and accurate detection of anomalies in rotating machinery. The goal is to minimize production interruptions in Industry 5.0. The study highlights the limitations of conventional vibration analysis approaches and traditional deep learning techniques, emphasizing the need for innovative solutions. VIBT incorporates transformers and a filter bank convolution (FBC) module for effective denoising, as well as an adaptive wavelet transformation (WTA) mechanism for dynamic feature fusion at various scales, thereby addressing the challenges posed by non-stationary and noisy signals. Extensive testing on the Mafaulda dataset reveals that VIBT achieves 98.1% precision and 98.8% accuracy, significantly outperforming existing standard models. The results suggest that VIBT not only improves fault detection capabilities but also optimizes maintenance strategies in industrial applications, paving the way for future research on semi-supervised learning based on transformers and the integration of intermodal data.

Keywords: Anomaly detection, Faults, Prediction, Precision, Transformers, Vibration.

1. Introduction

The development of diagnostic and anomaly detection technologies, combined with artificial intelligence, has promoted the adoption of data-driven methodologies for predicting failures in rotating machinery [1]. Vibration analysis, although effective, requires better integration of AI [2], [3]. Failures result in high costs [4], [5]. Deep learning (DL) technologies show promising performance but depend

on large amounts of data and present variability challenges [6]. Solutions, such as convolutional neural networks, are emerging, but feature extraction issues remain [7], [8]Signal processing techniques, such as wavelet transform and EMD, have limitations [9]. DL models, such as CNNs and LSTMs, struggle with long sequences [10].

We propose VIBT, a transformer-based method for analyzing non-stationary vibration data, incorporating a convolutional filtering block (CFB) for denoising and a wavelet attention mechanism (WTA). This model achieves 98.1% and 98.8% accuracy [11]. We use the Mafaulda dataset to enhance the robustness of the model [12], [13]. The article also addresses the challenges of DL models [14], [15] and sound analysis to reveal anomalies [16], [17].

We evaluate VIBT with precision, accuracy, and recall. The article presents similar research (section 2), the method (section 3), the results (section 4), and the conclusions (section 5), with references (section 6).

2. Related Work

Spot inspections and measurements, although essential for detecting faults in rotating machinery, are time-consuming and do not provide continuous coverage [18]. Traditional and image recognition methods, although effective, remain costly and sensitive to environmental factors. Distributed acoustic sensors (DAS) offer a viable alternative with signal processing techniques such as wavelet transform and empirical mode decomposition (EMD), but they often suffer from limitations [10].

Deep learning models, such as LLM4TS and ParInfoGPT, have improved anomaly detection in DAS time series [15], [17]. However, CNNs struggle with long sequences, and RNNs can lack long-term context. Hybrid techniques combining wavelets and LSTMs have shown advantages [19].

Recent research has applied LLMs to time series analysis, such as the Voice2Series model by Yang et al. [20] and the unified framework by Zhou et al. [21]. Chang et al. [17] introduced LLM4TS, while Hegselmann et al. [22] proposed an LLM for tabular data.

Classic transformers may not capture non-stationary fluctuations or suppress noise, but they offer powerful modeling. Our solution, VIBT, fills this gap by integrating wavelet attention modules and filter bank convolution into a Transformer encoder for real-time identification of deformation hazards.

3. Methods

The methods used to create the model are described here. To facilitate understanding of the relationships between the steps in the methodology, we have included a diagram of the steps in Figure 1 to illustrate this. Then, as shown in Figure 2 - 4 we use transformers to identify the type of fault and predict the state of the machine.

3.1 Data description

The vibration data in our study comes from the public Mafaulda dataset used in [23]. This dataset comprises 1,951 multivariate time series representing six machine conditions: normal operation, horizontal misalignment, vertical misalignment, imbalance, overhang, and bearing failure. The data were acquired with three unbalanced loads: 6(g), 20(g), and 35(g).

The dataset exhibits an imbalance in the distribution of samples; for example, the “Bearing” class has 558 samples, while horizontal misalignment has 197. This imbalance was taken into account when interpreting overall performance, hence the use of accuracy and complementary metrics for a more robust evaluation. Each file must contain complete loads without being too large.

Table 1: Defects in the data sets used

Faults	Measurements
Horizontal misalignment	197

Vertical misalignment	301
Unbalance	333
Overhangs	513
Bearings	558

For the training and evaluation of our model, the total dataset was split into a training set and a test set, with a proportion of 70% for training and 30% for testing, respectively. This split is illustrated in Table 2.

Table 2: Breakdown of the data set (Training/Testing)

Type of set	Percentage	Number of samples
Training	70 %	1366
Test	30 %	585
Total	100 %	1951

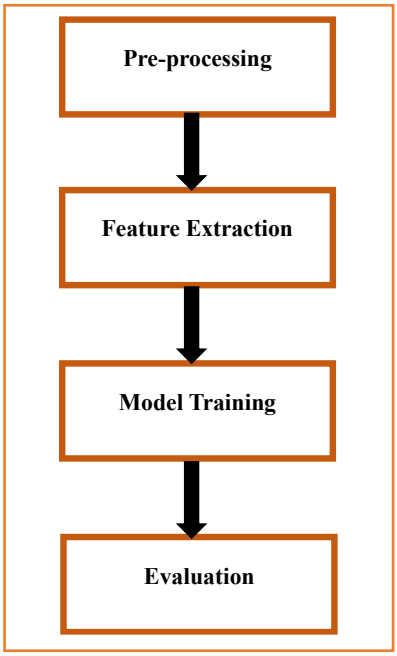

Figure. 1. flowchart of the methodology steps.

3.1.1 Pre-processing and Normalization

Preprocessing was performed using Python's NumPy to load CSV datasets and visualize vibration signals in order to identify anomalies and ensure the correct association of sound classes and fault labels. We removed extreme values that could skew the analysis and applied a band-stop filter to reduce noise.

Normalization was then performed by scaling the data within a defined range, using min-max normalization based on the minimum and maximum values of each feature, as shown in Equation 1.

$$X_n = \frac{X - Min(X)}{Max(X) - Min(X)} \quad (1)$$

- X : normalize. Is the vector of raw data;
- Xn: is a vector in the form of an array; Represents the vector of normalised data.
- Min(X): it is the lowest value of the vector X;
- Max(X): it is the highest value of the vector X;

3.1.2 Signal segmentation

Vibration signals can have a fixed or variable length, and sliding windows overlap to capture transitions between states. Segmentation is defined by equations (2) and (3), with a window length and a sliding step determining the distance between windows.For each segment, the start of the window can be calculated as follows:

$$t_i = i \,.\, S \quad (2)$$

For i=0, 1, 2, …, N

Where N is the total number of segments, determined by the relationship:

$$N = \left[\frac{T - L}{S}\right] + 1 \quad (3)$$

- t_i : Represents the start of the i-th segmentation window;
- i : Is the index of the window, varying from 0 to N;
- S : This is the sliding step between the start of two consecutive windows;
- N : Represents the total number of segments;
- T : Designates the total duration of the signal;
- L : Designates the length of the window (duration of each segment);

Windows are formed by segmenting the signal x(t) and then multiplying it by a window w(t), which can be chosen from several types of windows (for example, Hamming, Hanning, etc.). The windowing formula becomes. as illustrated by equation 4:

$$y_i\,(t) = x(t)\,.\,w\,(t - t_i) \quad (4)$$

For $t \in [t_i, t_i + L]$

- $y_i\,(t)$: Represents the windowed signal at time t for the i-th segment;
- $x(t)$: Is the original signal at time t;
- $w\,(t - t_i)$: Represents the function of the applied window, offset by t_i ;
- $t \in [t_i, t_i + L]$: Indicates that time t is within the interval of the i-th segment;

3.2 Feature extraction

This phase prepares the data for classification in two steps: the first converts the temporal data into frequency and time-frequency domains via FFT, and the second uses DFT [24] to reduce dimensionality and identify signal features, with Equation (5) for processing and Equation (6) for reconstruction..

$$X\,[K] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}km} \tag{5}$$

$$x\,[K] = \frac{1}{N}\sum_{k=0}^{N-1} X[k] e^{-j\frac{2\pi}{N}km} \tag{6}$$

- $X\,[K]$: is the DFT coefficient at index k;
- $x[n]$: Is the input signal with index n;
- N : Is the total number of signal samples;
- j : Is the imaginary unit;
- k Varies from o to N-1;
- m : Represents the time or frequency index in the summation;

We perform three types of analyses to generate functionalities in the time-domain, frequency-domain, and time-frequency domain.

3.2.1 Feature selection and Partitioning of the dataset

We perform a correlation analysis to eliminate redundant features, using SelectKBest to select the most relevant ones and PCA to reduce dimensionality. The dataset is divided into two sections: test and training. The test samples evaluate the model, while the training samples train it, with a division of 70% for training and 30% for testing [24].

3.3 Vibration detection transformer

We propose a transformative framework for feature extraction and risk assessment in non-stationary vibration data from cable tunnels. VIBT comprises two modules: FBC for seasonality and WTA for non-stationary patterns.

3.3.1. Filter bank convolution module

Equations (7), (8), and (9) show that the FBC module learns frequency-specific representations of vibration signals using convolutional neural networks and finite impulse response filters. Figure 2 illustrates its structure.

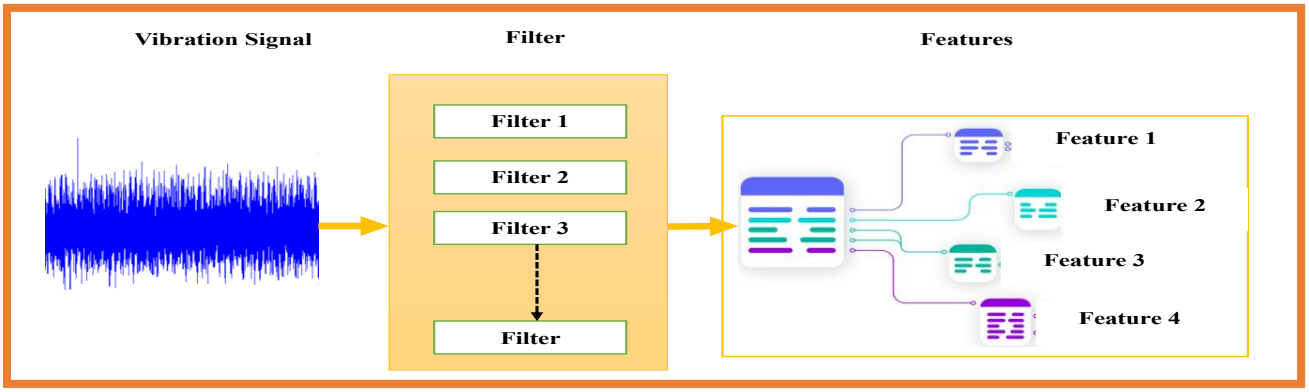

Figure. 2. Structure of FBC module.

The one-dimensional vibration signal $X(t) \in \mathbb{R}^T$ must be the vibration signal input. Composed of K parallel learnable filters, the FBC module

$h_k(t) \in \mathbb{R}^N$. In order to generate a filtered feature sequence, each filter convolves with the input:

$$F_k(t) = (x \times h_k)(t) = \sum_{n=0}^{N-1} h_k(n) \cdot x(t-n), \qquad k = 1, 2, \ldots\ldots\ldots, K \tag{7}$$

A stacked matrix of every *K* channel is the FBC module's output:

$$X_{FBC} = [F_1(t); F_2(t); \ldots\ldots; F_K(t)] \in \mathbb{R}^{K \times T} \tag{8}$$

The filters are set up to evenly span the frequency band $\left[0, \frac{f_s}{2}\right]$ in order to guarantee frequency-domain complementarity. Backpropagation is used to optimize all filter parameters $h_k(n)$ during training.

A regularization term and a task-specific loss L_{task} are included in the total loss function to encourage balanced frequency coverage:

$$L_{FBC} = L_{task} + \lambda \sum_{k=1}^{K} \left(\int |H_k(f)|^2 \, df - \frac{1}{K} C \right)^2 \tag{9}$$

in which λ is a regularization coefficient, C is a constant, and $H_k(f)$ is the frequency response of the kth filter.

3.3.2. Wavelet transform-based adaptive multi-scale attention

Our proposal, WTA, combines the discrete wavelet transform (DWT) with an attention mechanism for dynamic fusion of multi-scale features obtained using equations (10), (11), (12), to effectively represent the non-stationary and time-varying nature of vibration signals. The WTA structure is illustrated in Figure 3.

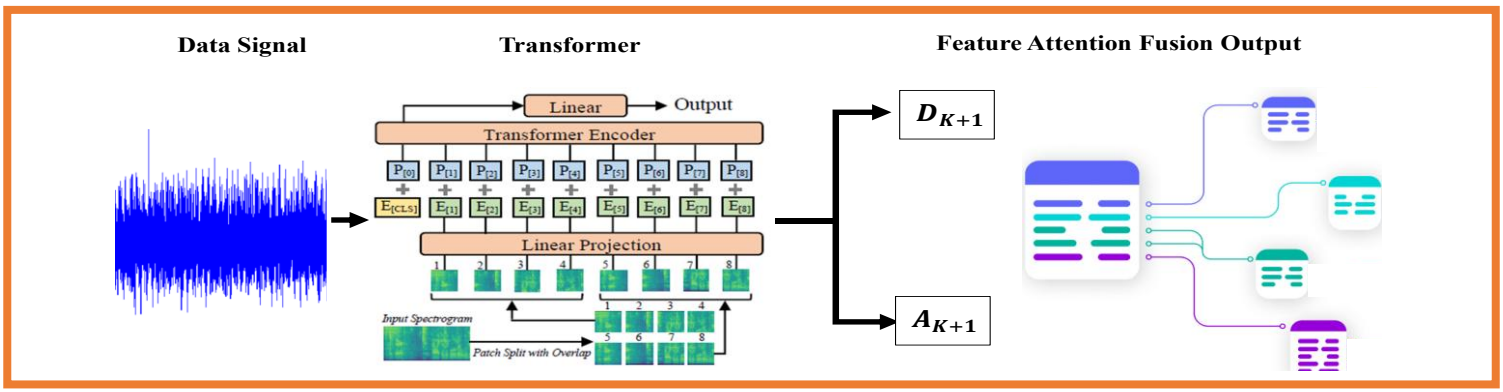

Figure. 3. Structure of WTA module.

Decomposition of the input signal $X(t)$ lieds several detail coefficients across D_j across J levels and one approximation coefficient A_j:

$$X(t) \xrightarrow{DWT} \{A_j, D_j, D_{j-1}, \dots\dots, D_1\} \tag{10}$$

Where A_j represents the coarsest approximation and Dj denotes the detail coefficients at level j.

Using grid-search ablation across {2,3, 4, 5, 6}, we identify the ideal number of levels J. The greatest Score-F1 on the validation set is obtained with J = 3.

A learnable attention network is used to calculate the attention weight α_j for each set of detail coefficients D_j:

$$\alpha_j = \frac{e^{\left(W^T \sigma\left(WD_J + b\right)\right)}}{\sum_{i=1}^{J} e^{\left(W^T \sigma\left(WD_l + b\right)\right)}} \tag{11}$$

Through a weighted sum across scales, the final fused representation is produced:

$$\hat{X}(t) = \sum_{j=1}^{J} \alpha_j D_j\,(t) \tag{12}$$

The nonlinear activation function (such as ReLU or tanh) is represented by $\alpha(.)$, and the learnable parameters are W, w, and b.

In equation (11), we use an attention network to calculate the attention weights, denoted α_e. These weights are essential for weighting the contributions of detail coefficients in the feature fusion process.

- **W** represents the weights of the attention network, which are learned during training. These weights determine the relative importance of each detail coefficient in the final representation of the signal.
- **b** is the bias associated with the weights W, allowing for additional flexibility in model adjustment.
- **σ** denotes a nonlinear activation function, such as ReLU or tanh, which is applied to the outputs of the vector product between W and the detail coefficients. The use of this activation function introduces nonlinearity into the model, which is crucial for capturing complex relationships in the data.

3.4. VIBT model architecture

For vibration anomaly detection and high-resolution feature learning, the VIBT framework combines the FBC and WTA modules into a Transformer encoder. Figure 4 shows the VIBT's construction.

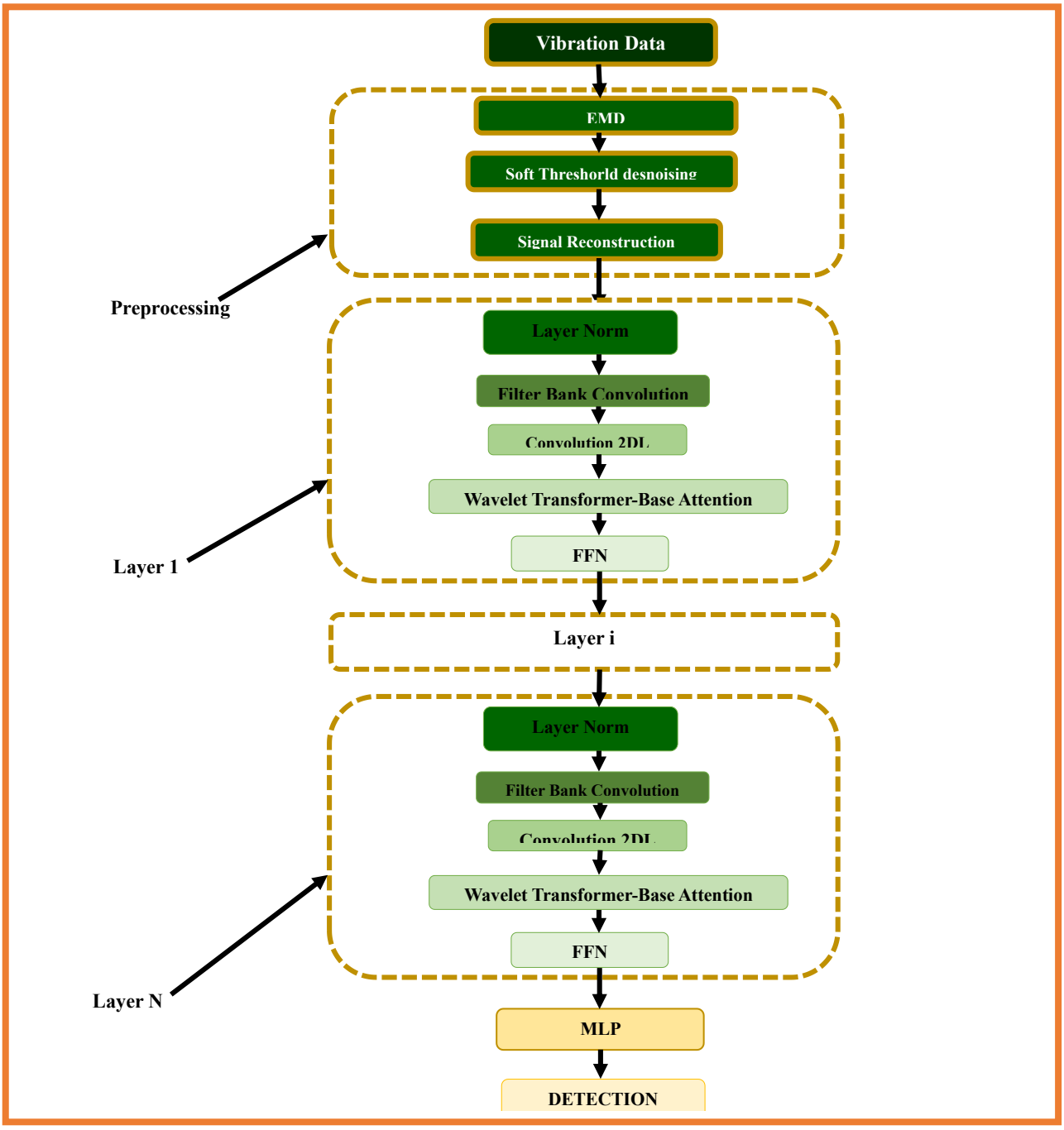

Figure 4. VIBT Structure .

Three primary steps comprise the model:

To create frequency-sensitive multi-channel functions, the FBC module first processes the input signal using equation (13).

$\boldsymbol{X_{input}} \in \mathbb{R}^{\boldsymbol{T}}$.

$$X_{FBC} = FBC(X_{input}) \in \mathbb{R}^{K \times T} \quad (13)$$

equation (14) provides improved temporal representations; the WTA module is applied to each feature channel:

$$\mathrm{X_{WTA}} = \mathrm{WTA(X_{FBC})} \in \mathbb{R}^{\mathrm{K\times T}} \tag{14}$$

equation (15) gives a Transformer encoder, which incorporates positional encoding and residual connections for depth modeling, receives the finite element sequence as input.

$$\mathrm{X_{out}} = \text{Transformer Encoder } (\mathrm{X_{WTA}}) \tag{15}$$

In order to identify or classify vibration-related risks, equation (16) is used, and the output X_{out} is then sent to a classification or regression head:

$$\mathrm{X_{out}} = \text{Transformer Encoder } (\mathrm{X_{WTA}}), \qquad \hat{y} = \mathrm{Head(X_{out})} \tag{16}$$

3.5. Downstream learning module

We test three different downstream learners using the identical feature inputs in order to confirm that the representations learnt by VIBT are capable of generalization. The first step is to implement a two-layer MLP head: we flatten the encoder output $X_{input} \in \mathbb{R}^{K\times T}$ into a vector of dimension D, apply a fully-connected layer with weight $W_1 \in \mathbb{R}^{d_h\times D}$ and bias $b_1 \in \mathbb{R}^{d_h}$, pass the result through a ReLU activation $\sigma(\cdot)$, and then project with $W_1 \in \mathbb{R}^{d_h\times D}$, $b_1 \in \mathbb{R}^{C}$, and softmax to generate class probabilities equation (17):

$$\hat{y} = \mathrm{Softmax(W_2\ \sigma\ (W_1 Flatten(X_{out})\ +\ b_1)\ +\ b_2)} \tag{17}$$

Lastly, we feed the flattened feature vector into a Stochastic Configuration Network (SCN) [25], where the output weights are chosen using least squares and the hidden-layer weights and biases are created at random using a supervisory process. This results in equation (18).

$$\hat{y} = \mathrm{SCN\big(Flatten(X_{out})\big)} \tag{18}$$

SCN offers robust resistance against gradient vanishing along with quick convergence.

3.6. Performance evaluation

Performance measures include precision, accuracy, and recall. Four metrics showed improvements: accuracy, recall, precision, and F1 score (equations 19-22). Recall evaluates the detection of positive results [23, 24, 25], precision evaluates positive predictions, and F1 score provides an overall assessment in cases of class imbalance. These metrics require true positives (TP), false positives (FP), and false negatives (FN) per class.TP: The true-positive of a class is the total number of correct predictions for this labeled class.

FP: The false-positive of a class is the total number of incorrect predictions that predicted this class.

FN: The false-negative of a class is the total number of false predictions for the data labeled in that class.

Accuracy, as indicated in equations (19), measures the proportion of correct predictions out of the total number of observations. It is an overall measure of model performance.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (19)$$

- TP: true positives
- TN: true negatives
- FP: false positives
- FN: false negatives

Precision, calculated using equation (20), is the percentage of true positives. It evaluates the ratio of correctly diagnosed defects. TP refers to detected class I samples, and FP refers to misclassified samples.

$$\text{Pecision} = \frac{TP}{TP+FP} \quad (20)$$

Recall is the ratio of true positives (TP) to actual positive cases, indicating the model's success rate. Combined with precision, recall measures identified defects, while precision evaluates predicted positive cases, as in equation (21).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (21)$$

The F1 score evaluates the performance of a classification model by combining precision and recall, which is essential for reducing false alarms and avoiding costly interruptions. It is calculated using equation (22).

$$\text{Score F1} = \frac{2 \times \text{Precision} \times \text{Sensibilite}}{\text{Precision} + \text{Sensibilite}} \quad (22)$$

AUC measures the model's ability to distinguish between positive and negative classes. A value of 1 indicates perfect separation, while a value of 0.5 indicates a random model.

4. Result and Discussion

This section evaluates the effectiveness of VIBT by analyzing performance on an Intel(R) Core i7-3320M 3 GHz workstation and an INTEL(R) 11 GB graphics card. The optimal hyperparameters are shown in Table 3.

4.1. Dataset construction

The vibration data comes from the Mafaulda dataset [23], with 1951 time series and loads of 6 (g), 20 (g), and 35 (g) [24]. It contains five faults and one normal state, with an accuracy of 98.8% and a recall of 92.4%.

The set shows an imbalance between classes (Table 2). Figure 5 illustrates each defect, while Figures 5 and 6 evaluate the reliability of the machine.

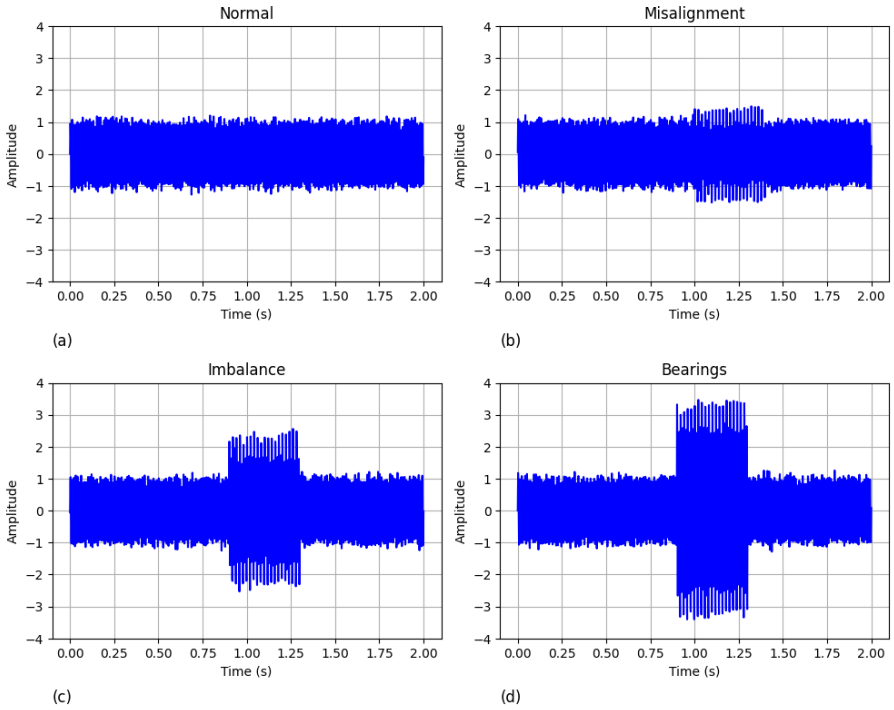

Fig. 5. The following waveforms represent vibration signals: (a) normal (no fault); (b) misaligned; (c) imbalanced; and (d) bearing faulty.

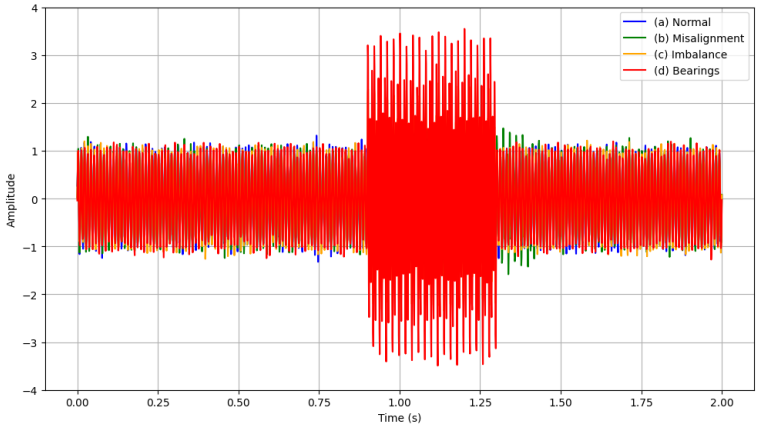

Figure 6: Signal representing all faults

After min-max normalization, the data stream was divided into 1,000 one-second segments with a 0.5-second overlap. Two engineers selected 600 deformation segments based on criteria such as frequency variations exceeding 5 Hz/s. Sub-sets were created for testing, validation, and training. The Mafaulda dataset is confidential and used solely for model validation.

4.2. Evaluation metrics

The following assessment indicators were employed in order to thoroughly assess the model's performance: F1 score, precision, and accuracy: the arithmetic mean of F1 scores across all classes, which gauges class balance. Evaluation of the model's precision and recall in preventing false alarms

and missed detections. Furthermore, we test robustness in varying signal-to-noise scenarios to measure real-world adaptability.

4.3. Baseline models

We contrast the suggested VIBT model with a number of sample models to show its efficacy:

- ❖LLM4TS [17]: This work proposed a LLM-based method for time series prediction using a pre-trained large speech model, which consisted of supervised fine-tuning and downstream fine-tuning.
- ❖ParInfoGPT [15]: An LLM-based two-stage framework for reliability assessment of rotating machine under partial information
- ❖FPT [21]: This work utilized a pre-trained LLM for time series analysis, which froze the self-attention and feedforward layers while fine-tuning the model to adapt to various tasks for time series analysis.
- ❖FEDforme [26]: This work proposed a frequency enhanced decomposed Transformer, which combined Transformer and Fourier Transform for extracting time series features

Every model was trained and assessed using the same data splits and training conditions. Table 3 provides a summary of the VIBT model design and training information.

Table 3: Training details and model configuration.

Parameter	**Value**
Optimizer	SGD
Number epochs	100
Encoder layer	2
Batch size	16
Rndom seed	42
Initiel leaning rate	1×10^{-3}

4.4 Experimental results

Table 5 provides an overview of each model's classification performance on the test set (see to Figure 7). Figure 7 shows the ROC curves for the FEDforme, FPT, ParInfoGPT, LLM4TS, and VIBT models, with AUC values ranging from 0.8013 for FEDforme to 0.9765 for VIBT, indicating the best performance of VIBT.

VIBT outperforms all models. Ablation experiments evaluated each module: without FBC, without WTA, and the reference with only the Transformer backbone. In our 6-fold cross-validation, we searched for the optimal number of wavelet decomposition levels in the WTA module. The results are shown in Table 4.

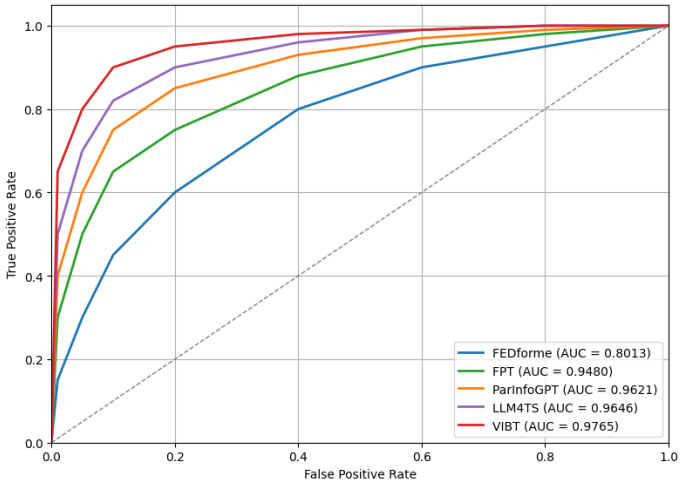

Figure. 7. ROC for models under comparison.

Table 4 shows the performance of the models in terms of accuracy, precision, recall, and F1 score. Model 4 excels in accuracy (0.921) and precision (0.925), while model 3 has the best recall (0.891) and F1 score (0.902). Model 5 has the lowest scores, indicating trade-offs. The table highlights the strengths and weaknesses of each model.

Table 4: Performance cross-validation for various decomposition levels *J*

J	Accuracy	Precision	Recall	Score FI
2	0,905	0,892	0,86	0,891
3	0,918	0,902	0,891	0,902
4	0,921	0,925	0,88	0,895
5	0,897	0,901	0,874	0,881
6	0,919	0,91	0,90	0,905

4.5 Discussion

Table 5 shows that VIBT has the best accuracy (0.988), precision (0.981), and recall (0.924). LLM4TS (0.965, 0.962, 0.906) and ParInfoGPT (0.964, 0.961, 0.904) perform well, with ParInfoGPT having the best F1 score (0.914). FEDforme has lower results (0.662, 0.652, 0.653) and is less suitable. In summary, VIBT is the optimal method, with LLM4TS and ParInfoGPT offering specific advantages.

Table 5: Results of five-fold cross-validation on the test sets.

Methode	Accuracy	Precision	Recall	Score FI
LLM4TS	0,965	0,962	0,906	0,905
ParInfoGPT	0,964	0,961	0,904	0,914
FPT	0,950	0,954	0,952	0,949
FEDforme	0,662	0,652	0,653	0,644
VIBT	0,988	0,981	0,924	0,929

The FBC module and WTA mechanism have been integrated to improve the performance of the VIBT model. Models such as LLM4TS and FPT do not handle non-stationarity, which limits their ability to detect frequency components, whereas the FBC module excels in this area. WTA identifies high-frequency transient disturbances, flagging anomalies that traditional models struggle to detect.

The implementation of VIBT could significantly reduce downtime in industry, with 98.8% accuracy in anomaly detection. By preventing up to 70% of failures, savings could reach 10 to 21 hours per machine per year, representing considerable financial gains. VIBT also integrates into predictive maintenance strategies for continuous monitoring and optimization of interventions.

Minimizing false negatives is crucial for safety. To improve recall, strategies include adjusting the decision threshold, using ensemble techniques such as bagging or boosting, and improving training data. These approaches show promise for enhancing industrial safety.

4.6- Ablation Study

Table 6 shows that **VIBT (Full)** achieves the best results (accuracy = 0.989, precision = 0.983, recall = 0.936). The **baseline** is lower (0.872/0.862/0.848), while **VIBT w/o FBC** (0.901/0.873/0.890) and **VIBT w/o WTA** (accuracy = 0.889) perform better than the baseline but remain below the full model. **VIBT (Full)** is the most reliable method.

Table 6: Study of ablation (complete metrics)

Methodes	Accuracy	Precision	Recall	Score FI
VIBT baseline	0,872	0,862	0,848	0,859
VIBT w/o WTA	0,889	0,861	0,871	0,884
VIBT w/o FBC	0,901	0,873	0,890	0,887
VIBT (Full)	0,989	0,983	0,936	0,946

The VIBT model performs well in classification despite a low signal-to-noise ratio (Figure 8), with an AUC of 0.9821. The J=4 method is the most balanced (precision of 0.921, recall of 0.88, FI of 0.895) (Table 4). The final configuration uses four levels of wavelet decomposition to optimize feature capture and reduce noise.

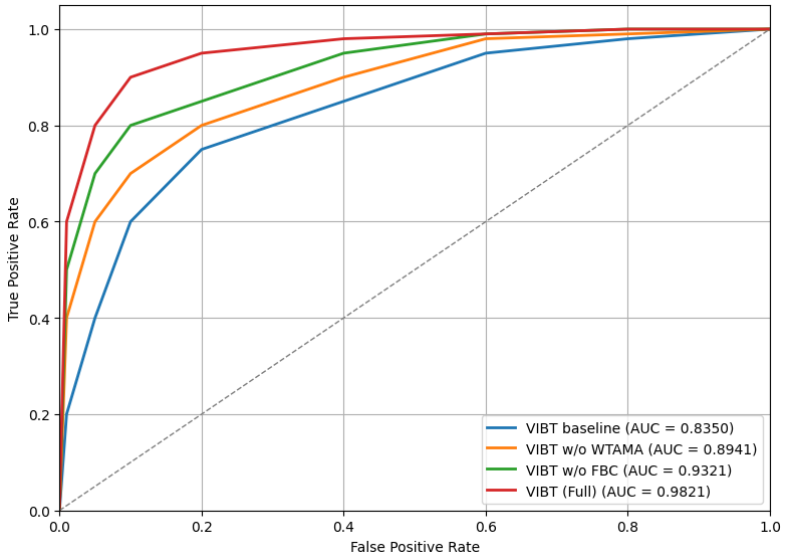

Figure. 8. ROC across several ablation models.

In conclusion, the VIBT model's exceptional ability to handle noise and non-stationary signal characteristics directly contributes to its enhanced anomaly identification. The model may capture small, transient high-frequency components that are indicative of structural deformations, which are limited in traditional time-series models, by utilizing both the FBC and WTA modules.

4.7 Limitations

The VIBT model, although effective for detecting vibration faults, has limitations. The Mafaulda dataset may not cover all failure scenarios, which affects its generalizability. Its complexity requires significant computing resources, making it difficult to deploy on less powerful machines. The interpretability of decisions also remains a challenge, highlighting the need for research to improve its robustness and adaptability.

5. Conclusion

This study presents VIBT, a Transformer-based framework for anomaly detection in rotating machinery, aimed at reducing losses in Industry 5.0. With 98.1% accuracy and 98.8% precision, VIBT overcomes the challenges of non-stationary signals.

We plan to extend the method to other systems and explore transfer learning. However, limitations remain: the Mafaulda dataset may not cover all failure scenarios, and the complexity of the model requires significant resources. The interpretability of decisions remains a challenge.

In summary, our research opens up new perspectives for the diagnosis of rotating machines and identifies avenues for future work.

Author Contributions: Conceptualization, TOUKAP NONO. and TOKOUE NGATCHA; methodology, TOUKAP NONO, TOKOUE NGATCHA and OFFOLE FLORENCE.; software, TOUKAP NONO and TOKOUE NGATCHA.; validation, OFFOLE FLORENCE and MOUZONG PEMI; investigation, TOUKAP NONO.; resources, TOKOUE NGATCHA, OFFOLE FLORENCE and MOUZONG PEMI; data curation, TOUKAP NONO.; writing original draft preparation, TOUKAP NONO. and TOKOUE NGATCHA; project administration, TOKOUE NGATCHA, OFFOLE FLORENCE and MOUZONG PEMI. All authors have read and approved the submitted version of the manuscript.

Funding: This research received no external or internal funding.

Conflict of Interest: The authors declare that they have no conflict of interest.

5.1. Future Work

In our future work, we will optimize the architecture of the VIBT model to reduce its complexity, using compression techniques such as quantization and pruning. We will explore transfer learning to improve its robustness and develop interpretability tools.

We will extend testing to other datasets and real-world data from sectors such as aerospace and energy to assess its adaptability. Integrating this data will validate the model's performance in industrial settings and optimize accuracy.

These efforts will strengthen user confidence in VIBT and open up new avenues of research for its continuous improvement.

5.2 a Glossary of Acronyms

- **AI: Artificial Intelligence**
  Artificial intelligence, a field of computer science that simulates human intelligence.
- **CNN: Convolutional Neural Network**
  Convolutional neural network, used primarily for image recognition and visual data analysis.
- **DAS: Distributed Acoustic Sensing**
  Distributed acoustic sensing system, used to measure vibrations over long distances using optical fibers.
- **DFT: Discrete Fourier Transform**
  Discrete Fourier transform, used to analyze the frequencies of a signal.
- **EMD: Empirical Mode Decomposition**
  Empirical mode decomposition, a method of analyzing nonlinear and non-stationary signals.
- **FBC: Filter Bank Convolution**
  Filter bank convolution, a module used for denoising vibration signals.
- **LLM: Large Language Model**
  Large language model, used for various natural language processing tasks.
- **MLP: Multi-Layer Perceptron**
  Multi-layer perceptron, a type of artificial neural network composed of several layers.
- **VIBT: Vibration Detection Transformer**
  Transformer for vibration detection, the method proposed in the article for analyzing vibration data.
- **WTA: Wavelet Transform Attention**
  Wavelet transform-based attention, an attention mechanism for dynamic fusion of multi-scale features..

Data Availability : The datasets used and analyzed in the current study are available in references [23] and [24]. Available at the link: https://www02.smt.ufrj.br/~offshore/mfs/page_01.html

References

[1] Z. Xu et J. H. Saleh, « Machine learning for reliability engineering and safety applications: Review of current status and future opportunities », *Reliability Engineering & System Safety*, vol. 211, p. 107530, 2021.

[2] W. J. Lee, H. Wu, H. Yun, H. Kim, M. B. Jun, et J. W. Sutherland, « Predictive maintenance of machine tool systems using artificial intelligence techniques applied to machine condition data », *Procedia Cirp*, vol. 80, p. 506-511, 2019.

[3] S. Liu *et al.*, « Fault diagnosis study of hydraulic pump based on improved symplectic geometry reconstruction data enhancement method », *Advanced Engineering Informatics*, vol. 61, p. 102459, 2024.

[4] H. Shi, Y. Miao, C. Li, et X. Gu, « A novel bearing intelligent fault diagnosis method based on spectrum sparse deep deconvolution », *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108102, 2024.

[5] S. Xia, W. Huang, et J. Zhang, « A novel fault diagnosis method based on nonlinear-CWT and improved YOLOv8 for axial piston pump using output pressure signal », *Advanced Engineering Informatics*, vol. 64, p. 103041, 2025.

[6] P. Liang, L. Xu, H. Shuai, X. Yuan, B. Wang, et L. Zhang, « Semisupervised subdomain adaptation graph convolutional network for fault transfer diagnosis of rotating machinery under time-varying speeds », *IEEE/ASME Transactions on Mechatronics*, vol. 29, nº 1, p. 730-741, 2023.

[7] Y. Li *et al.*, « A novel interpretable semi-supervised graph learning model for intelligent fault diagnosis of hydraulic pumps », *Knowledge-Based Systems*, vol. 305, p. 112598, 2024.

[8] P. Ong, Y. K. Tan, K. H. Lai, et C. K. Sia, « A deep convolutional neural network for vibration-based health-monitoring of rotating machinery », *Decision Analytics Journal*, vol. 7, p. 100219, 2023.

[9] Y. Liu, Y.-A. Zhang, M. Zeng, et J. Zhao, « A novel distance measure based on dynamic time warping to improve time series classification », *Information Sciences*, vol. 656, p. 119921, 2024.

[10] C. Gao, N. Zhang, Y. Li, F. Bian, et H. Wan, « Self-attention-based time-variant neural networks for multi-step time series forecasting », *Neural Computing and Applications*, vol. 34, nº 11, p. 8737-8754, 2022.

[11] X. Tang *et al.*, « Intelligent fault diagnosis of helical gearboxes with compressive sensing based non-contact measurements », *ISA transactions*, vol. 133, p. 559-574, 2023.

[12] Y. Li, J. Yang, S.-L. Shih, W.-T. Shih, C.-K. Wen, et S. Jin, « Efficient iot devices localization through wi-fi csi feature fusion and anomaly detection », *IEEE Internet of Things Journal*, vol. 11, nº 24, p. 39306-39322, 2024.

[13] S. Wang *et al.*, « Few-shot fault diagnosis of axial piston pump based on prior knowledge-embedded meta learning vision transformer under variable operating conditions », *Expert Systems with Applications*, vol. 269, p. 126452, 2025.

[14] K. Li, Y. Song, L.-R. Dai, I. McLoughlin, X. Fang, et L. Liu, « Ast-sed: An effective sound event detection method based on audio spectrogram transformer », in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, p. 1-5. Consulté le: 29 août 2025. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/10096853/

[15] Z. Pang, Y. Luan, J. Chen, et T. Li, « Parinfogpt: An llm-based two-stage framework for reliability assessment of rotating machine under partial information », *Reliability Engineering & System Safety*, vol. 250, p. 110312, 2024.

[16] M. Tami, S. Masri, A. Hasasneh, et C. Tadj, « Transformer-based approach to pathology diagnosis using audio spectrogram », *Information*, vol. 15, nº 5, p. 253, 2024.

[17] C. Chang, W.-Y. Wang, W.-C. Peng, et T.-F. Chen, « LLM4TS: Aligning Pre-Trained LLMs as Data-Efficient Time-Series Forecasters », *ACM Trans. Intell. Syst. Technol.*, vol. 16, nº 3, p. 1-20, juin 2025, doi: 10.1145/3719207.

[18] C. Li, Q. Hu, J. Xiong, et S. Ma, « Feature Entropy Recognition Based on Dual-Channel-Multi-Scale 1DCNN Model for Intelligent Compound Fault Diagnosis of Bearings », *IEEE Transactions on Instrumentation and Measurement*, 2025, Consulté le: 29 août 2025. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/10843157/

[19] X. Chen, B. Zhang, et D. Gao, « Bearing fault diagnosis base on multi-scale CNN and LSTM model », *J Intell Manuf*, vol. 32, nº 4, p. 971-987, avr. 2021, doi: 10.1007/s10845-020-01600-2.

[20] C.-H. H. Yang, Y.-Y. Tsai, et P.-Y. Chen, « Voice2series: Reprogramming acoustic models for time series classification », in *International conference on machine learning*, PMLR, 2021, p. 11808-11819. Consulté le: 29 août 2025. [En ligne]. Disponible sur: http://proceedings.mlr.press/v139/yang21j.html

[21] T. Zhou, P. Niu, L. Sun, et R. Jin, « One fits all: Power general time series analysis by pretrained lm », *Advances in neural information processing systems*, vol. 36, p. 43322-43355, 2023.

[22] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, et D. Sontag, « Tabllm: Few-shot classification of tabular data with large language models », in *International conference on artificial intelligence and statistics*, PMLR, 2023, p. 5549-5581. Consulté le: 29 août 2025. [En ligne]. Disponible sur: https://proceedings.mlr.press/v206/hegselmann23a.html

[23] M. Baptista, S. Sankararaman, I. P. de Medeiros, C. Nascimento Jr, H. Prendinger, et E. M. Henriques, « Forecasting fault events for predictive maintenance using data-driven techniques and ARMA modeling », *Computers & Industrial Engineering*, vol. 115, p. 41-53, 2018.

[24] R. M. Souza, E. G. Nascimento, U. A. Miranda, W. J. Silva, et H. A. Lepikson, « Deep learning for diagnosis and classification of faults in industrial rotating machinery », *Computers & Industrial Engineering*, vol. 153, p. 107060, 2021.

[25] D. Wang et M. Li, « Stochastic configuration networks: Fundamentals and algorithms », *IEEE transactions on cybernetics*, vol. 47, nº 10, p. 3466-3479, 2017.

[26] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, et R. Jin, « Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting », in *International conference on machine learning*, PMLR, 2022, p. 27268-27286. Consulté le: 29 août 2025. [En ligne]. Disponible sur: https://proceedings.mlr.press/v162/zhou22g