Heterogeneous vibration data preprocessing method for fault detection

Donatien Claeyssens, Dorsaf Zekri, Thierry Delot, Ait El Cadi Abdessamad

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 253 (2025) 2127–2136

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Industry 4.0 and Smart Manufacturing

Heterogeneous vibration data preprocessing method for fault detection

Donatien Claeyssens[a,e,*], Dorsaf Zekri[b,c], Delot Thierry[a], Ait El Cadi Abdessamad[a,d]

[a] *UPHF, CNRS, UMR 8201 - LAMIH, F-59313 Valenciennes, France*
[b] *CReSTIC EA 3804, Université de Reims Champagne-Ardenne, Reims, 51097, France.*
[c] *ReDCAD Laboratory, University of Sfax, Sfax, B.P. 1173, Tunisia.*
[d] *INSA Hauts-de-France, F-59313 Valenciennes, France*
[e] *I-care, Famars, France*

Abstract

The early detection of bearing faults is an important subject for maintenance and many research works have been conducted on laboratory-based homogeneous datasets. Industrial data sets are still uncommon, and so far, only one have been made publicly available so work on such industrial-based data sets are still rare. In this study, we focus on the use of an industrial-based data set and how well a model can classify new and unused data compared to a laboratory-based homogeneous data set. To do so, we propose a novel preprocessing method for heterogeneous vibration data set in order to extract as much information as possible for the model training. This preprocessing method is based on a sliding window method with a variable window length based on several factors of the data set such as the rotational speed of the machine, the duration of the measurement as well as the sampling frequency. Our preprocessing method ensures standardization of the dimensions, minimal loss of information and maximizes the quantity of data available. Because of the highly unbalanced nature of such heterogeneous data sets, we use spectrograms as the data representation and the CNN as a learning model. A validation method and experimental results are proposed to show the superiority of a model trained on heterogeneous data set over homogeneous data set.

Keywords: Fault detection; heterogeneous data; vibration data set; factory based data set; deep learning

1. Introduction

With the advent of Industry 4.0, each domain of manufacturing is introducing usage of computers and digitalization, and maintenance is one of them [1, 2]. Maintenance is essential because it enhances the lifespan of a device. Moreover, it should be planned in advance with an accurate estimation of the machine failure period to reduce the

* Donatien Claeyssens Beaupere
E-mail address: Donatien.Claeyssensobeaupere@uphf.fr

1877-0509
10.1016/j.procs.2025.01.273

risk of accidents, financial losses, and human casualties. This field is now known as predictive maintenance (PM). It is a technique where maintenance is conveniently scheduled according to the condition of a machine in terms of its degree of degradation and probability of failure. Applying predictive maintenance in production environments brings several benefits and also involves overcoming several challenges. On one side, benefits of PM include productivity improvement, reduction of system faults, financial losses and human casualties, minimization of unplanned downtimes and risk of accidents, increased efficiency in the use of financial and human resources, and the optimization in planning the maintenance interventions. For predictive maintenance, AI has been increasingly used in predicting the requirement and planning of maintenance operations for machines [3]. Indeed, machine learning techniques are gaining popularity especially predictive algorithms which can represent a viable solution in order to analyze data, predict trends, behavior patterns, and correlations by statistical or machine learning models for anticipating pending failures in advance to improve the decision-making process for the maintenance activity avoiding mainly the downtime.

On the other hand, overcoming challenges include the need to integrate heterogeneous data from various sources and systems within a facility which is important to gather accurate information to construct a machine learning model [4] or data-driven real-world systems [5]. Indeed, in order that these techniques be implemented and effective, a large amount of good data must be available [6]. These data can be in several formats and types for example electrical, pressure thermal or vibration data which is the most used type of data for fault diagnosis and prediction due to their noninvasive nature [7] and their effectiveness in identifying eccentricities [8]. These data sets can be obtained from the industry real world or/and laboratory environments [9, 10, 11, 12, 13, 14] which do not have the same characteristics. Indeed, the laboratory data sets typically do not represent the heterogeneity present in real data sets since industrial facilities don't always use the same type of sensors and the same parameters for each measurement device. The preprocessing of these heterogeneous data, particularly in industrial environments, presents unique challenges due to variations in data acquisition conditions, such as sampling frequency and machine rotational speed. Traditional preprocessing methods often fall short when applied to such data. These methods typically assume uniform data characteristics, which is rarely the case in industrial settings where different machines operate under varying conditions.

Furthermore, industrial data may be prone to erroneous measurements, due to harsh environmental conditions or sensor faults. Imbalanced data is a common challenge encountered in both real and laboratory data sets. It occurs when there is an uneven distribution of observations in the target class, with one class label having a significantly higher number of observations compared to the other. The presence of imbalanced data can have a notable impact on the performance and reliability of fault detection models. This is due to the difficulties that machine learning models may encounter when trying to accurately learn and predict the minority classes with limited training examples.

The main objective of this article is to propose a solution for detecting bearing defaults using heterogeneous vibration data collected from various sources in real factory environments. The originality of the work presented in this article is to propose a new preprocessing method for heterogeneous vibration data based on the sliding window technique. Our preprocessing method ensures standardization of the dimensions, minimal loss of information and maximizes the quantity of data available. Thus, our contributions in this article are the following:

- We present a novel preprocessing method for industrial-based heterogeneous data set based on the sliding window technique using a variable window length depending on several factor such as the machine rotational speed, the sampling frequency and the measurement duration.
- Then we present a validation method for deep learning model trained on industrial-based heterogeneous data set or laboratory-based homogeneous data set aiming to test the classification performance of the model on data not previously used during the training and validation phases.

The rest of this paper is organized as follows. In Section 2, we discuss about different research works related to our study. Section 3 details our approach with a preprocessing method for heterogeneous data set and a deep fault diagnosis model. In Section 4, an experimental study is proposed to validate our approach. Finally, we give our conclusions and introduce future works in Section 5.

2. Related works

In the context of predictive maintenance fault detection models are usually data-driven models that require a variety of data streams provided by multiple real-time and offline sources. In recent years, various physical entities, such as the air gap flux [15], torque [16], vibration [17], acoustics [18], stray flux [19], and more, have found applications in fault diagnosis.

The existing literature underscores the significance of vibration signals for fault diagnosis, primarily due to their noninvasive nature [7]. In addition, vibration signals have proven effective in identifying eccentricities as in [8]. In the literature several public data sets related to vibration data are available and were used as the main data sources by many research papers. These data sets can be categorized into two main types: laboratory and industry data sets.

Several public laboratory data sets focused on the mechanical defects of rotation machines either by offering several measurement of the machine running with and without a specific defect. Considering the sampling frequency, most of the laboratory data sets are not heterogeneous [20, 11, 21, 10, 9]. It provide signals acquired at one sampling frequency for just one machine. Some laboratory data sets, such as [22] and [23], offer two sets of data with two different sampling frequencies. However, both sets of measurements pertain to the same machine and consequently represent the same vibration signal.

A few public industry data sets are recently published [12] representing different mechanical systems with different measurements. As example, the data set published by Lundtröm et al. in [12] gathered bearing vibration data from industrial real-world over 4 years. Moreover, they captured 11 broken bearing on 11 different machines with different rotational speeds and sampling frequencies. This data set is not balanced and the ball defect is under represented compared to the other types of defects contained. The existence of imbalanced data sets can significantly affect the performance and reliability of fault detection models. This is because machine learning models may struggle to accurately learn and predict the minority classes when there are limited training examples available.

In the literature, researchers have proposed several processing methods for vibration data. These methods make the vibration data usable and useful for machine or deep learning techniques aimed at predicting or detecting faults. Among the proposed methods we can find method use on homogeneous laboratory-based data set as in [24] where authors proposed a method based on the Empirical Mod Decomposition method that allow, like wavelet transform and Short Term Fourier Transform method, time-frequency features to be extracted from the data source bu creating a certain number of Intrinsic Mode Function describing the different frequency that composed the signal studied. Another method proposed by Minh Tuan Pham et al. [25] relies on the use of the well known Short-Term Fourier Transform to generate spectrograms based on the time series of the original data. Another processing method is proposed in [26] with the Hilbert transform as the foundation of their method and images generation afterward with a fixed window length for their operation on the data. Some researchers have also proposed custom methods representing the data in one or more data representation domain. For example, [27] and [28] have proposed to transform time series in grey scale images by transposing them into tables of fixed dimensions and converting it in grey scale pixel. As an alternative to these methods Li et al. [29] proposes to represent their data both in the time domain, frequency domain and time-frequency domain. They have done this through data processing methods like Fast Fourier Transform and Short-Term Fourier Transform and transposing it into different fixed size table. Each of the fixed size table was used as a color layers for a RGB Images.

Other researchers have proposed different method of processing and tested them on data set with some degree of heterogeneity as the MAFAULDA data set[20] where the speed widely vary from a measure to another without getting under 700rpm, or by using the whole CWRU data set[22] with both the data sampled at 12,000Hz and the data sampled at 48,000Hz. Souza et al. proposed in [30] to use a method based on the generation of images using the fast Fourier transform of the vibration signal of the MAFAULDA data set but also with the CWRU data set using bot the 12,000Hz and 48,000Hz data. Nascimiento et al. in [31] also used a similar method with fast Fourier transform as their entry data for the transformer model.

Most of the methods cited above as [28, 27, 25, 29, 32] use the sliding window technique but with a fixed size and fail to account for both the speed and sampling frequency of each measurement in the data set. Therefore when used with heterogeneous industrial-based data set, they cannot accurately extract features due to the data's heterogeneity. Consequently, these data processing techniques tested on homogeneous data set may not perform effectively on data sets with heterogeneous rotating speed and sampling frequency.

3. Methodology

In the following, we explain, in Section 3.1, our proposed method for the preprocessing of the factory-based heterogeneous data set. Then we detail, in Section 3.2, our deep learning model for fault classification based on a convolutional neural network.

3.1. Preprocessing of heterogeneous dataset

The factory-based heterogeneous data set presents two primary challenges that must be addressed using the preprocessing method proposed in this paper. The first and foremost challenge is that the data set encompasses measurements from machines operating at various rotating speeds as well as various sampling frequencies. The second challenge is the limited amount of data available for each defect represented in the data set. In order to address these challenges, we introduce in this article a method that focuses on preprocessing the vibration data using a spectrogram generation technique. Common preprocessing methods have been used on homogeneous data sets or partially heterogeneous data sets, such as the Mafaulda data set [20]. However, they often fail to take into consideration rotational speed and sampling rate when processing different measurements. As a result, important information may be overlooked when applied to measurements from low-speed machines operating below 600 rpm or to measurements with low sampling rates that can occur in industrial setups.

To effectively address the first challenge concerning varying rotating speeds and sampling frequencies, we proposed a method that adapts the window size used for spectrogram generation. This adaptive window size is determined by the rotating speed of each machine and the measurement duration, which is based on the sampling frequency of each measurement. The choice of specific parameters in the sliding window technique is crucial for ensuring accurate feature extraction from heterogeneous vibration data. The rotating speed and the decision to use at least five rotations were determined based on established practices in vibration analysis and fault detection. This value is based on the work of Girdhar and Scheffer [33], which suggests that a minimum of 5 rotations is needed to analyze vibration data.

In our case, we consider a minimum of 5 rotations. We, also, based our windowing method on a specific rotating speed value X allowing us to adapt the window length according to both the speed and the duration of the measurement. If the observed machine's rotating speed is lower than the specified X value, our method uses the entire duration of the time series, provided it is long enough to capture a minimum of 5 rotations. Conversely, if the rotating speed exceeds the X value, our method uses a window size equivalent to one second of the signal or the same number of data samples as the sampling frequency. For the second challenge regarding the quantity of data available and the volume needed for training, we decided to add at our method the sliding window enables to increase the number of sample extracted from the original data as shown in Figure 1. In order to maximize the number of segment while minimizing redundancy, we decided to use a shifting factor equal to half the sliding window size. Given our goal to use an heterogeneous data set with varying rotating speeds and sampling frequencies, we needed a solution to standardize the dimensions of our processed data. After reviewing various representation utilized in the literature, we decided to generate spectrograms from the different waveform and waveform segments extracted by our method. The spectrogram generation method through Short-Term Fourier Transform is a well known technique allowing for time-frequency characteristics extraction and representation with a readable format. Thus allowing us to verify the performance of a deep learning model with those images.

3.2. A fault deep classification model

For fault classification with vibration data several type of neural network have been proposed and tested with success. In this study, considering the use of image as the data format in our proposed preprocessing method, we decided to use a convolutional neural network as our test model because it performs well with images and has proven effective for fault classification and detection using vibration data. This type of neural network offer good classification and defect detection performance for ball bearing fault and defect detection in general. The model used in this study is constructed as follows :

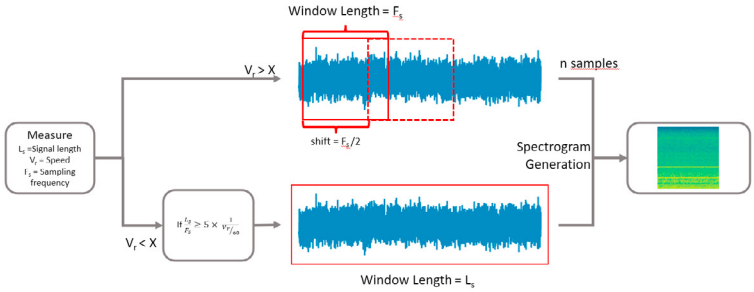


Fig. 1. (a) Pre processing method schematic

- Four convolutional layers with a ReLu activation, a kernel size of 3 by 3, 20 filters for the two first layers and 40 for the two last layers. Each convolutional layers is followed by one max-pooling layer and one dropout layers with a dropout factor of 0.3.
- One dropout layers with a dropout factor of 0.4 selected after optimisation
- Three fully connected Dense layers with 1536, 256 and 2 units respectively with Relu Activation function for the first two and Softmax for the last one for classification.
- The number of epochs was set to 10 and the batch size to 64

4. Experimentation and Results

In the following, as a case study, two data sets were chosen to explain the experimental set-up. We start with detailing in the different data sets used 4.1 then with detailing 4.2 the process for the preprocessing of heterogeneous data set and then illustrate in 4.3 the faults classification result of the deep learning model.

4.1. Data sets

For our experiment we wanted to show the impact of an heterogeneous factory based data set on the training and the classification capabilities of a deep learning model on external data and compared it with a model trained on an homogeneous laboratory based data set. To do that, we decided to used two public data sets :

1. The laboratory data set provided by The Case Western Reserve University (CWRU) [22], a very well known data set that have been use several times before. This data set offer a large number of measurement over different severity of defect for bearing fault varying from 0.007" to 0.028" in fault diameter. It is measure at two sampling frequency 12,000Hz and 48,000Hz and two different place. They provide measurement for the three main faults of a bearing i.e. ball, inner ring and outer ring defects at different speed from 1720rpm to 1797rpm. In order to recreate an homogeneous data set we decided to limit this data set to the 12,000Hz part being the one with the most measurement made on. We ended with 83 different measurements over the 4 states.

2. The real factory data set provided by real factory of Svenska Cellulosa Aktiebolaget (SCA) [12]. This data set is based on eleven different machines and offer heterogeneous specification with different sampling frequency and rotational speed. The detail of this data set and its composition can be found in the table 1. This table contains 4 columns with the case number in the first one, the sampling rate corresponding to the number of vibration measurements made each seconds in the second one, the average rotating speed of each machine and then the type of bearing defect measured on this machine. In our study, we exclude the 11^{th} machine because it does not indicate the presence of a bearing-related defect. This data set is composed of five different labels corresponding to a different state. The different faults labels are as follow:

- 0 : No defect in the bearing
- 1 : Defect in the inner race of the bearing
- 2 : Defect in one or several balls of the bearing
- 3 : Defect in the outer race of the bearing
- -1 : Machine offline

In our study, we exclude the class -1 since the machine turn off so it cannot present any defects.

Table 1. SCA bearing dataset description

Case number	Sampling Rate (Hz)	Average rotation speed	Fault type
1	640	1120.6	Inner ring
2	5120	1162.0	Outer ring
3	512	34.6	Inner ring
4	8192	1100.3	Inner ring
5	12800	2483.5	Ball
6	6400	1208.5	Inner ring
7	4096	700	Inner ring
8	5120	1105.9	Outer ring
9	5120	1162.0	Outer ring
10	5120	189.3	Outer ring
11	5120,12800	189.3	Not bearing

To train our classification model, we used 85% of the data set, while the remaining data (15%) is used to evaluate the model's ability to classify machine faults.

4.2. Processing Results

To overcome the low quantity of data in the publicly available heterogeneous factory-based data set, named SCA Bearing[12], we use a sliding window technique explained in Section 3. In our method, in order to overcome both the low quantity of data and the severe imbalance profile, we make the sliding window method more flexible by the addition of a parameter allowing the window size to be modulated according to the signal length, the rotating speed and the sampling frequency of the signal observed. As explained in [33] the number of full rotation needed in a measure in order to do an analysis should be at least 5 and up to 10 full rotations so we decided to set the specific rotating speed value X in the data set at 600rpm or 10 rotation per second allowing us then to produce samples with enough information in each of them for model training.

Given a value of rotation speed X equal to 600rpm, our processing method will either use a one-second window or the entire signal length for spectrogram generation. With this value of 600rpm we were able to guaranty that for every case with a rotational speed higher than 600rpm, we had at least 10 full rotation in each spectrogram. For the other cases, by using the whole length of the signal and ensuring sufficient duration, we guaranteed a minimum of 5 complete rotations in every spectrogram used for the training of our model.

By applying our preprocessing method on the real factory data set SCA, we obtain 23630 spectrograms distributed on the 4 faults classes (1,2,3 and 4) listed in Section 4.1. Considering the huge imbalance profile of this data set after spectrogram generation, we decided to only use the classes 0, 1 and 3 respectively "No Defect", "Inner Race Defect" and "Outer Race Defect". The fault class 2 was be represented only by 23 spectrograms even with the sliding windows method. Therefore, it is considered insufficiently representative and was excluded from training in this study. We apply also our preprocessing method on the laboratory data set CWRU, we obtain 3066 spectrograms for the 3 classes "no defect", "inner ring fault" and "outer ring fault". To highlight the effectiveness of our preprocessing method, we compared, in Table 2, the resulting number of generated spectrograms using our preprocessing method with the baseline model without the sliding window in spectrogram generation. At a second stage, we compared, in Table 2, our method to the model proposed by [32] which we found the most related to our work since this work aims, as well, to use the sliding window technique but with a fixed window length of 5000. According to the results

presented in table 2 our method increased the number of spectrogram images generated by 7.11 times compared to the baseline method and by 1.49 times compared to the fixed sliding window of 5000 samples [32]. This proves the effectiveness of the sliding window with variable length to maximize the quantity of spectrograms for training and testing in the heterogeneous dataset. We also compared the number of spectrograms in each class and found that our method reduced the imbalance characteristics between the two subsets we will consider in Section 4.2 namely, "without defect" and "with defect", as shown in Table 3.

Table 2. Total number of generated spectrograms in the preprocessing step

Methods	Real factory data set (SCA)	Laboratory data set (CWRU)
Baseline method without sliding window	3318	81
Fixed sliding window (5000 samples)[32]	15801	9452
Our method with sliding window	23607	3066

Table 3. Number of generated spectrograms for SCA data set per class

Classes	SCA : Our method	SCA : baseline method	SCA : sliding window with fixed length (5000 samples)
0	15367	2470	12458
1	6665	372	1737
2	23	23	115
3	1575	453	2127

As show in table3 our method gave us 6.22x more images for the first class, 17.91x more for the second, 3.47x more for the fourth and the same amount for the third class. Compared to the other method used here our method gave us 1.23x more images for the first class, 3.83x more for the second, a reduction by a factor 5 for the third and a reduction by 1.35 for the fourth. The reduction can be explained by the low speed and high sample rate of each case.

4.3. Classification results

Once the preprocessing step is completed and the spectrograms are generated, we propose an experimental study to evaluate the generalization capabilities of our faults classification system. Therefore, we considerate only two classes (1) without defect and (2) with defect including both the inner ring fault and the outer ring fault. Then we repeated the complete faults classification process explained in Section 3.2 in two different ways:

- The first way consists of training our model on the 85% of heterogeneous factory based data set (SCA) afterwards we validate the model on the last 15% of the data set. Then we do a classification test of the trained model on the homogeneous data set (CWRU) to see how well it can generalize its learning on new data.
- The second way consists of training the same proposed model on the homogeneous data set (CWRU) with the same split parameter for training and validation. Then the trained model will be tested on the heterogeneous data set (SCA) by doing classification of all the sprectrograms of the data set.

On each of the validation part of the data sets used for the training we achieve an accuracy of 98.89% and 100% for respectively heterogeneous and homogeneous data sets confirming that our model is properly trained.

We proceed to the second part of the experiment, where each model will be tested on the other data set. The confusion matrix of each test consisting of taking the model trained on one data set and testing on the other one can be found in the figure 2. For the performance and classification metrics we use several well known metrics such as accuracy, precision, Recall, F1 score and AUC ROC using respectively equations (1), (2), (3), (4) and (5).

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

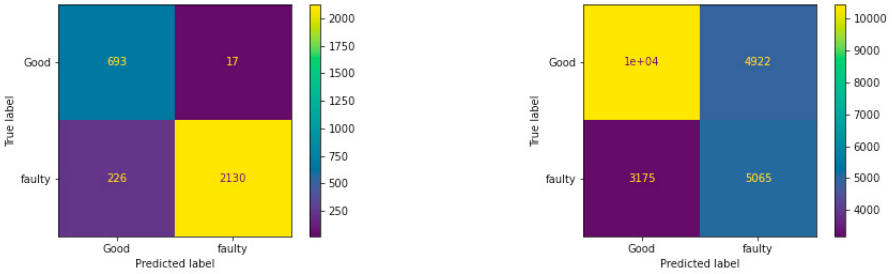
$$F1\_Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

$$AUC\_ROC = \int TPR\, dFPR \tag{5}$$

Where TP is True Positive, TN is True Negative, FP is False Positive, FN is False Negative, TPR is True Positive Rate and FPR is False Positive Rate.

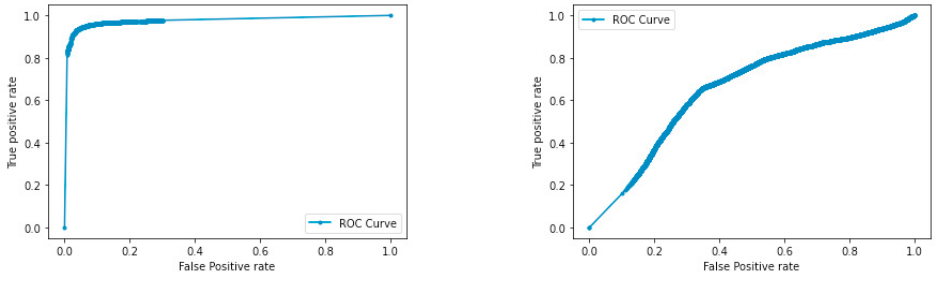
Table 4. cross prediction results.

Data sets	SCA-trained tested on CWRU	CWRU-trained tested on SCA
Accuracy	92.07%	65.7%
Recall	89.78%	61.4%
Precision	99.23%	50.72%
F-score	94.60%	55.588%
AUC ROC	0.938	0.647



(a) SCA trained (b) CWRU trained

Fig. 2. Confusion matrix of the two 2-classes model



(a) SCA trained (b) CWRU trained

Fig. 3. ROC Curve of the two 2-classes model

Table 4 shows the performance metrics of cross-evaluation (the heterogeneous trained model applied to homogeneous data and vice versa): precision, recall, precision, F1 score, and AUC ROC. Each one provides insights into the effectiveness of the proposed method in fault detection.

- The heterogeneous trained model achieves an accuracy of 92.07% on the homogeneous data set and the homogeneous trained model achieves an accuracy of only 65,72%, indicating its robustness in handling both controlled and variable conditions.
- Precision reflects the proportion of correctly identified faults among all predicted faults. The high precision (99.23% vs. 50.72%) shows the reliability of the method in reducing false alarms, which is crucial in industrial applications.
- Recall indicates the model's ability to detect actual faults. The results (89.78% vs. 61.4%) demonstrate effective fault detection under both homogeneous and challenging heterogeneous conditions.
- F1 Score balances precision and recall, with values of 94.60% vs. 55.58%. This metric shows the ability of the method to manage both false positives and false negatives effectively.
- AUC ROC measures the model's ability to distinguish between faulty and non-faulty states. The scores of 0.938 vs. 0.647 highlight the strength in classifying faults accurately.

In the evaluation of our classification model it is crucial to measure the ability to distinguish between different outcome classes accurately. In addition to the metrics in Table 4, we add the ROC Curve in fig.3 that allows to see the relation between the True Positive Rate and the False Positive Rate of our model over several thresholds.

5. Conclusions

In this paper, we introduce a novel preprocessing method for heterogeneous vibration data based on the sliding window technique. The window size may change depending on several factors like rotational speed, measurement duration or sampling frequency. The originality of this work is to standardize heterogeneous vibration dataset, minimal loss of information and maximizes the quantity of data available. We have validated our approach by conducting an experimental study addressing a factory-based heterogeneous data set SCA and laboratory-based homogeneous data set CWRU. This study have showed that our method outperforms the baseline model without sliding window. In the future work we focus on the gathering of data set from industrial or factory environment for bearing faults, gear and gearbox faults and many other type of mechanical defect in order to develop further both the preprocessing method and new model capable of learning from such data and offering both good performance and good efficiency for industrial needs.

6. Discussion

The proposed method for utilizing a heterogeneous industrial dataset for model training in fault classification related to bearings demonstrates promising performance. Furthermore, this approach can be extended to address various fault types, including gear faults, misalignment faults, and others. It is essential to continue refining preprocessing techniques to ensure their compatibility across different fault categories. Future work could focus on improving the preprocessing methods, such as the one presented here, to better accommodate diverse data types while providing more granular control over the preprocessing parameters used.

Regarding the dataset analyzed in this paper, it would be beneficial to explore other deep learning models for fault detection, such as transformer models, in addition to the CNN-based approach we employed. These models have the potential to reduce the necessity for data separation based on rotational speed, making them more suitable for heterogeneous data sets like the one used here.

References

[1] Mustufa Abidi, Hisham Alkhalefah, and Usama Umer. Fuzzy harmony search based optimal control strategy for wireless cyber physical system with industry 4.0. *Journal of Intelligent Manufacturing*, 33, 08 2022.

[2] Praveen Kumar Reddy Maddikunta, Quoc-Viet Pham, Prabadevi B, N Deepa, Kapal Dev, Thippa Reddy Gadekallu, Rukhsana Ruby, and Madhusanka Liyanage. Industry 5.0: A survey on enabling technologies and potential applications. *Journal of Industrial Information Integration*, 26:100257, 2022.

[3] H Kamel. Artificial intelligence for predictive maintenance. *Journal of Physics: Conference Series*, 2299(1):012001, jul 2022.

[4] Fazel Ansari, Robert Glawar, and Wilfried Sihn. Prescriptive maintenance of cpps by integrating multimodal data with dynamic bayesian networks. In Jürgen Beyerer, Alexander Maier, and Oliver Niggemann, editors, *Machine Learning for Cyber Physical Systems*, pages 1–8, Berlin, Heidelberg, 2020. Springer Berlin Heidelberg.

[5] Iqbal Sarker, Moshiul Hoque, Kafil Uddin, and Tawfeeq Alsanoosy. Mobile data science and intelligent apps: Concepts, ai-based modeling and research directions. *Mobile Networks and Applications*, 26, 02 2021.

[6] Tiago Zonta, Cristiano André da Costa, Rodrigo da Rosa Righi, Miromar José de Lima, Eduardo Silveira da Trindade, and Guann Pyng Li. Predictive maintenance in the industry 4.0: A systematic literature review. *Computers and Industrial Engineering*, 150:106889, 2020.

[7] Iman Sadeghi, Hossein Ehya, Jawad Faiz, and Hossein Ostovar. Online fault diagnosis of large electrical machines using vibration signal-a review. In *2017 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM) and 2017 Intl Aegean Conference on Electrical Machines and Power Electronics (ACEMP)*, pages 470–475, 2017.

[8] Yu-Ling He, Yong Li, Wen Zhang, Ming-Xing Xu, Yi-Fan Bai, Xiao-Long Wang, Shan-Zhe Shi, and David Gerada. Analysis of stator vibration characteristics in synchronous generators considering inclined static air gap eccentricity. *IEEE Access*, 11:7794–7807, 2023.

[9] Christian Lessmeier et al. KAT-Datacenter https://mb.uni-paderborn.de/kat/forschung/kat-datacenter/bearing-datacenter/data-sets-and-download. Accessed: 2024-04-27.

[10] Dumond Patrick Sehri Mert. University of ottawa rolling-element dataset – vibration and acoustic faults under constant load and speed conditions (uored-vafcls).

[11] Femto bearing data set. https://www.nasa.gov/content/prognostics-center-of-excellence-data-set-repository. Accessed: 2024-04-27.

[12] Adam Lundström and Mattias O'Nils. Factory-based vibration data for bearing-fault detection. *Data*, 8(7), 2023.

[13] Mohammed Hakim, Abdoulhdi A. Borhana Omran, Ali Najah Ahmed, Muhannad Al-Waily, and Abdallah Abdellatif. A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations. *Ain Shams Engineering Journal*, 14(4):101945, 2023.

[14] Shen Zhang, Shibo Zhang, Bingnan Wang, and Thomas G. Habetler. Deep learning algorithms for bearing fault diagnostics—a comprehensive review. *IEEE Access*, 8:29857–29881, 2020.

[15] Hind C. Dirani, Arezki Merkhouf, Anne-Marie Giroux, Bachir Kedjar, and Kamal Al-Haddad. Impact of real air-gap nonuniformity on the electromagnetic forces of a large hydro-generator. *IEEE Transactions on Industrial Electronics*, 65(11):8464–8475, 2018.

[16] Aoran Gao, Zhipeng Feng, and Ming Liang. Permanent magnet synchronous generator stator current am-fm model and joint signature analysis for planetary gearbox fault diagnosis. *Mechanical Systems and Signal Processing*, 149:107331, 2021.

[17] Rony Ibrahim, Ryad Zemouri, Bachir Kedjar, Arezki Merkhouf, Antoine Tahan, Kamal Al-Haddad, and François Lafleur. Non-invasive detection of rotor inter-turn short circuit of a hydrogenerator using ai-based variational autoencoder. *IEEE Transactions on Industry Applications*, 60(1):28–37, 2024.

[18] Fang Dao, Yun Zeng, Yidong Zou, Xiang Li, and Jing Qian. Acoustic vibration approach for detecting faults in hydroelectric units: A review. *Energies*, 14(23), 2021.

[19] Helene Bechara, Ryad Zemouri, Bachir Kedjar, Arezki Merkhouf, Kamal Al-Haddad, and Antoine Tahan. Non-invasive detection of rotor inter-turn short circuit in large hydrogenerators by using stray flux measurement combined with convolutional variational autoencoder analysis (cvae). *IEEE Transactions on Industry Applications*, 60(1):196–205, 2024.

[20] https://www02.smt.ufrj.br/~offshore/mfs. Accessed: 2023-09-27.

[21] Natalie Baddour Huan Huang. Bearing vibration data under time-varying rotational speed conditions.

[22] https://engineering.case.edu/bearingdatacenter/welcome. Accessed: 2024-05-27.

[23] Dr Eric Bechhoefer. Condition based maintenance fault database for testing of diagnostic and prognostics algorithms. https://www.mfpt.org/fault-data-sets. Accessed: 2024-04-27.

[24] Rafia Nishat Toma, Cheol-Hong Kim, and Jong-Myon Kim. Bearing fault classification using ensemble empirical mode decomposition and convolutional neural network. *Electronics*, 10(11), 2021.

[25] Minh Tuan Pham, Jong-Myon Kim, and Cheol Hong Kim. Accurate bearing fault diagnosis under variable shaft speed using convolutional neural networks and vibration spectrogram. *Applied Sciences*, 10(18), 2020.

[26] M. Zabin, Choi, HJ., and J Uddin. Hybrid deep transfer learning architecture for industrial fault diagnosis using hilbert transform and dcnn–lstm. *J Supercomput*, 79:5181–5200, 2023.

[27] Jing Zhao, Shaopu Yang, Qiang Li, Yongqiang Liu, Xiaohui Gu, and Wenpeng Liu. A new bearing fault diagnosis method based on signal-to-image mapping and convolutional neural network. *Measurement*, 176:109088, 2021.

[28] Yahui Zhang, Taotao Zhou, Xufeng Huang, Longchao Cao, and Qi Zhou. Fault diagnosis of rotating machinery based on recurrent neural networks. *Measurement*, 171:108774, 2021.

[29] Jun Li, Yongbao Liu, and Qijie Li. Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method. *Measurement*, 189:110500, 2022.

[30] Roberto M. Souza, Erick G.S. Nascimento, Ubatan A. Miranda, Wenisten J.D. Silva, and Herman A. Lepikson. Deep learning for diagnosis and classification of faults in industrial rotating machinery. *Computers and Industrial Engineering*, 153:107060, 2021.

[31] Erick Giovani Sperandio Nascimento, Julian Santana Liang, Ilan Sousa Figueiredo, and Lilian Lefol Nani Guarieiro. T4pdm: a deep neural network based on the transformer architecture for fault diagnosis of rotating machinery, 2022.

[32] Liangwei Zhang, Qi Fan, Jing Lin, Zhicong Zhang, Xiaohui Yan, and Chuan Li. A nearly end-to-end deep learning approach to fault diagnosis of wind turbine gearboxes under nonstationary conditions. *Engineering Applications of Artificial Intelligence*, 119:105735, 2023.

[33] Paresh Girdhar and C. Scheffer. 4 - signal processing, applications and representations. In Paresh Girdhar and C. Scheffer, editors, *Practical Machinery Vibration Analysis and Predictive Maintenance*, pages 55–88. Newnes, Oxford, 2004.