

Exploring Rolling Element Bearing Data Collection and Algorithm Hyperparameters for Machine Learning-Based Fault Diagnosis

Mert Sehri¹, and Patrick Dumond¹

¹ *University of Ottawa, Ottawa, ON, K1N 6N5, Canada*

msehr006@uottawa.ca

pdumond@uottawa.ca

ABSTRACT

This paper explores rolling element bearing data collection and hyperparameter tuning for machine learning-based fault diagnosis to aid in the development of modern condition monitoring systems. The integration of industrial internet of things (IIoT) products and cloud databases has led to an increased interest in utilizing artificial intelligence (AI) models, including artificial neural networks (ANNs) and convolutional neural networks (CNNs), to diagnose machine faults. However, the development of AI methodologies in smart monitoring is hindered by a lack of publicly available industry data, as well as limitations involved in the collection and storage of large high-dimensional datasets. Combining machine learning (ML) methods, such as traditional learning (TL), deep learning (DL), and bearing signature theory, will allow for a better understanding of data collection and hyperparameter tuning. Moreover, considering how high-dimensional datasets for rolling element bearing fault diagnosis affect ML algorithms has yet to be explored in the literature, providing little robustness for analysis. Concerns around the way data has been collected and used historically for both TL and DL are raised. Therefore, recommendations for data collection specifically suited to TL and DL methods for rolling element bearing fault diagnosis are proposed by analyzing existing lab-based datasets. The recommendations proposed combine knowledge of these methodologies to aid in selecting an appropriate sampling rate, as well as the ideal number of samples, stride, duration of each sample, and resolution for rolling element bearing fault diagnosis. The goal is to increase efficiency and reduce setup and collection time when selecting the design parameters for creating new

rolling element bearing datasets. To achieve this, the study applied a structured approach with the use of multiple datasets to determine a threshold accuracy of 95% for fault diagnosis. Furthermore, the results of this study will help IIoT companies re-evaluate the constraints imposed by the limited data storage and transmission of their devices when used for ML. This paper will also help improve the efficiency and effectiveness of AI methodologies in smart monitoring systems by establishing data collection recommendations. This work will hopefully motivate the vast collection of open-access data that can be used by researchers to further develop ML-based methods for rolling element fault diagnosis.

1. INTRODUCTION

In the heavy machinery industry, IIoT devices, as well as TL and DL methods, collectively known as ML methods, have recently gained popularity for tracking machine health and performing condition monitoring. However, obtaining quality data to train these ML-based techniques presents a substantial challenge (Chandrvanshi et al., 2024; Rahman et al., 2023). So far, a few high-dimensional datasets have been collected by researchers in the lab without describing their reasoning for the data collection method they used, especially when related to their use with ML techniques for fault diagnosis. However, a methodological approach to data collection is crucial for the continued development of useful ML algorithms (Chandrvanshi et al., 2024; Soomro et al., 2024). This study intends to clarify the important characteristics associated with data collection parameters, as well as the hyperparameters, that affect the training of ML algorithms to overcome existing constraints and enhance the accuracy and effectiveness of rolling element bearing analysis when using ML. The parameters selected are used to identify a threshold accuracy based on computational efficiency and storage requirements to allow IIoT

Mert Sehri and Patrick Dumond. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

<https://doi.org/10.36001/IJPHM.2025.v16i2.4372>

companies to integrate ML fault diagnosis in their devices and systems.

The proposed approach recommends data collection parameters for rolling element bearings by combining ML approaches and bearing signature theory. ML algorithms, such as convolutional neural networks (CNNs), enable more rapid and accurate fault diagnosis when provided with an adequate input. The selection of parameters such as sampling rate, number of fault harmonics, number of samples, data collection duration, frequency resolution, window length, and stride is made easier with the help of existing research based on bearing dataset benchmarks. Bearing signature theory, which sheds light on the distinctive patterns of defects in roller element bearings, allows for identifying the number of fault harmonics contained within a signal. Researchers are encouraged to use the findings in this paper to collect data for rolling element bearings specifically intended for use with ML algorithms, which will encourage the collection of a greater number of datasets and will also help improve AI-based fault diagnosis techniques for IIoT applications.

The proposed recommendations for collecting data used with ML approaches for rolling element bearings open a wide range of possibilities. Importantly, it encourages IIoT companies to develop devices that can be used with ML algorithms, as well as allowing researchers to collect data based on bearing signature theory, ensuring that the data gathered is in line with the specific requirements of the current problem. Storage space limitations can be overcome by optimizing the sample rate and quantity of data collected. The proposed recommendations establish a foundation for advancements in IIoT products and services and serve as a useful tool for further study in data-driven fault diagnosis.

1.1. Motivation

ML-based bearing fault diagnosis is a widely studied topic, particularly with the emergence of data driven methods such as neural networks (Sehri et al., 2023; Sehri, Varejão, et al., 2025). However, many existing studies omit critical details regarding the computational time, software frameworks, hyperparameters (Wu et al., 2019), and hardware configurations used to implement these methods, making reproducibility and fair comparison difficult (Vashishtha et al., 2025). This lack of transparency in the literature hinders the ability to evaluate algorithmic performance in real-world conditions (Sehri, Hua, et al., 2025). This work acknowledges this gap and incorporates a brief explanation to justify experimental design choices. Additionally, a majority of publicly available rolling element bearing datasets were created before the emergence of DL algorithms, often making them non-ideal for training modern DL models. This work identifies and addresses these limitations by proposing optimized data usage strategies for these datasets. This paper also highlights the importance of understanding both the data and AI-based

condition monitoring processes, aiming to bridge the gap between ML researchers and signal processing engineers. To ensure clarity and accessibility for a broader audience, this work includes an overview of neural networks and their application to fault analysis in rolling element bearings. The remainder of this paper is organized as follows: Section 2 describes the background for different methods and models used. Section 3 outlines the results for bearing fault diagnosis analysis using DL. Section 4 presents the discussion, and Section 5 concludes with key takeaways and future work.

2. BACKGROUND

A significant problem in developing ML models is identifying the dimension and size of the dataset required to properly train a particular model. The dimensionality of a dataset refers to the characteristics of each set of data, for instance, how many columns of data are present and how many samples are collected. The size refers to the number of distinct objects (e.g., rolling element bearings) that are used in collecting the data. The size of the dataset helps in assessing whether a dataset is large or robust. Some misunderstand the difference between large datasets and high-dimensional datasets. Large datasets are preferred over high-dimensional datasets because of their ability to obtain higher accuracy results when using ML algorithms (Muñoz-Terol et al., 2020). However, due to a lack of guiding principles for data collection, data scientists tend to collect data without consideration for these differences (Mazhar, 2021; Soomro et al., 2024). Moreover, most datasets that are publicly accessible have been collected in a lab environment, where typical data collected, consisting of load, temperature, vibration, and/or acoustic data, is obtained from sensors at the maximum sampling rate of the particular configuration used, rather than for any reasons related to ML algorithm training (Nasir & Sassani, 2021; Soomro et al., 2024). Due to limited resources, this often leads to the creation of high-dimensional datasets, but not large datasets with many distinct objects. This leads to the question of whether there exists a minimal set of parameters (e.g., sampling rate, number of samples, data collection duration, and stride) for the effective development and implementation of ML methodologies. To address this question, these parameters are considered in this study. A structured evaluation is carried out using multiple bearing datasets, where a 95% accuracy threshold is used as a benchmark to determine the minimum parameter values required for effective fault diagnosis.

With the introduction of new machinery datasets, a need to assess data reliability is raised. t-distributed stochastic neighbor embedding (t-SNE) is the most useful method for determining whether high-dimensional data is reliable, as it helps to cluster data into two-dimensional visualizations (van der Maaten & Hinton, 2008). This method can be used to visualize either raw data or data that

has been processed using an ML algorithm. Numerous studies use the t-SNE method to determine the reliability of rolling element bearing datasets (Duo et al., 2021; Jiang et al., 2022; F. Xie et al., 2023).

As a branch of ML, DL can be used for data analysis. Data classification accuracy for DL models are based on high-quality and large quantities of training data (Cho et al., 2016). The question then becomes how much data is required to train adequate ML models? Cho et al. evaluate the quantity of data needed to train DL algorithms by using medical image (CT scans) recognition to identify medical concerns. They conclude that small window lengths lead to high misclassification, while dataset sizes consisting of 100 to 200 images had similarly high accuracies, indicating that there is a minimum amount of data that must be collected for satisfactory results (Cho et al., 2016). Their findings indicate that it may be possible to define a particular set of data collection parameters for other ML tasks to achieve adequate results as well. On the other hand, Perez et al. concluded that for TL, more harmonics (968) outperformed the use of 8 fault signatures using k-nearest neighbours (KNN) and support vector machine (SVM) algorithms (Duque-Perez et al., 2019). Nonetheless, when artificial neural networks (ANNs) are compared to KNN and SVM algorithms, Ren concludes that ANNs outperform both (Ren, 2012). In this study, ANNs and CNNs are used as benchmarks for understanding the dataset requirements of TL and DL methods, respectively. Both TL and DL methods require significant training data to achieve high accuracy. Although most TL studies indicate that more fault signatures are needed for higher accuracy, they do not provide enough information to determine the minimum sampling rate, number of bearing fault harmonics, number of samples, data collection duration, signal resolution, window length, stride, and 2D CNN image resolution needed for “high enough” ML accuracies.

2.1. Parameter Selection

The selection of appropriate data collection parameters plays an essential role in creating a quality rolling element bearing dataset. Once bearing hardware has been selected, researchers must understand bearing fault frequencies to select accelerometers that will capture enough data to distinguish different fault types.

2.1.1. Bearing Fault Frequencies

A key component of data collection is understanding the signal of interest. Bearing test rigs often focus on the fundamental bearing fault frequencies since the controlled research lab environment allows for relatively clean data collection with minimal interference. In such cases, high-frequency resonant bands, which are typically much higher than the fundamental fault frequencies, can sometimes be overlooked. Unfortunately, in industry, fundamental bearing fault frequencies are often buried in machine and

environmental noise. Therefore, early detection often relies on more distinct high-frequency signal content. As highlighted by Randall and Antoni (Randall & Antoni, 2011), bearing faults often excite high-frequency structural resonances, which depend more on the machine setup, such as the test rig, rather than the bearing type itself. However, in lab test rigs, the data is generally less noisy, allowing researchers to focus on the fundamental bearing frequencies with fewer concerns about high-frequency resonance interference.

Nevertheless, moving to real world industrial environments introduces significant operational noise, making it crucial to consider these high-frequency resonant bands. In such cases, the high-frequency bursts caused by impacts due to faults excite resonant frequencies, and these signals are further modulated by the load and transmission path. Ignoring these high-frequency components can result in improper fault detection, as the fundamental fault frequencies and their harmonics can easily be masked by noise from other machinery (Randall & Antoni, 2011). Techniques like envelope analysis and spectral kurtosis become essential for capturing these high-frequency signals, which carry significant diagnostic information in noisy industrial environments (Randall & Antoni, 2011). While lab test rigs might allow for a focus on fundamental frequencies, ensuring that the frequency range in industry data collection includes these high-frequency resonant bands is currently crucial for effective fault diagnosis. However, due to upper frequency limitations in existing IIoT devices, as well as the lack of publicly available industrial datasets, this paper seeks to understand how ML algorithms interact with bearing fundamental frequencies to provide fault diagnosis. It hypothesized that ML algorithms will provide additional capacity to filter through noisy industry data rather than require high-frequency resonant bands to provide adequate performance.

Rolling element bearing fault frequencies play an important role in identifying bearing fault types. In most cases, researchers are interested in classifying bearing fault types. Sacerdoti et al. analyzed the signal obtained from their intelligent maintenance system (IMS) test rig using the bearing characteristic equations (1)-(4) (Sacerdoti et al., 2023). They specified that each sample obtained from the IMS test rig had a sampling rate of 20,480 Hz collected over 1 s at 2,000 RPM. 26,690 N of the load was applied radially to the bearing via a spring (Sacerdoti et al., 2023). Although the authors focused on bearing signals, they did not justify their reasons for using their sampling rate or collection time. To better understand bearing faults, Figure 1 a), b), and c) are provided to visualize the characteristic features of rolling element bearings.

Characteristic bearing fault frequency equations can be written as, where BPFO is the ball pass frequency of the outer race, BPFI is the ball pass frequency of the inner

race, FTF is the fundamental train frequency, and BSF is the ball spin frequency.:

$$BPFO = \frac{nf_r}{2} \left(1 - \frac{d}{D} \cos\phi\right) \quad (1)$$

$$BPFI = \frac{nf_r}{2} \left(1 + \frac{d}{D} \cos\phi\right) \quad (2)$$

$$FTF = \frac{f_r}{2} \left(1 - \frac{d}{D} \cos\phi\right) \quad (3)$$

$$BSF = \frac{f_r D}{2d} \left[1 - \left(\frac{d}{D} \cos\phi\right)^2\right] \quad (4)$$

where D represents the pitch diameter of the bearing, d is the rolling element (ball) diameter, n is the number of rolling elements and ϕ is the load contact angle.

Characteristic bearing frequencies are also usually associated with a set of harmonic frequencies in integer intervals of the characteristic frequency. Bearing fault frequency harmonics are known to be useful for understanding the type and severity of the faults contained within a signal. Bearing faults come from the formation of cracks or pits in bearing surfaces that cause impacts, which are reflected in vibration signals as a spike (Smith & Randall, 2015). These faults can also generate harmonics that are primarily driven by non-linear loads or conditions. Since these harmonics play such a key role in identifying differences between similar characteristic bearing fault values, providing enough data to identify these bearing fault harmonics is important. The number of fault harmonics available in a signal depends on the value of the characteristic bearing fault frequency and the sampling rate. The total number of fault harmonics contained in a signal can be calculated as half the sampling rate divided by the highest identified bearing fault frequency, as shown in equation (5).

$$\# \text{ of Fault Harmonics} = \frac{F_s/2}{F_{Hbearing}} \quad (5)$$

In the tutorial paper for “Rolling element bearing diagnostics” provided by Randall and Antoni, the use of 10 or more harmonics in capturing bearing fault frequencies is essential for identifying bearing faults using traditional methods (Randall & Antoni, 2011). Thus, within this tutorial framework, 10 fault harmonics serve as a minimum requirement for fault diagnosis by an expert when collecting data without the intended use of ML. As such, there is a similar need to determine the minimum number of harmonics required for performing bearing fault diagnosis via simple ML algorithms. This helps define the frequency range of interest during data collection.

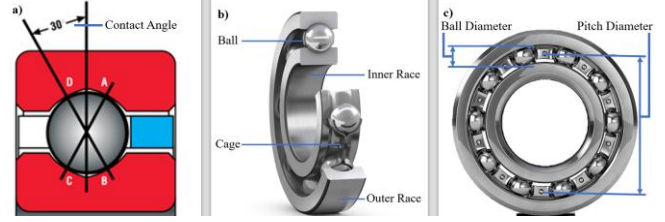


Figure 1. a) Contact Angle (“BALL BEARING BASICS AND TYPES,” n.d.), b) Rolling Element Component Identification (*Deep Groove Ball Bearings* | SKF, n.d.), c) Important Dimensions of the Rolling Element (*Skf Ball Bearing*, n.d.)

To examine the impact of bearing fault frequencies on data collection, a range of Svenska Kullagerfabriken (SKF) bearings with varying bore diameters were selected. This selection includes the entire spectrum of commonly available bearings, starting from the smallest bore size and extending to the largest bore diameter available. The goal was to observe how different bearing frequencies (i.e., BPFI, BPFO, BSF, and FTF) would affect the data collection process.

The SKF 6000 bearing is selected to be the smallest diameter used frequently in the industry. This helps set the frequency range possible (6000 - *Deep Groove Ball Bearings* | SKF, n.d.). Although smaller bearings are manufactured by SKF and are used for special applications requiring high speeds, they are not often used in industry. The SKF 6000 dimensions include a 26 mm outer diameter, a 10 mm bore diameter, a pitch diameter of 18 mm, and a ball diameter of 6 mm. The bearing contains 9 balls and is rated for a maximum rotation speed of 40,000 RPM. Loads are assumed to be applied only in the radial direction at a maximum motor speed. In this case, characteristic bearing frequencies include a BPFO of 2,166.7 Hz, a BPFI of 3,833.3 Hz, an FTF of 240.7 Hz, and a BSF of 1,107.4 Hz. SKF’s largest single-row ball bearing in common use, the 708/1250AMB (*708/1250 AMB - Angular Contact Ball Bearings* | SKF, n.d.), has an outer diameter of 1500 mm, a bore diameter of 1250 mm, a pitch diameter of 1375 mm, and a ball diameter of 55 mm. It contains 14 balls and is rated for a maximum speed of 280 RPM. Loads are assumed to be applied only in the radial direction at 280 RPM. Based on this information, the largest bearing will have a BPFO of 31.4 Hz, a BPFI of 34.0 Hz, an FTF of 2.2 Hz, and a BSF of 58.2 Hz. From equations (1)-(4), rotational speed can be seen to have the most significant effect on bearing frequencies for data collection, especially since the number of balls cannot be adjusted due to each manufacturer’s specifications.

The fault frequency range of SKF bearings is obtained by calculating the characteristic bearing frequencies in the bore range of 10 to 1250 mm. Tests were only conducted at typical industry speeds to determine the bearing fault with the highest frequency. Figure 2 provides a sub-set of results

for the highest fault characteristic frequencies obtained at common AC and DC motor speeds (Nagel, 2018) used in industry. Since the largest common bearings operate well below 900 RPM and have lower fault frequencies, they are not included in this figure. The inner race is specifically chosen due to it having the highest frequency among the four fault signatures. Figure 2 exhibits the highest frequency of 490 Hz at 5000 RPM, representing a single fault harmonic. This frequency aligns reasonably well with industry expectations and can be considered as the typical maximum fault signature frequency. It is important not to forget, though, that fault harmonics play an important role in the diagnosis of bearing faults when using ML algorithms. A few factors must be considered: differentiating between different fault types and noise, reliability, and redundancy. Differentiating between different fault types and noise in bearing data is important so that enough fault harmonics are available to classify differences between each class type when trying to obtain high ML algorithm accuracies. Reliability and redundancy refer to gathering enough fault harmonics to have a clear, distinguished signal. This is crucial for effectively differentiating between potential noise in data and bearing fault signatures. The goal is to ensure that there are enough fault harmonics included in the signal to enable the network to recognize and accurately respond to various fault conditions. Reliability involves the consistent and accurate performance of the system in identifying and distinguishing between healthy and faulty conditions. Redundancy, on the other hand, consists of additional components and measures beyond the essential requirements, providing backup functionality to strengthen overall system reliability. In this way, the combination of reliability and redundancy contributes to a robust fault detection mechanism.

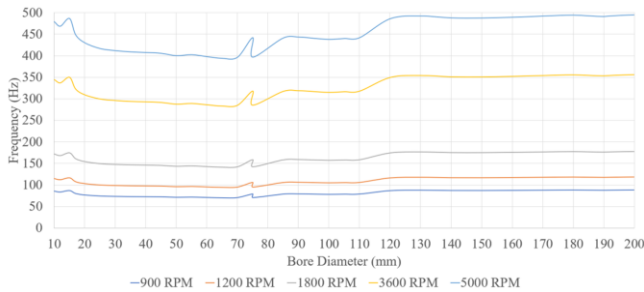


Figure 2. Frequency vs. Bore Diameter for SKF Bearing Inner Race Frequencies with Various Motor Speeds, 10 mm to 200 mm

2.1.2. Sampling Rate

The sampling rate (sampling rate) indicates the rate at which a continuous analog signal is converted to a discrete digital signal. Continuous, discrete signals are signals obtained from sensors and measured in differences of voltages. The sampling rate is an important parameter for sensor selection (e.g., accelerometers and microphones) (Delaunay et al., 1994). For bearing data collection applications at lower speeds, a low pass filter (LPF) is

required before converting the analog signal to remove higher frequency components from the signal that do not aid in understanding the health state of the bearing (X. Zhang et al., 2012). This helps to avoid aliasing by removing high frequency artifacts. If an LPF is not applied, noise (including electrical interference) or other signal disturbances can occur, as well as artifacts introduced by aliasing of the signal. Aliasing occurs when the sampling rate is less than the minimum sampling rate required to represent the analog (original) signal obtained from the sensor.

To prevent aliasing and to select an appropriate sampling rate, the Nyquist-Shannon sampling theorem must be applied to the signal (Nyquist, 1924; Shannon, 1948). According to the theorem, the sampling rate must be greater than two times the analog input's highest frequency of interest to prevent aliasing (Nyquist criterion) (Kester, 2023). If samples are collected below this frequency, as seen in Figure 3, aliasing will corrupt the digitized signal. For instance, if a sensor (e.g., accelerometer) has a maximum frequency capacity of 15 kHz, and all information is intended to be captured, then according to the Nyquist criterion, the sampling rate should be greater than 30 kHz. This paper investigates the effect of sampling rate on the fault diagnosis accuracy by analyzing different existing bearing datasets that have different sampling frequencies using ML algorithms. Setting a threshold accuracy is required to ensure a reduction in the requirement for high-dimensional datasets to avoid prolonged algorithm running times and lower hardware requirements (Dini et al., 2024; Rahman et al., 2023; X. Zhang et al., 2021). There are only a few ML applications of IIoT-bearing diagnosis that can train or update ML models within less than 5 minutes with high accuracy, as most articles refrain from publishing computational times (Z. Chen et al., 2024; Shao et al., 2018; Y. Wang et al., 2020).

According to Xie et al., DL CNN models should achieve a minimum accuracy of 95% or higher in both training and validation datasets (W. Xie et al., 2022). Although the threshold accuracy is case-dependent for selected machinery components, when it comes to bearing diagnosis, numerous IIoT research articles indicate that a validation accuracy above 95% is satisfactory based on results achieved (Asutkar & Tallur, 2023; Djaballah et al., 2024; Kumar et al., 2022; Sun & Gao, 2024; J. Xu et al., 2022). Bearing's operate in industry under high-stake environments in terms of machine failure, leading to costly downtimes and potential equipment damage when they fail (Bloch & Geitner, 2012; Karabay & Uzman, 2009; Theissler et al., 2021). Therefore, fault diagnosis models should be performed with high accuracy to minimize incorrect detection (Z. Gao et al., 2015; Isermann, 2006). This paper will seek to justify the threshold accuracy that can be reasonably attained for bearing fault diagnosis using standard ML algorithms based on test accuracies of bearing conditions and computational times. This will be obtained

by comparing TL and DL algorithms with different existing bearing datasets as benchmarks. For instance, if a specific architecture cannot reach a targeted threshold accuracy on noisy bearing data but can on clean bearing data, it shows the limitation of the model in handling real world data, which is a step towards understanding noisy industrial scenarios. It is expected that more complex algorithms should provide better fault diagnosis results, but these often come at the cost of hardware requirements and additional computational time. Therefore, for the remainder of this study, ML algorithm performance will be considered satisfactory based on the results obtained from the algorithms described herein.

2.1.3. Number of Samples and Data Collection Duration

The number of samples depends on the accelerometer selected, the number of harmonics, machine component frequencies, and data from the scientist's expertise. The maximum frequency capacity of a sensor can be determined from the manufacturer's specifications or the required maximum frequency of interest, and by applying the Nyquist criterion, the minimum sampling rate is determined. Using the resolution formula, different numbers of samples are selected and tested for existing datasets. The number of samples and the duration of data collection are related to each other via the sampling rate. To know the duration of data collection required (in seconds) to collect the appropriate number of samples, the number of samples required must simply be divided by the sampling rate. Data quantities and durations are investigated by performing bearing fault diagnosis using different network types to attain a minimum threshold accuracy.

2.1.4. Frequency Resolution

In machine condition monitoring, converting time-domain signals into the frequency domain is crucial for gaining additional insight from a mechanical system. This transformation is frequently done using techniques such as the fast Fourier transform (FFT) or other spectral analysis. These methods transform time-domain signals (which show a signal's amplitude over time) into frequency components. This conversion allows for a more extensive evaluation of the signals' spectral content, providing a better understanding of the root cause of a machinery's dynamics.

Frequency resolution plays an essential role in identifying intricate details within the gathered bearing data that has been converted to the frequency domain. For bearing diagnosis, frequency resolution holds significant importance, impacting the precision of bearing fault analysis directly. Notably, increasing resolution helps differentiate between fault types such as inner race, outer race, ball, and cage faults. In the context of data collection for ML purposes, one must acknowledge that data resolution stands as a critical parameter, inherently dependent upon the sampling rate and the number of samples collected. The

sampling rate is the number of points required to convert an analog signal to a digital signal, while the quantity of samples defines the dimensionality of the dataset, both of which have a large effect on the total dataset size. According to Duan et al., increasing the number of samples results in a finer frequency resolution, as indicated by equation (6) (Duan et al., 2019).

$$\Delta\omega = \frac{F_s}{N} \quad (6)$$

where $\Delta\omega$ is the frequency resolution, F_s is the sampling rate, and N is the total number of samples. Consequently, different window lengths will be tested to account for this relationship. To have an enhanced resolution in the frequency domain, the number of samples collected in the time domain should be increased while maintaining a sampling rate that satisfies the Nyquist criterion. Having an enhanced resolution means a finer detailed signal, which can help distinguish between different bearing fault frequencies. However, the goal is to identify if there is a minimum resolution required to reduce the number of samples collected while ensuring high accuracy. Different bearing datasets are analyzed using ML algorithms to observe if a causal relation exists between the effects of frequency resolution and the network's accuracy. To understand how frequency resolution affects fault identification accuracy, ANN and 1D-CNN algorithms are utilized.

2.2. ML Algorithms

2.2.1. Neural Network Specifications

Basic TL and DL algorithms used for this study are summarized in Table 1, including references to the algorithms used for each implementation. The table also provides information on the number of hidden layers and network classifier type. Based on referenced ML research, obtaining high training and validation accuracies relies on the quality of the data itself. Therefore, sensor placement must be carefully considered to minimize noise and to obtain reliable signals.

Additionally, for algorithm selection in Table 1, an ANN is categorized as a TL algorithm, while methodologies employing a 1D-CNN or 2D-CNN are classified as DL algorithms. They are chosen to explore algorithm hyperparameter tuning by employing the simplest configuration with minimal hidden layers.

Batch size refers to the number of training samples processed together in one iteration during ML algorithm training. It helps in balancing computational efficiency with training stability and convergence speed. Larger batch sizes can expedite training but may require more memory, while smaller batch sizes may lead to faster convergence but could increase training time. Selecting a batch size involves finding a balance between these factors to ensure efficient and effective model training. Based on ML researcher findings, the following hyperparameters are selected; the

batch size set to 64 (C.-C. Chen et al., 2021; T. Kim & Chai, 2021; Neupane & Seok, 2020; Wu et al., 2020; R. Zhang & Gu, 2022), classifier as Softmax (Hoang & Kang, 2019; C. Lu et al., 2017; Neupane & Seok, 2020; Shao et al., 2018; J. Zhang et al., 2020; W. Zhang et al., 2017; Zilong & Wei, 2018), 100 epochs (Althobiani, 2024; Che et al., 2021; Wei et al., 2021; Wu et al., 2020), and a learning rate of 0.01 (Gu et al., 2022; Wen et al., 2021; R. Zhang & Gu, 2022; Zhiwei, 2022) which are determined based on high accuracies achieved with the CWRU bearing dataset. A batch size of 64 balances memory efficiency and training stability, allowing for faster convergence without overloading system memory. A Softmax classifier is better for multi-class problems because it provides normalized probabilities for each class, making sure that the model predicts each class with the highest confidence. 100 epochs provides a sufficient number of iterations for the model to learn complex patterns while avoiding overfitting of data. Lastly, a learning rate of 0.01 controls the speed at which the model updates its weights, providing a balance between fast convergence and avoiding overshooting. This learning rate is preferred over 0.001 or 0.0001 since it allows for faster convergence while maintaining pattern identification for bearing classes. Therefore, different batch sizes, classifiers, learning rates, and number of epochs are not explored in this study.

In all networks, the ReLU activation function was used in the hidden layers to introduce non-linearity and prevent vanishing gradients, given the two shallow hidden layers of the networks studied. The Softmax activation function was used only at the output layer to handle multi-class classification by providing class probabilities. Although deeper networks may require more careful handling of gradient flow, the shallow configurations used in this study did not exhibit vanishing gradient issues.

Table 1. ML Algorithms Used for TL and DL

Ref	Algo.	# of Hidden Layers	Batch Size	Classifier	Epoch	Learning Rate
(Samanta & Al-balushi, 2003)	ANN	2	64	Softmax	100	0.01
(X. Wang et al., 2021)	1D-CNN					
(S. Yang et al., 2022)	2D-CNN, grayscale images					

* Ref- references, Algo- algorithms

2.2.2. Window Length

The window length is a time series hyperparameter used in ML algorithms that represents the input size for training neural networks. The window length is what enables the capture of patterns created by bearing faults when training neural networks. In ANNs and 1D-CNNs for

time domain analysis, window length refers to the number of data point segments that are fed into the network at any given time for training purposes. For 2D-CNNs using grayscale images, window length is the filter or kernel's dimensions that are used to define the image characteristics. Matrix dimensions are a common way of defining a kernel's size. For instance, a kernel can measure 20x20, meaning the image created has 20 pixels along each axis. The created images are then fed into the convolutional layer for processing. The window length is used to define how many samples are pushed into the neural network for training and testing of the data. A window length can impact the validation accuracy results of a network due to insufficient features captured if not selected correctly.

2.2.3. Stride

Stride is a hyperparameter used in ML algorithms that reduces the spatial dimensions of processed data by allowing overlapping of the window to occur on the existing signal during ANN and CNN operations (Hendriks et al., 2022). Stride is useful when handling datasets with limited quantities of specific samples (small-sized datasets), such as those obtained from bearings tested in a lab. Stride is measured as the number of time domain segment points that overlap the window length for ANNs and 1D-CNNs, as well as the number of pixels that the kernel shifts over the image during each convolution for 2D CNN grayscale images. A smaller stride allows the usage of smaller datasets. A smaller stride value means a higher overlap in the data that is used as input during network training. As the stride is reduced, it is assumed that the network accuracy will increase so that the network should be able to identify more similarities between images at the expense of computation time. However, an ideal stride should be determined based on different bearing datasets to reach a certain network threshold accuracy. Stride will be investigated by selecting different stride values and applying them to different bearing datasets used for training different types of neural networks. Suggested values for stride will be determined based on the minimum threshold accuracy.

2.2.4. Image Resolution for 2D CNNs

When raw bearing data is converted to grayscale images in the frequency domain, the resolution is then dependent on the quality of the created image and can change as the data is passed through the network from one convolutional layer to the other, as the data works through operations such as pooling and stride. These operations reduce the dimensions of the feature maps. Therefore, the resulting resolution of the output images is directly dependent on the network architecture used by the algorithm and cannot be determined explicitly without knowledge of the specific structure used. The number of grayscale images per class for testing is justified based on existing research on the CWRU dataset by identifying research results that have

the highest obtained fault diagnosis accuracies. The number of grayscale images used for training and testing that have the highest accuracies per class ranges from 100 to 500 images (Han & Jeong, 2020; He et al., 2023; She et al., 2024; Wei et al., 2021). Additionally, 10 to 100 images are also tested for comparison in an attempt to further reduce computational requirements. Evidently, as the number of images created increases, so does computational time.

2.2.5. Training and Testing Split of the Data

For each bearing dataset used in an algorithm, the data is split using Hendriks et al.'s framework for training and testing (Hendriks et al., 2022; Unal et al., 2014). This framework is designed to prevent data leakage by using test samples that come from different operational conditions or machines than those used for training, thereby providing a more realistic evaluation of model generalization. Unlike traditional k-fold cross-validation, which may not respect distribution shifts, this approach enforces domain separation to mimic real-world scenarios.

2.3. Existing Datasets

Table 2 reveals that most rolling element bearing datasets that are currently openly accessible and typically used in the literature were created before the publication of LeCun et al.'s seminal DL paper in 2015 (LeCun et al., 2015). This suggests that DL was not considered when determining the number of samples and sampling frequencies used in the creation of these datasets. In all cases, the emphasis was on collecting a substantial number of samples without undergoing a justification process related to ML. For instance, in the case of bearing data under constant load and speed conditions (1400-1948 RPMs), CWRU and UORED-VAFCLS data collection was conducted for 10s without providing a justification for this timing. This extended duration was likely chosen to ensure that a wide range of features were captured within the signals, resulting in high-dimensional datasets rather than large and robust datasets.

The three datasets were chosen for analysis based on specific criteria: data available for performing diagnosis, constant speeds, and data collection duration of at least 10s to ensure a wide range of available test configurations. The CWRU and UORED-VAFCLS datasets were selected because they exhibit similar speeds and bearing characteristics, making them suitable for comparison. Additionally, the HUST dataset was included due to its diverse range of bearing dimensions while also providing a different constant speed variation for testing purposes. Due to the overall similarity among the three selected bearing

datasets in terms of signal characteristics and fault types, the assignment of each dataset to a specific parameter optimization task was made arbitrarily. The UORED-VAFCLS dataset was used to determine the optimal stride and window length parameters, the CWRU dataset was used to determine the optimal number of fault harmonics, and the HUST dataset was used to identify the optimal duration and frequency resolution parameters. While each dataset could theoretically serve any of these roles, this distribution was chosen to balance the experimental workload and ensure diversity in validation sources.

Table 2 displays a range of sampling frequencies, spanning 12 to 51.2 kHz, and data collection times, depending on the dataset. To further optimize bearing fault diagnosis, a more in-depth analysis is conducted on the UORED-VAFCLS, CWRU, and HUST datasets to determine the number of data samples that should be collected to attain the threshold accuracy. The values for data collection duration are chosen based on those most used in the literature by ML researchers, including values between 1 (Sacerdoti et al., 2023), 4 (*Konstruktions- und Antriebstechnik (KAT) - Data Sets and Download (Universität Paderborn)*, n.d.), and 10s (Gousseau et al., 2016) (Table 2). 2, 3, and 5 s collection times are also included for completeness. Three stride values are selected (3% (Hendriks et al., 2022), 10% (De la Fuente et al., 2021), and 50% (De la Fuente et al., 2021)) to assess whether comparable accuracies can be achieved using less data. Specifically, smaller stride values indicate a larger overlap that should lead to higher accuracy, which leads to more data usage when training networks. In this case, 10% and 50% are selected in trying to use the least amount of data to train the networks, while 3% is chosen to see the effect of using more data on the network's accuracy. However, it is worth noting that when converting data into grayscale images, the resolution changes significantly based on image size, making it more challenging for neural networks to detect faults accurately.

3. METHODOLOGY

The background section provided a concise overview of established concepts and datasets necessary for understanding this study, such as neural networks, fault analysis of rolling element bearings, and data collection parameters. This section now turns to the original contributions of the work. It begins with the research procedure, outlining the approach used to evaluate dataset requirements and algorithm performance. This is followed by the selection and justification of hyperparameters for ANN and CNN models, which are tailored to the goals of this study. Together, these elements define the experimental

Table 2. Comparison of Bearing Fault Signature Datasets.

Dataset	Date	Signal type	SF ^a (kHz)	# of samples	Collection time (s)	Motor Speed Range (RPM)	Analysis type	Fault type (A ^b , N ^c)	Bearing ID
CWRU	2012	V ^d	12.0	121,265	10	1720- 1797	Diagnosis	A	6205, 6203
CWRU	2012	V ^d	48.0	487,384	10	1720- 1797	Diagnosis	A	6205, 6203
HUST	2023	V ^d	51.2	512,000	10	1400-1500	Diagnosis	A	6204, 6205, 6206, 6207, 6208
UORED-VAFCLS	2023	V ^d , A ^e , T ^f	42.0	420,000	10	1700- 1948	Diagnosis	N	Nippon Seiko Kabushiki- 8 gaisha (NSK) & FAFNIR 6203

Legend: SF- Sampling frequency, ^b A- artificial, ^c N- natural, ^d V- vibration, ^e A-acoustic, ^f T- temperature

framework that links the theoretical considerations introduced earlier with the results and discussion that follow.

3.1. Research Procedure

The first step is to define the hyperparameters for evaluating model performance. Subsequently, a diverse range of stride and window lengths are chosen, covering a spectrum from small to large, to comprehend how model performance scales with data volume. The procedure utilized during this study is developed to determine the minimal parameter values required to train ANNs and CNNs for attaining a threshold accuracy when performing rolling element bearing fault diagnosis, while also considering the balance between data efficiency and the risk of overfitting. These values are intended as lower bounds for effective training rather than optimal values. To determine the optimal stride and window length for ML algorithms, including both ANNs and CNNs, the UORED-VAFCLS dataset is used to achieve the threshold accuracy needed to refine these parameters. The CWRU dataset is then used to establish the number of fault harmonics required to attain this same threshold accuracy. Finally, the HUST dataset is used to fine-tune the data collection duration and frequency resolution, employing the same ANN and CNN algorithms. The study also considers specific sampling rates, the number of samples, and image resolution requirements for 2D CNNs. The robustness of these findings is assessed for generalizability to other datasets and similar problems. The iterative nature of the methodology allows for refinement and adjustment as needed.

3.2. Hyperparameters for Bearing Fault Frequency Analysis Using TL and DL Algorithms

In section 2.3, Table 2 tabulates openly accessible datasets that can be used to analyze bearing fault frequencies. Among these datasets, inner race faults are consistently responsible for the highest characteristic bearing fault frequency at 150 Hz, as determined by the bearing frequency equations (1)-(4). The largest common fault harmonic is identified by determining the fault frequencies of all bearings using the methodology described in section 2.1.1. This is achieved by examining the datasets provided below and considering the bearings and speeds they correspond to. The sampling rate is divided by the characteristic bearing fault frequency, and after applying the Nyquist criterion in equation (5), the resulting number of fault harmonics that are likely captured in each dataset are shown in Table 3.

The sampling rate of the setup is a crucial factor, which varies depending on the researcher's choice and the manufacturer's specifications. Different sampling rates are used for the datasets considered in this study, and these are detailed in Table 3 as well. This study seeks to determine

whether the lowest sampling rate used for these datasets is sufficient for fault identification using ML algorithms.

Data sample collection durations, which are another essential parameter, were selected based on commonly used durations in the literature by ML researchers. These intervals include 1 (Sacerdoti et al., 2023), 4 (*Konstruktions- und Antriebstechnik (KAt) - Data Sets and Download (Universität Paderborn)*, n.d.), and 10s (*Konstruktions- und Antriebstechnik (KAt) - Data Sets and Download (Universität Paderborn)*, n.d.) as primary choices. For a comprehensive analysis, we have also included intervals of 2, 3, and 5 s. The number of samples in each dataset is directly proportional to the data sample collection duration and is calculated as the product of the sampling rate and the data collection duration. The resolution of each dataset is dependent on the sampling rate and number of samples, as described in Table 3.

In section 2.2.1, three configurations of traditional and DL algorithms were identified to evaluate performance in analyzing bearing fault frequencies using the datasets in Table 3. Among these configurations, the window length and stride were highlighted by researchers for their distinct effects on each model's ability to detect fault

Table 3. Characteristics of the Rolling Element Fault Frequencies for a Subset of Openly Accessible Datasets.

Dataset	Number of Samples	Sampling rate	Resolution	Duration of Data Intervals (s)	Number of Fault Harmonics
UORED-VAFCLS (Sehri et al., 2023; Sehri & Dumond, 2023)	42,000	42,000	1.00	1	140
	84,000		0.50	2	
	126,000		0.33	3	
	168,000		0.25	4	
	210,000		0.20	5	
	420,000		0.10	10	
CWRU (Apparatus & Procedures Case School of Engineering Case Western Reserve University, 2021)	12,000	12,000	1.00	1	40
	24,000		0.50	2	
	36,000		0.33	3	
	48,000		0.25	4	
	60,000		0.20	5	
	120,000		0.10	10	
CWRU (Apparatus & Procedures Case School of Engineering Case Western Reserve University, 2021)	48,000	48,000	1.00	1	160
	96,000		0.50	2	
	144,000		0.33	3	
	192,000		0.25	4	
	236,000		0.20	5	
	480,000		0.10	10	
HUST (Thuan & Hong, 2023a)	51,200	51,200	1.00	1	170
	102,400		0.50	2	
	153,600		0.33	3	
	204,800		0.25	4	
	256,000		0.20	5	

harmonics, as determined by the input sizes and data training/testing splits described in Table 4.

Window length is an important parameter for 2D grayscale image-based CNNs. Therefore, different window lengths are selected for analysis (Table 4). These include window lengths of 512, 1024, 2048, and 4096 samples,

which are based on the most used parameters by scientists to contain sufficient information regarding bearing signals (Ruan et al., 2023; Y. Yang et al., 2018; S. Zhang et al., 2021). Additionally, stride percentages of 3%, 10%, and 50% were chosen for the evaluation mentioned in section 2.3.

The input size for each model was directly influenced by window length and stride combinations, resulting in four distinct configurations, as presented in Table 4. Training and testing data splits were chosen to be 80% and 20%, respectively, to ensure a balance between model training and evaluation (Chuya et al., 2022; Pham et al., 2020).

The number of images generated for training each model varied based on window length and stride configurations, with values ranging from 10 to 500 images, providing a wide range of values for exploring their effect on performance. Furthermore, the number of epochs, representing the number of training iterations over the entire dataset, was fixed at 100 for all configurations to ensure consistent convergence of the models (Althobiani, 2024).

The parameters listed in Table 4 correspond to the ML models defined in Table 1. For TL, the raw vibration signals were flattened and used directly as inputs to the ANN model. For DL, the 1D-CNN used windowed raw time-series inputs, while the 2D-CNN was trained using grayscale spectrogram images derived from the same signals. The input sizes listed here reflect the dimensions fed into each model, depending on the type of preprocessing used.

Table 4. Parameters of the TL and DL Algorithms Used in This Study

Window Length	Stride (%)	# of Images	Input Size
512	3, 10, 50	10, 20, 30, 40,	25 x 20
1024		50, 60, 70, 80,	20 x 50
2048		90, 100, 150,	40 x 50
4096		200, 250, 500	50 x 80

4. RESULTS

In this section, the aim is to narrow down the values selected in Table 3 to attain a threshold accuracy when using the UORED-VAFCLS, CWRU, and HUST datasets. provides the characteristic values for the three datasets used for this test.

For the UORED-VAFCLS dataset, to capture acoustic sensor data within the human audible range of 20 Hz to 20 kHz and by applying the Nyquist Theorem (Kester, 2023), a sampling rate of 42,000 Hz is selected by Sehri et al. (Sehri & Dumond, 2023). On the other hand, the CWRU test rig uses data collection sampling rates of 12 and 48 kHz, while the HUST dataset uses 51.2 kHz. All three datasets need to be checked for reliability. While the CWRU (S. Gao et al., 2022; He et al., 2021; Hou et al., 2023; H. Lu et al., 2023; Mahesh et al., 2024; W. Xu & Li, 2022) and HUST (Hou et al., 2023; F. Lu et al., 2024; Y. Xu et al., 2024) dataset reliabilities have already been measured in

previous studies, the UORED-VAFCLS dataset has yet to be verified.

Figure 3 shows UORED-VAFCLS inner race data from different bearing tests using the t-SNE technique on ANN results. This statistical method is used to arrange high-dimensional data in a 2-dimensional space. By using t-SNE, embedded features can be visualized, allowing for the assessment of the effectiveness of ML methodologies (Duo et al., 2021). Moreover, t-SNE helps to determine the statistical robustness of a dataset by ensuring that the dataset is compactly separated without overlapping clusters (Arora et al., 2018).

Based on a visual analysis of Figure 3, the clusters appear compact and well-separated, with only a few overlapping outliers among the inner race fault data points. Similar results are obtained for the other fault types. Hence, it can be concluded that the UORED-VAFCLS dataset is statistically robust and reliable when using the ANN algorithm.

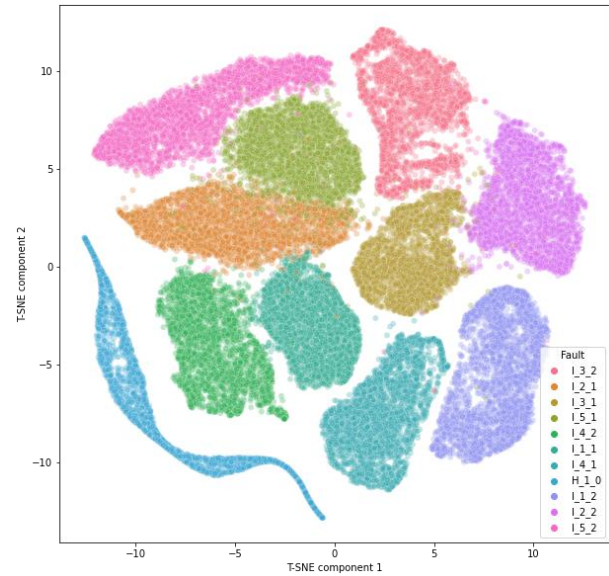


Figure 3. UORED-VAFCLS: Accelerometer Inner Race t-SNE Plot When Using an ANN

4.1. Validation Accuracy Values for Window Length and Stride Using the UORED-VAFCLS Dataset

This section aims to narrow down window length and stride by using ANN and CNN algorithms in conjunction with the UORED-VAFCLS dataset. The dataset includes various bearing fault types contained within accelerometer and acoustic sensor data. By conducting tests with different window lengths and strides and assessing their impact on the performance of ML models, the aim is to assess the effects of window length and stride on fault diagnosis accuracy. This investigation starts with an exploration of dataset characteristics, followed by an analysis of how varying window lengths and strides influence the accuracy

of basic ML algorithms.

Table 5 shows the UORED-VAFCLS inner race results for the number of fault harmonics, number of samples, data collection duration, frequency resolution, and window length while also considering a 3%, 10%, and 50% stride. To achieve threshold accuracy levels above 95% for both accelerometer and acoustic data when using ML, a total of 140 fault harmonics, 210,000 data samples for each set of data, a data collection duration of 5 s, and a resolution of 0.20 are necessary for both ANN and 1D-CNN models. Specifically, a window length of 1024 is required, with a 10% stride. Notably, when the stride is increased to 50%, the accuracy of the algorithms significantly decreases, indicating that a smaller data stride yields more accurate results for this case due to more overlap. Conversely, a 3% stride demonstrates a minimal accuracy improvement at the expense of a significant increase in computational time and resources.

Inner race results obtained from applying an ANN [65] to the UORED-VAFCLS are presented in Figure 4. The validation accuracy of the model reaches 95.13% after 100

significantly after 20 epochs. The results were obtained for a sample that used a 10% stride a window size of, 1024 samples, and a split of 80% (336,000 samples) for training and 20% for validation (84,000 samples). The training process utilized a batch size of 64.

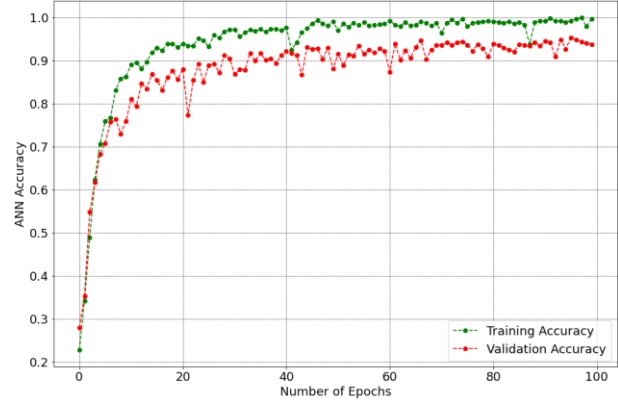


Figure 4. UORED-VAFCLS: Accuracy Results of Accelerometer Data for Inner Race Faults and Healthy Data

Table 5. UORED-VAFCLS Inner Race Dataset Results and Test Cases for Features in the Time Domain for ANN and 1D-CNN

# of Fault Harmonics [a], Number of samples [b], Data Collection Duration (s) [c], Freq. Resolution [d], Window Length [e]					ANN, BPFL, Stride %, Validation Accuracy (%), Average Run Time (minutes)								1D-CNN, BPFL, Stride % Validation Accuracy (%), Average Run Time (minutes)						
a	b	c	d	e	Acc., 3%	Aco., 3%	Acc., 10%	Aco., 10%	Acc., 50%	Aco., 50%	Time (min)	Acc., 3%	Aco., 3%	Acc., 10%	Aco., 10%	Acc., 50%	Aco., 50%	Time (min)	
140	42,000	1	1.00	512	79.26	77.02	77.60	73.06	39.07	44.54	1.53	38.49	66.18	20.81	57.89	25.65	49.73	2.26	
	84,000	2	0.50		90.51	85.45	87.67	82.01	54.56	53.47	3.32	89.37	90.12	86.36	75.91	33.06	55.65	4.51	
	126,000	3	0.33		91.77	89.37	88.54	86.54	52.65	58.28	4.51	87.13	89.75	86.96	71.91	35.10	67.15	6.24	
	168,000	4	0.25		95.99	89.25	94.23	87.01	66.42	66.28	6.10	89.84	91.84	90.94	81.38	24.76	59.19	8.10	
	210,000	5	0.20		95.13	89.84	95.10	89.43	67.35	71.85	8.26	93.77	90.18	91.45	83.58	28.96	70.28	10.36	
	420,000	10	0.10		95.41	95.55	95.68	93.29	81.88	77.74	9.57	94.59	92.41	93.46	85.94	68.50	71.07	12.12	
140	42,000	1	1.00	1024	67.58	66.79	63.41	63.53	27.07	25.97	1.42	73.45	80.67	61.20	70.29	33.70	39.78	2.12	
	84,000	2	0.50		79.37	78.66	77.11	79.41	34.43	41.26	2.15	78.60	84.13	62.61	62.65	14.48	51.64	3.46	
	126,000	3	0.33		88.91	87.41	85.78	86.80	33.61	50.36	2.50	85.81	87.76	71.30	73.82	17.49	62.00	3.57	
	168,000	4	0.25		92.16	93.18	88.65	90.94	54.97	57.96	3.34	91.53	94.17	86.12	85.33	51.84	65.71	4.10	
	210,000	5	0.20		95.61	95.78	95.13	95.04	59.24	64.78	3.41	95.85	95.69	95.78	95.15	60.11	59.89	4.51	
	420,000	10	0.10		95.43	95.84	95.46	95.91	70.07	79.01	6.32	95.13	95.73	95.87	95.68	69.58	73.70	7.14	
140	42,000	1	1.00	2048	53.26	42.71	49.09	38.64	30.68	26.14	1.35	57.13	60.38	35.00	55.00	18.18	29.55	1.58	
	84,000	2	0.50		64.52	67.24	61.31	65.19	35.36	28.18	2.05	60.54	67.71	55.65	70.95	35.91	40.88	2.12	
	126,000	3	0.33		67.28	74.05	60.86	74.05	25.41	33.70	2.41	63.87	71.43	58.76	74.12	42.54	57.88	3.25	
	168,000	4	0.25		82.01	87.11	80.67	85.16	36.61	42.08	3.13	67.89	74.46	44.47	71.47	56.28	60.38	3.57	
	210,000	5	0.20		84.34	93.45	82.52	92.07	64.48	49.13	3.40	92.88	86.74	85.09	79.72	35.15	67.90	5.01	
	420,000	10	0.10		88.71	95.83	85.95	95.54	77.64	70.65	6.10	94.71	89.67	93.30	86.39	71.52	74.46	6.37	
140	84,000	2	0.50	4096	42.45	41.66	39.77	40.23	14.86	11.36	1.33	20.23	62.47	16.84	58.45	18.18	31.82	1.35	
	126,000	3	0.33		53.49	55.78	50.35	52.46	25.12	22.22	2.03	42.51	64.18	38.86	64.01	34.92	40.00	2.01	
	168,000	4	0.25		81.22	71.93	77.17	67.18	34.56	23.20	2.36	61.49	71.33	56.36	67.49	45.36	42.54	3.24	
	210,000	5	0.20		91.47	86.34	88.92	83.77	61.47	30.84	3.38	81.72	79.44	75.20	70.48	68.45	55.07	4.35	
	420,000	10	0.10		95.11	95.70	95.47	95.81	79.56	44.54	6.03	89.66	83.35	87.37	70.39	72.89	78.56	6.15	

Legend: Acc. Is the accelerometer, Aco. is acoustic.

epochs. Notably, improvements in accuracy reduce

vs Number of Epochs Using an ANN

Table 6. UORED-VAFCLS Inner Race Dataset Results for Image and Test Case Features for 2D-CNN

# of Fault Harmonics [a], Number of samples [b], Data Collection Duration (s) [c], # of Images per Class [d], Input Size [e]					2D-CNN, BPFI, Stride %, Validation Accuracy (%), Average Run Time (minutes)						
a	b	c	d	e	Acc., 3%	Aco., 3%	Acc., 10%	Aco., 10%	Acc., 50%	Aco., 50%	Time (min)
140	42,000	1	50	25x20	70.31	67.44	78.42	73.71	36.74	52.46	17.33
	84,000	2	100		90.16	88.49	89.52	85.68	45.90	54.97	19.17
	126,000	3	150		87.13	89.75	88.89	89.17	58.23	63.08	20.35
	168,000	4	200		88.84	90.71	92.29	90.61	74.29	67.30	21.22
	210,000	5	250		95.72	93.35	95.41	91.69	75.05	72.78	23.12
	420,000	10	500		95.43	94.77	95.85	93.48	86.62	82.45	25.34
140	42,000	1	50	20x50	87.81	83.79	84.59	75.61	41.99	48.62	15.13
	84,000	2	100		89.16	87.81	87.45	86.14	50.55	58.74	18.55
	126,000	3	150		92.27	93.48	89.43	90.00	52.19	66.18	19.13
	168,000	4	200		95.42	96.49	95.24	95.16	66.12	68.84	20.51
	210,000	5	250		95.88	95.41	95.13	95.89	72.72	75.54	22.56
	420,000	10	500		95.45	95.70	95.40	95.74	80.86	82.81	23.27
140	42,000	1	20	40x50	80.44	79.79	76.59	74.77	34.09	54.55	14.57
	84,000	2	40		91.56	90.13	89.91	88.25	56.93	59.12	17.55
	126,000	3	60		93.71	92.22	91.91	90.18	58.01	64.47	19.04
	168,000	4	80		95.45	94.48	95.40	91.26	62.02	72.40	20.31
	210,000	5	100		95.87	95.19	95.76	94.71	63.32	72.71	21.46
	420,000	10	250		95.81	95.87	95.19	95.49	81.41	90.98	22.10
140	84,000	2	20	50x80	74.16	69.71	68.86	64.11	42.71	35.71	15.47
	126,000	3	30		92.46	94.62	89.45	95.98	54.48	52.27	16.13
	168,000	4	40		94.74	95.81	91.68	95.90	59.86	70.72	20.09
	210,000	5	50		95.71	95.46	95.56	95.59	65.23	75.89	21.06
	420,000	10	100		95.55	95.87	95.78	95.81	84.65	90.14	21.17

Legend: Acc. Is the accelerometer, Aco. is acoustic.

2D-CNNs are used to obtain results by converting raw acceleration and acoustic data into grayscale images. Table 6 tabulates the number of fault harmonics, number of samples, data collection duration, number of images per class, and input size per image. It is observed that the resolution of grayscale images increases with larger window lengths. The results presented in Table 6 indicate that validation accuracies plateau around 95% for both accelerometer and acoustic sensor data when 140 fault harmonics, 168,000 samples, a data collection duration of 4 s, 200 images per class, and an input size of 20x50 when using either a 3% or 10% stride. Like Table 5, it can be concluded that increasing the stride adversely affects the accuracy of the network, resulting in a significant drop in accuracy.

Table 7 summarizes the values required to achieve the results found in Table 5 and Table 6 and extends them to include other fault types. The results indicate that after

including a certain number of samples, validation accuracy plateaus for all fault types when using 140 fault harmonics (highest fault frequency identified), 168,000 (1D-CNN), and 210,000 samples (2D-CNN), 4 and 5 s data collection durations, 0.25 and 0.20 frequency resolutions, a window length of 1024, 200 images, and a 10% stride. For the remainder of dataset testing, a window length of 1024 and a 10% stride will be used to identify appropriate threshold accuracies, as these values were determined to be consistently adequate with different fault types in each testing scenario.

4.2. Validation Accuracy Values for Sampling Rate and Fault Harmonics Using the CWRU Dataset

The CWRU dataset is analyzed to understand the effect of sampling rate and number of fault harmonic values on identifying threshold accuracies by using ML algorithms.

Table 7. UORED-VAFCLS Classification Accuracy Results Using a Sampling Rate of 42kHz and a Stride of 10%

Sensor Type		Accelerometer	Acoustic
Sampling Rate		42,000 Hz	42,000 Hz
# Fault Harmonics		140	140
Training Data Size		1024	1024
# of Images		200	200
Average Run Time ANN (minutes)		4.05	4.31
Average Run Time 1D-CNN (minutes)		4.47	4.45
Average Run Time 2D-CNN (minutes)		20.36	20.47
Bearing Faults using 1D-CNNs, 210,000 Samples, sample duration of 5 seconds, 0.20 frequency resolution	Inner Race	95.13 %	95.04 %
	Outer Race	97.48 %	96.88 %
	Cage	98.76 %	96.46 %
	Ball	96.87 %	97.79 %
Average Total		97.06 %	96.54 %
Bearing Faults using 2D-CNNs, 168,000 Samples, sample duration of 4 seconds, 0.25 frequency resolution	Inner Race	95.24 %	95.16 %
	Outer Race	98.18 %	93.55 %
	Cage	99.59 %	95.18 %
	Ball	97.43 %	91.65 %
Average Total		97.86 %	93.89 %

Table 8. CWRU Inner Race Dataset Results

Sampling Rate (Hz) [a], # of Fault Harmonics [b], number of samples [c], Data Collection Duration (s) [d], Frequency Resolution [e], Window Length [f], # of Images per Class [g], Input Size [h]								ML Models, BPFI, Validation Accuracy (%), Average Run Time (minutes)					
a	b	c	d	e	f	g	h	ANN Acc., 10% stride (%)	Time (min)	1D 2-layer CNN Acc., 10% stride (%)	Time (min)	2D-CNN Acc., 10% stride (%)	Time (min)
12,000	40	12,000	1	1.00	1024	10	20x50	70.08	0.57	73.86	2.24	76.89	7.22
		24,000	2	0.50		20		83.51	1.15	85.87	2.33	86.41	10.45
		36,000	3	0.33		30		91.38	1.35	90.71	2.58	93.44	13.33
		48,000	4	0.25		40		95.11	1.45	95.16	3.15	95.13	15.47
		60,000	5	0.20		50		95.16	2.21	95.06	4.31	95.16	16.13
		120,000	10	0.10		100		95.21	3.11	95.10	5.31	95.08	18.51
		48,000	1	1.00		50		77.50	1.59	79.56	3.42	88.45	17.51
48,000	160	96,000	2	0.50	1024	100	20x50	85.66	2.41	86.83	4.17	93.96	19.13
		144,000	3	0.33		150		93.45	3.23	90.26	4.23	94.92	21.11
		192,000	4	0.25		200		95.08	4.11	95.71	4.51	95.50	25.18
		236,000	5	0.20		250		95.18	5.21	95.56	6.17	95.51	26.35
		480,000	10	0.10		500		95.13	7.16	95.80	8.35	95.84	27.17
		48,000	1	1.00		50		77.50	1.59	79.56	3.42	88.45	17.51
		96,000	2	0.50		100		85.66	2.41	86.83	4.17	93.96	19.13

The CWRU dataset is assessed using ANNs, 1D-CNNs, and 2D-CNNs, and results are provided in Table 8. 2D-CNNs are used with grayscale images as inputs to train the network. Table 8 reveals that results plateau around 95% as well for the CWRU 48 kHz sampling rate dataset, when 160 fault harmonics, 192,000 samples, a data collection duration of 4 s, a 0.25 frequency resolution, a 1024 window length, 200 grayscale images per class, and an input image size of 20x50 are required. On the other hand, if a 12 kHz sampling rate is used, the dataset only needs 40 fault harmonics, 48,000 samples, a data collection duration of 4 s, 0.25 frequency resolution, 40 grayscale images per class, and 20x50 image input size to achieve similar accuracies.

Table 9 demonstrates that either a 12 kHz or 48 kHz sampling rate can be utilized when reducing the number of samples during data collection since, after a certain number of sample, the accuracy does not change. This suggests that fewer samples can be collected for bearing fault diagnosis. While a minimum sampling rate is required to capture enough fault information so that an ML algorithm can assess a specific fault, these results indicate that once this minimum threshold is identified, sampling rates have little

effect on achieving a high accuracy. In fact, other parameters can be adjusted to reduce the total amount of data needed to achieve adequate performance.

Finally, the CWRU dataset also shows that a minimum of 40 fault harmonics still attains an adequate accuracy when other parameters are adjusted accordingly. Therefore, further study would be required to better understand how ML algorithms utilize fault harmonics in assessing faults by reducing the sampling frequency further to find the lower limit.

4.3. Validation Accuracy Values for Data Collection Duration and Frequency Resolution Using the HUST Dataset

Table 10 illustrates that the HUST dataset yields unsatisfactory results for both the ANN and 1D-CNN models, especially when analyzing data from the bearing sets ranging from 6204 to 6208. However, when the datasets are processed using a 2D-CNN with grayscale images, a threshold accuracy of around 95% can be successfully achieved. This was accomplished using 170 fault harmonics, 204,800 samples, a data collection duration of 4 s, a 0.24

Table 9. CWRU Classification Accuracy Results of 2D-CNNs at Sampling Frequencies of 12 kHz and 48 kHz

Sensor Type		Accelerometer
Data Collection Duration (s)		4
Frequency Resolution		0.25
Window Length		1024
Stride		10 %
Average Run Time [12 kHz] (minutes)		15.34
Average Run Time [48 kHz] (minutes)		25.51
Bearing Faults (12 kHz, 40 fault harmonics, 48,000 samples, 40 images per class)	Inner Race	95.13 %
	Outer Race	95.36 %
	Ball	95.75 %
Average Total		95.41 %
Bearing Faults (48 kHz, 160 fault harmonics, 192,000 samples, 200 images per class)	Inner Race	95.50 %
	Outer Race	95.90 %
	Ball	95.46 %
Average Total		95.62 %

Table 10. HUST Inner Race Dataset Results

Bearing Name [a], # of Fault Harmonics [b], # of samples [c], Data Collection Duration (s) [d], Frequency Resolution [e], Window Length [f], # of Images per Class [g], Input Size [h]								ML Models, BPFI, Validation Accuracy (%), Average Run Time (minutes)					
a	b	c	d	e	f	g	h	ANN Acc., 10% stride (%)	Time (min)	1D 2-layer CNN Acc., 10% stride (%)	Time (min)	2D-CNN Acc., 10% stride (%)	Time (min)
6204	170	51,200	1	1.00	1024	50	20x50	64.24	2.21	70.13	3.55	81.46	18.44
		102,400	2	0.50		100		69.79	2.55	76.37	4.31	89.33	20.15
		153,600	3	0.33		150		72.93	3.57	77.18	4.36	92.45	22.34
		204,800	4	0.25		200		71.23	4.49	75.22	4.59	95.44	26.50
		256,000	5	0.20		250		71.76	5.55	77.79	5.73	95.32	27.19
6205	170	51,200	1	1.00	1024	50	20x50	63.54	2.26	75.45	3.48	80.88	18.40
		102,400	2	0.50		100		67.95	2.51	77.61	4.21	85.43	20.11
		153,600	3	0.33		150		73.45	3.51	80.47	4.41	92.19	22.21
		204,800	4	0.25		200		72.46	4.43	74.98	4.49	95.23	26.47
		256,000	5	0.20		250		71.43	6.31	75.11	6.35	95.01	27.26
6206	170	51,200	1	1.00	1024	50	20x50	61.77	2.16	78.55	3.40	79.46	18.38
		102,400	2	0.50		100		66.13	2.43	79.43	4.26	83.81	20.29
		153,600	3	0.33		150		70.64	3.51	76.15	4.47	91.47	22.47
		204,800	4	0.25		200		71.55	4.41	77.69	4.40	95.28	26.31
		256,000	5	0.20		250		76.49	6.13	74.11	6.31	95.10	27.50
6207	170	51,200	1	1.00	1024	50	20x50	59.99	2.16	75.49	3.57	84.61	18.22
		102,400	2	0.50		100		65.56	2.51	77.13	4.51	86.13	20.18
		153,600	3	0.33		150		64.83	3.50	79.84	4.58	90.44	22.36
		204,800	4	0.25		200		65.95	4.41	74.21	5.21	95.06	26.44
		256,000	5	0.20		250		70.51	6.27	75.87	7.42	95.15	27.31
6208	170	51,200	1	1.00	1024	50	20x50	60.45	2.30	77.41	3.23	85.69	18.31
		102,400	2	0.50		100		62.88	2.45	75.89	3.30	90.12	20.26
		153,600	3	0.33		150		65.41	3.42	73.86	3.45	93.88	22.33
		204,800	4	0.25		200		68.87	4.33	74.77	4.01	95.17	26.58
		256,000	5	0.20		250		72.59	7.01	77.67	6.07	95.84	27.07

frequency resolution, a 1024 window length, 200 grayscale images per class, and a 20x50 image input size. Bearing 6204 is selected for further analysis using a 2D-CNN as bearings in the range of 6204 to 6208 all had similar results.

Table 11 demonstrates that the final accuracy climbs to 95.65% when considering different fault conditions with the threshold accuracy parameters obtained from Table 10. The results indicate that to reach an adequate accuracy, the data collection duration for a 2D-CNN should be a minimum of 4 s and a frequency resolution of at least 0.25.

Table 11. HUST Classification Accuracy Results at a Sampling Rate of 51.2 kHz for the 6204 Bearing

Sensor Type		Accelerometer
Sampling Rate		51,200
# of Fault Harmonics		170
# of Samples		204,800
Data Collection Duration (s)		4
Window Length		1024
# of Images Per Class		200
Input Size		20x50
Stride		10 %
Average Run Time (minutes)		26.46
Bearing Faults	Inner Race	95.44 %
	Outer Race	95.75 %
	Ball	95.77 %
Average Total		95.65 %

5. DISCUSSIONS

Figures 5, 6, and 7 consist of validation accuracy vs computational times for the ANN, 1D-CNN, and 2D-CNN, respectively. In general, as the number of samples used as an input to each algorithm increases, so does computation time and accuracy, up to a plateau value. In Figure 5, the UORED-VAFCLS and CWRU datasets approach plateau accuracies of approximately 95% in less than 5 minutes. However, the HUST dataset does not meet this accuracy, suggesting that the architecture is not able to diagnose noisy

data, which is essential for transitioning into industry.

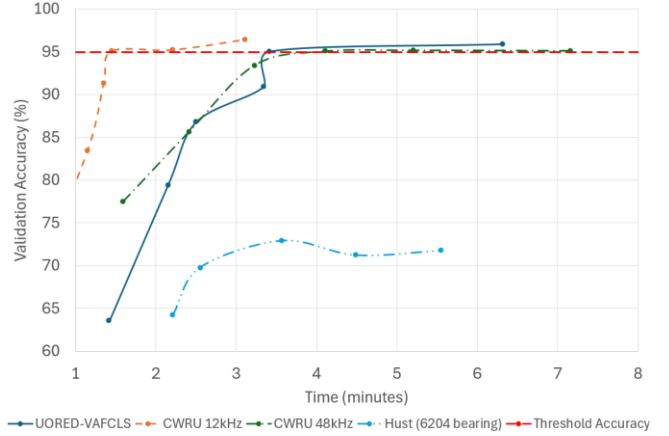
**Figure 5.** ANN Validation Accuracy vs Computational Times for the Datasets Considered

Figure 6 displays a similar pattern, where the UORED-VAFCLS and CWRU datasets both reach plateau accuracies of approximately 95% in less than 5 minutes. The HUST dataset, again, lags behind due to the dataset being noisy. This causes the validation accuracy to fluctuate, indicating that noisy data cannot be used to achieve a satisfactory accuracy with simple ML algorithms when using time domain data.

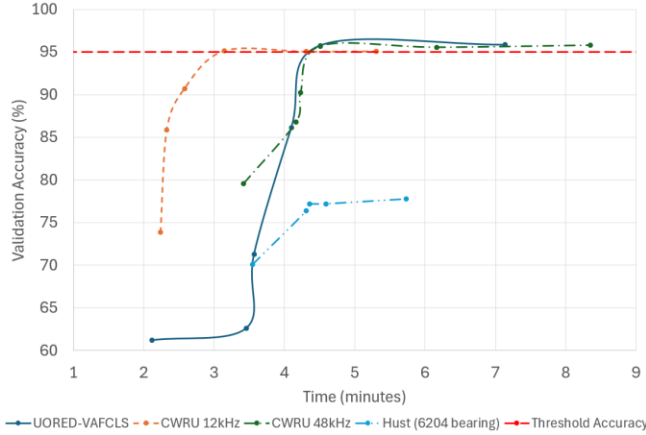


Figure 6. 1D-CNN Validation Accuracy vs Computational Times for the Datasets Considered

Finally, in Figure 7, all datasets, including the noisy HUST dataset, surpass the threshold accuracy. The 2D-CNN architecture demonstrates its robustness in handling noisy data, potentially making it more suitable for industrial datasets, which are often noisy. This observation points to the potential of integrating 2D-CNNs into real world industrial applications for fault diagnosis or machine monitoring, where data noise is a common issue. To overcome the noise issue in industrial datasets, the frequency domain should be considered, as demonstrated by many studies (Feng et al., 2013; Y. Kim & Kim, 2023; Q. Wang et al., 2009).

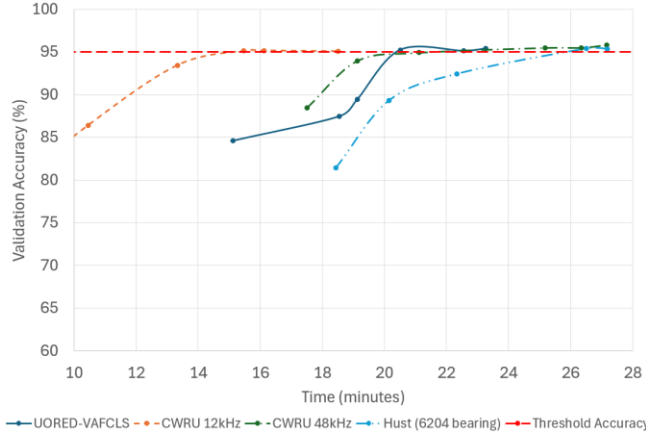


Figure 7. 2D-CNN Validation Accuracy vs Computational Times for the Datasets Considered

With all simple ML algorithms considered herein, validation accuracy seems to plateau around 95%, justifying this value as the threshold accuracy that should be achieved for satisfactory fault diagnosis results. In fact, when considering validation accuracy versus computation time, 95% provides the best value for IIoT systems. Therefore, when testing new combinations of parameters or datasets, a minimum 95% threshold accuracy can be used as a satisfactory yet reasonable benchmark. With the development of more complex and more efficient algorithms, it is expected that the threshold accuracy for satisfactory results should increase.

5.1.1. Accelerometer Data

To achieve an adequate validation accuracy, all three datasets used in this study (i.e., UORED-VAFCLS, CWRU, and HUST) required similar sample collection durations, regardless of sampling rate differences. For all three datasets 1, 2, 3, 4, 5, and 10s durations were tested. Nonetheless, the minimum data collection duration for TL and 1D-CNN rolling element bearing diagnosis was found to be 5 s, as seen in Table 12 for clean data, while for 2D-CNNs, 4 s was found to be sufficient for both clean and noisy data. It was not possible to achieve a satisfactory accuracy for the HUST dataset when using an ANN or a 1D-CNN. This may be due to a potential lack of data quality, as mentioned by the researchers who published the HUST dataset themselves, which is an indication of potential poor data collection (Thuan & Hong, 2023b). However, data collection duration and frequency resolution were found to be sufficient when using constant values of 4 s and 0.25, respectively while using a 2D-CNN.

The average run times in Table 12 show that 2D-CNN models are generally the slowest in processing data across all datasets despite achieving high accuracy. For example, when applied to the UORED-VAFCLS dataset, 2D-CNNs have an average run time of 20.36 minutes, while 1D-CNNs take 4.47 minutes, and ANNs are the fastest at 4.05 minutes. A similar trend is seen in the CWRU and HUST datasets, where the 2D-CNN takes the longest time to process (up to 26.46 minutes with HUST), while 1D-CNNs are moderately faster, and ANNs consistently process the data the fastest, as expected.

Table 12. Accuracy Sample Duration Based on Accelerometer Sampling Rate Needed to Achieve a Network Accuracy Greater than 95%

Dataset Name	Sampling Rate	# of Fault Harmonics	Number of Samples	Sample Duration	Frequency Resolution	Window Length	# of Images per Class	Input Size	Stride (%)	Average Run Time (minutes)	Network
UORED-VAFCLS	42 kHz	140	210,000	5 s	0.20	1024	N/A	N/A	10	4.05	ANN
CWRU	12 kHz	40	48,000	4 s	0.25					1.45	
CWRU	48 kHz	160	192,000	4 s	0.25					4.11	
HUST	51.2 kHz	N/A	N/A	N/A	N/A					4.41	
UORED-VAFCLS	42 kHz	140	210,000	5 s	0.20	1024	N/A	N/A	10	4.47	1D-CNN
CWRU	12 kHz	40	48,000	4 s	0.25					3.15	
CWRU	48 kHz	160	192,000	4 s	0.25					4.51	
HUST	51.2 kHz	N/A	N/A	N/A	N/A					4.46	
UORED-VAFCLS	42 kHz	140	168,000	4 s	0.25	1024	200	20x50	10	20.36	2D-CNN
CWRU	12 kHz	40	48,000	4 s	0.25		40			15.34	
CWRU	48 kHz	160	192,000	4 s	0.25		200			25.51	
HUST	51.2 kHz	170	204,800	4 s	0.25		200			26.46	

This performance discrepancy is crucial for IIoT researchers and companies that prioritize efficiency in both accuracy and runtime, especially in real-time applications. Although 2D-CNNs achieve the highest accuracy across all datasets, their higher computational cost might not be ideal for cases requiring quicker responses, as evidenced by the 25.51 minute runtime for the CWRU dataset using the 2D-CNN.

Figure 6 shows that high accuracies (greater than 95%) can still be obtained by increasing the number of samples, but for applications where runtime is critical, ANNs and 1D-CNNs may be preferred. For instance, the HUST dataset processes much faster using an ANN (4.41 minutes) compared to a 2D-CNN (26.46 minutes) but does not achieve satisfactory results, thus making 2D-CNNs the only viable solution for both accuracy and computational time (even though it is slightly longer).

Table 12 includes a summary of best performing 2D-CNN networks (above 95% accuracies). This is found to occur when the number of samples is four times larger than the sampling rate. Interestingly, a relationship related to resolution, equation (5), can be observed when looking at all the results in this study. Evidently, as the number of samples collected increases, so does the resolution. Then, the data collection duration can be used to determine how much data is required to achieve an accuracy over 95%. For example, if a data scientist selects an accelerometer that operates at 15 kHz (50 fault harmonics) to collect data, then to reach a network accuracy of at least 95% when using a 2D-CNN with grayscale images, 60,000 samples collected over 4 s would be required. Additionally, ANNs are found to be less reliable than CNNs if the data collection process is poor, as shown with the HUST dataset when compared to cleaner datasets (UORED-VAFCLS and CWRU).

Thus, while ANNs and 1D-CNNs provide faster runtimes in terms of computation, they do not reach a satisfactory threshold accuracy for all datasets. Therefore,

IIoT companies should consider 2D-CNNs by sacrificing some computational runtime for achieving validation accuracies above 95% for proper condition monitoring, even under noisy environments.

The average run times shown in Table 12 were obtained using an Intel® Core™ i7-1255U processor, up to 16 GB of RAM, 1 TB of storage, Intel® UHD Graphics, and Jupyter notebook. Additionally, average run times consist of how long it takes to run 100 epochs.

5.1.2. Acoustic Data

Results also suggest that for TL and DL algorithms, the data sampling duration should be 4 s (2D-CNNs) and 5 s (ANNs, 1D-CNNs) for microphone data, as observed in the UORED-VAFCLS dataset. Unfortunately, the CWRU and HUST datasets did not include acoustic data, making conclusions for this type of data difficult. However, a recommendation is made based on the UORED-VAFCLS data results. This involves collecting samples for a duration of 5 s to obtain a 2D-CNN network accuracy of more than 95% when using a microphone. Notably, acoustic sensors pick up more noise during signal collection than accelerometers, which could explain why a longer data collection duration is needed. More datasets need to be collected for acoustic data to provide greater insight using traditional and DL.

5.1.3. Summary of Results and Bearing Data Collection Recommendations

Based on Table 13's summary of the results, the minimum required resolution for ML diagnosis can be calculated using equation (5) for each dataset, using all three algorithms used in this study. For ANNs and 1D-CNNs, a resolution of 0.20 (5 seconds of data collection) is needed, while for 2D-CNNs, a lower resolution of 0.25 (4 seconds of data collection) is required when using simple ML

Table 13. Resolution (Hz) for Each Dataset

Dataset Name	Algorithm	Sampling Rate	Number of Samples Required	Sample Duration Required	Resolution
UORED-VAFCLS	ANN	42 kHz	210,000	5 s	0.20
CWRU		12 kHz	48,000	4 s	0.25
CWRU		48 kHz	192,000	4 s	0.25
HUST		51.2 kHz	N/A	N/A	N/A
UORED-VAFCLS	1D-CNN	42 kHz	210,000	5 s	0.20
CWRU		12 kHz	48,000	4 s	0.25
CWRU		48 kHz	192,000	4 s	0.25
HUST		51.2 kHz	N/A	N/A	N/A
UORED-VAFCLS	2D-CNN	42 kHz	168,000	4 s	0.25
CWRU		12 kHz	48,000	4 s	0.25
CWRU		48 kHz	192,000	4 s	0.25
HUST		51.2 kHz	204,800	4 s	0.25

algorithms. It is expected that more complex algorithms could require parameters with lesser values to attain a satisfactory accuracy after a certain number of samples.

To ensure optimal ML performance in rolling element bearing fault diagnosis, this study explored data collection parameters that balance model accuracy with the amount of bearing data used. Only 2D-CNNs were able to achieve a satisfactory threshold accuracy for all three datasets. For this case, results indicate that a minimum sampling rate of 12 kHz and a data collection duration of 4 s are effective in capturing bearing conditions, allowing 2D-CNNs to reach satisfactory accuracies while maintaining relatively low computational times. When using 2D-CNN grayscale images in the frequency domain, all three datasets analyzed herein can achieve a satisfactory threshold accuracy at the expense of increased computational times. The number of harmonics, which is defined using bearing signal theory, is a critical factor for diagnosis. Based on the results, 40 fault signatures are enough to obtain a satisfactory accuracy. Additionally, the suggested minimum image resolution is 20x50 pixels, and a stride of 10% provides better accuracies for fault diagnosis without excessive computational times, while an input signal window length of 1024 is satisfactory in all cases. The provided configurations support an efficient yet accurate set of parameters for IIoT applications where data storage and processing constraints are considered. These guidelines provide a methodological approach to data collection, which will help support accurate fault diagnosis in industrial settings. This will aid in developing efficient ML-based monitoring systems for IIoT companies.

Both the CWRU and HUST datasets contain data for three bearings of each fault type, including inner race, outer race, and ball faults. On the other hand, the UORED-VAFCLS dataset has data for five bearings with natural faults of each fault type, including inner race, outer race, ball, and cage faults. To enhance the robustness of the dataset, multiple sets of 5 of the same fault type but on different bearings of the same size and from the same manufacturer are recommended. This will facilitate effective training of ML algorithms by allowing for larger datasets rather than high-dimensional datasets. Lower dimensions and larger datasets are essential for transitioning research in this field into industry. Moreover, this study demonstrates that frequency-domain data can be used with simple ML algorithms to determine rolling element bearing faults with 95% accuracy. Unfortunately, most IIoT products currently available in the market pre-process their data due to storage and communication limitations of their devices. As such, these devices would need to be able to send time series data directly if ML algorithms are to be integrated with IIoT products.

5.1.4. Limitations

It is important to note that due to the limited number of datasets available for this study, conclusions made herein

can be said to only apply to accelerometer sampling frequencies between 12 kHz and 51.2 kHz and speeds ranging between 1400 and 1900 RPM. While such ranges are common in practice, the extent to which these assumptions are restrictive should be further studied. Additionally, for 2D-CNNs using grayscale images, a window length of 1024 seems to yield the minimum required accuracy sought in this study. Although the paper provides general conclusions, more structured evidence is needed to support their applicability to other fault types and broader operating conditions.

It is important to note that hyperparameters such as window length, stride, and data duration were varied independently in this study to observe their isolated effects on model accuracy. However, potential interactions between these parameters (e.g., stride and window length affecting the number and overlap of samples) may impact performance.

6. CONCLUSIONS

The results indicate that a minimum number of samples is required to achieve a training and validation accuracy above 95%. This number varies depending on the sampling rate provided by each dataset.

When selecting the data collection duration for TL and DL, it is recommended that data be collected for a duration of 5s (resolution of 0.20) for ANNs and 1D-CNNs for clean data, and 4s (resolution of 0.25) for 2D-CNNs for clean and noisy data. The 2D-CNN finding is applicable for accelerometers that have a sampling rate range between 12 and 51.2 kHz and speeds ranging between 1400 and 1900 RPM.

Once the minimum threshold of samples is reached, it was observed that sampling rates have little effect on achieving a high classification accuracy. Additionally, for 2D-CNNs, the best performance was found to occur when the number of samples was approximately four times the sampling rate, establishing a practical relationship between these two key data collection parameters.

Finally, a 10% stride was found to provide sufficient information for achieving an ML accuracy of at least 95%. To increase the robustness of these results, further testing on more bearing datasets will be required at different constant speeds to see the effects on resolution.

The research succeeds in delivering general conclusions that can be directly applied across a wide variety of bearing conditions. The findings presented here are based on research lab-based bearing datasets, with sampling rates and machine speeds confined to specific ranges. As a result, the conclusions drawn do not fully capture the variability of all machine applications where accelerometers may operate outside the ranges studied.

Moreover, the study provides useful insights into how different ML models (ANNs, 1D-CNNs, and 2D-CNNs) perform under specific conditions. It has been shown that

for ball bearing fault diagnosis within the speed ranges and frequencies considered in this study, even with noisy data, the threshold accuracy can be reached using 2D-CNN grayscale images in the frequency domain by sacrificing computational time. The identification of a threshold accuracy will help data scientists and ML researchers to conduct research more efficiently and will allow the implementation of ML algorithms in IIoT systems by maximizing accuracy while reducing samples and computational times.

These results provide researchers a basis for setting their parameters during data collection but also serve as a first step for integrating ML algorithm-based rolling element bearing fault diagnosis into IIoT monitoring products by identifying the minimum amount of data needed to reach threshold fault diagnosis accuracies.

Future work will focus on exploring the interaction between hyperparameters such as window length, stride, and duration, rather than treating them independently. Additionally, future plans include examining statistical robustness by reporting mean performance metrics and standard deviations across repeated runs and investigating the impact of cross-validation versus domain-split validation frameworks. Expanding the scope to include variable speed conditions, a wider range of machine types, and more diverse sensor configurations will further enhance the applicability of these findings in real-world industrial settings.

NOMENCLATURE

<i>AI</i>	Artificial Intelligence
<i>ANN</i>	Artificial Neural Networks
<i>BPFO</i>	Ball Pass Frequency of Outer Race
<i>BPFI</i>	Ball Pass Frequency of Inner Race
<i>BSF</i>	Ball Spin Frequency
<i>CNN</i>	Convolutional Neural Networks
<i>DL</i>	Deep Learning
<i>FTF</i>	Fundamental Train Frequency
<i>IIoT</i>	Industrial Internet of Things
<i>IMS</i>	Intelligent Maintenance System
<i>KNN</i>	K-Nearest Neighbors Algorithm
<i>ML</i>	Machine Learning
<i>NSK</i>	Nippon Seiko Kabushiki-gaisha
<i>SVM</i>	Support Vector Machine
<i>SKF</i>	Svenska Kullagerfabriken
<i>TL</i>	Traditional Learning
<i>t-SNE</i>	T- distributed Stochastic Neighbor Embedding

7. REFERENCES

- 708/1250 *AMB - Angular contact ball bearings* | SKF. (n.d.). Retrieved June 1, 2023, from <https://www.skf.com/us/products/rolling-bearings/ball-bearings/angular-contact-ball-bearings/single-row-angular-contact-ball-bearings/productid-708%2F1250%20AMB>
- 6000—*Deep groove ball bearings* | SKF. (n.d.). Retrieved June 1, 2023, from <https://www.skf.com/in/products/rolling-bearings/ball-bearings/deep-groove-ball-bearings/productid-6000>
- Althobiani, F. (2024). A Novel Framework for Robust Bearing Fault Diagnosis: Preprocessing, Model Selection, and Performance Evaluation. *IEEE Access*, 12, 59018–59036. IEEE Access. <https://doi.org/10.1109/ACCESS.2024.3390234>
- Apparatus & Procedures* | Case School of Engineering | Case Western Reserve University. (2021, August 10). Case School of Engineering. <https://engineering.case.edu/bearingdatacenter/apparatus-and-procedures>
- Arora, S., Hu, W., & Kothari, P. K. (2018). *An Analysis of the t-SNE Algorithm for Data Visualization* (No. arXiv:1803.01768). arXiv. <https://doi.org/10.48550/arXiv.1803.01768>
- Asutkar, S., & Tallur, S. (2023). Deep transfer learning strategy for efficient domain generalisation in machine fault diagnosis. *Scientific Reports*, 13(1), Article 1. <https://doi.org/10.1038/s41598-023-33887-5>
- BALL BEARING BASICS AND TYPES. (n.d.). *ENGINEERING APPLICATIONS*. Retrieved August 16, 2023, from <https://www.hkdivedi.com/2016/02/ball-bearing-basics-and-types.html>
- Bloch, H. P., & Geitner, F. K. (2012). *Machinery Failure Analysis and Troubleshooting: Practical Machinery Management for Process Plants*. Butterworth-Heinemann.
- Chandrvanshi, S., Sharma, S., Singh, M. P., & Singh, R. (2024). Bearing Fault Diagnosis Using Machine Learning Models. In D. K. Sharma, S.-L. Peng, R. Sharma, & G. Jeon (Eds.), *Micro-Electronics and Telecommunication Engineering* (pp. 219–233). Springer Nature. https://doi.org/10.1007/978-981-99-9562-2_18
- Che, C., Wang, H., Fu, Q., & Ni, X. (2021). Intelligent fault prediction of rolling bearing based on gate recurrent unit and hybrid autoencoder. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 235(6), 1106–1114. <https://doi.org/10.1177/0954406220941037>
- Chen, C.-C., Liu, Z., Yang, G., Wu, C.-C., & Ye, Q. (2021). An Improved Fault Diagnosis Using 1D-Convolutional Neural Network Model. *Electronics*, 10(1), Article 1. <https://doi.org/10.3390/electronics10010059>
- Chen, Z., Zhu, L., Lu, H., Chen, S., Zhu, F., Liu, S., Han, Y., & Xiong, G. (2024). Research on bearing fault diagnosis based on improved genetic algorithm and BP neural network. *Scientific Reports*, 14(1), 15527. <https://doi.org/10.1038/s41598-024-66318-0>
- Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2016). *How much data is needed to train a medical image deep*

- learning system to achieve necessary high accuracy? (No. arXiv:1511.06348). arXiv. <https://doi.org/10.48550/arXiv.1511.06348>
- Chuya, J., Alonso-Valerdi, L., & Ibarra-Zarate, D. I. (2022). Deep-Learning Method Based on 1D Convolutional Neural Network for Intelligent Fault Diagnosis of Rotating Machines. *Applied Sciences*, 12, 2158. <https://doi.org/10.3390/app12042158>
- De la Fuente, C., Martinez-Valdes, E., Priego-Quesada, J. I., Weinstein, A., Valencia, O., Kunzler, M. R., Alvarez-Ruf, J., & Carpes, F. P. (2021). Understanding the effect of window length and overlap for assessing sEMG in dynamic fatiguing contractions: A non-linear dimensionality reduction and clustering. *Journal of Biomechanics*, 125, 110598. <https://doi.org/10.1016/j.jbiomech.2021.110598>
- Deep groove ball bearings | SKF. (n.d.). Retrieved August 16, 2023, from <https://www.skf.com/au/products/rolling-bearings/ball-bearings/deep-groove-ball-bearings>
- Delaunay, J. J., Rommel, M., & Geisler, J. (1994). The importance of the sampling frequency in determining short-time-averaged irradiance and illuminance for rapidly changing cloud cover. *Solar Energy*, 52(6), 541–545. [https://doi.org/10.1016/0038-092X\(94\)90662-9](https://doi.org/10.1016/0038-092X(94)90662-9)
- Dini, P., Diana, L., Elhanashi, A., & Saponara, S. (2024). Overview of AI-Models and Tools in Embedded IIoT Applications. *Electronics*, 13(12), Article 12. <https://doi.org/10.3390/electronics13122322>
- Djaballah, S., Saidi, L., Meftah, K., Hechifa, A., Bajaj, M., & Zaitsev, I. (2024). A hybrid LSTM random forest model with grey wolf optimization for enhanced detection of multiple bearing faults. *Scientific Reports*, 14(1), 23997. <https://doi.org/10.1038/s41598-024-75174-x>
- Duan, Y., Wang, C., Chen, Y., & Liu, P. (2019). Improving the Accuracy of Fault Frequency by Means of Local Mean Decomposition and Ratio Correction Method for Rolling Bearing Failure. *Applied Sciences*. <https://doi.org/10.3390/app9091888>
- Duo, W., Zhang, M., Xu, Y., Lu, W., Yang, J., & Zhang, T. (2021). Metric-based meta-learning model for few-shot fault diagnosis under multiple limited data conditions. *Mechanical Systems and Signal Processing*, 155. <https://doi.org/10.1016/j.ymssp.2020.107510>
- Duque-Perez, O., Del Pozo-Gallego, C., Morinigo-Sotelo, D., & Fontes Godoy, W. (2019). Condition Monitoring of Bearing Faults Using the Stator Current and Shrinkage Methods. *Energies*, 12(17), Article 17. <https://doi.org/10.3390/en12173392>
- Feng, Z., Liang, M., & Chu, F. (2013). Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples. *Mechanical Systems and Signal Processing*, 38(1), 165–205. <https://doi.org/10.1016/j.ymssp.2013.01.017>
- Gao, S., He, J., Pan, H., & Gong, T. (2022). A Multi-Scale and Lightweight Bearing Fault Diagnosis Model with Small Samples. *Symmetry*, 14(5), Article 5. <https://doi.org/10.3390/sym14050909>
- Gao, Z., Cecati, C., & Ding, S. X. (2015). A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches. *IEEE Transactions on Industrial Electronics*, 62(6), 3757–3767. <https://doi.org/10.1109/TIE.2015.2417501>
- Gousseau, W., Girardin, F., & Griffaton, J. (2016). Analysis of the Rolling Element Bearing data set of the Center for Intelligent Maintenance Systems of the University of Cincinnati.
- Gu, J., Peng, Y., Lu, H., Chang, X., & Chen, G. (2022). A novel fault diagnosis method of rotating machinery via VMD, CWT and improved CNN. *Measurement*, 200, 111635. <https://doi.org/10.1016/j.measurement.2022.111635>
- Han, S., & Jeong, J. (2020). An Weighted CNN Ensemble Model with Small Amount of Data for Bearing Fault Diagnosis. *Procedia Computer Science*, 175, 88–95. <https://doi.org/10.1016/j.procs.2020.07.015>
- He, J., Lv, Z., & Chen, X. (2023). Rolling bearing fault diagnosis method based on 2D grayscale images and Wasserstein Generative Adversarial Nets under unbalanced sample condition. *Complex Engineering Systems*, 3. <https://doi.org/10.20517/ces.2023.20>
- He, J., Wu, P., Tong, Y., Zhang, X., Lei, M., & Gao, J. (2021). Bearing Fault Diagnosis via Improved One-Dimensional Multi-Scale Dilated CNN. *Sensors (Basel, Switzerland)*, 21(21), 7319. <https://doi.org/10.3390/s21217319>
- Hendriks, J., Dumond, P., & Knox, D. A. (2022). Towards better benchmarking using the CWRU bearing fault dataset. *Mechanical Systems and Signal Processing*, 169, 108732. <https://doi.org/10.1016/j.ymssp.2021.108732>
- Hoang, D.-T., & Kang, H.-J. (2019). Rolling element bearing fault diagnosis using convolutional neural network and vibration image. *Cognitive Systems Research*, 53, 42–50. <https://doi.org/10.1016/j.cogsys.2018.03.002>
- Hou, P., Zhang, J., Jiang, Z., Tang, Y., & Lin, Y. (2023). A Bearing Fault Diagnosis Method Based on Dilated Convolution and Multi-Head Self-Attention Mechanism. *Applied Sciences*, 13(23), Article 23. <https://doi.org/10.3390/app132312770>
- Isermann, R. (2006). *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer Science & Business Media.
- Jiang, W., Xu, Y., Chen, Z., Zhang, N., & Zhou, J. (2022). Fault diagnosis for rolling bearing using a hybrid

- hierarchical method based on scale-variable dispersion entropy and parametric t-SNE algorithm. *Measurement*, 191, 110843. <https://doi.org/10.1016/j.measurement.2022.110843>
- Karabay, S., & Uzman, I. (2009). Importance of early detection of maintenance problems in rotating machines in management of plants: Case studies from wire and tyre plants. *Engineering Failure Analysis*, 16(1), 212–224. <https://doi.org/10.1016/j.engfailanal.2008.03.003>
- Kester, W. (2023). *MT-002 TUTORIAL What the Nyquist Criterion Means to Your Sampled Data System Design*.
- Kim, T., & Chai, J. (2021). Pre-Processing Method to Improve Cross-Domain Fault Diagnosis for Bearing. *Sensors*, 21(15), Article 15. <https://doi.org/10.3390/s21154970>
- Kim, Y., & Kim, Y.-K. (2023). Time-Frequency Multi-Domain 1D Convolutional Neural Network with Channel-Spatial Attention for Noise-Robust Bearing Fault Diagnosis. *Sensors*, 23(23), Article 23. <https://doi.org/10.3390/s23239311>
- Konstruktions- und Antriebstechnik (KAt)—Data Sets and Download (Universität Paderborn). (n.d.). Retrieved June 15, 2023, from <https://mb.uni-paderborn.de/kat/forschung/kat-datacenter/bearing-datacenter/data-sets-and-download>
- Kumar, D., Ujjan, S. M., Dev, K., Khawaja, S. A., Bhatti, N. A., & Hussain, T. (2022). Towards soft real-time fault diagnosis for edge devices in industrial IoT using deep domain adaptation training strategy. *Journal of Parallel and Distributed Computing*, 160, 90–99. <https://doi.org/10.1016/j.jpdc.2021.10.005>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), Article 7553. <https://doi.org/10.1038/nature14539>
- Lu, C., Wang, Z.-Y., Qin, W.-L., & Ma, J. (2017). Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. *Signal Processing*, 130, 377–388. <https://doi.org/10.1016/j.sigpro.2016.07.028>
- Lu, F., Tong, Q., Xu, J., Feng, Z., Wang, X., Huo, J., & Wan, Q. (2024). Towards multi-scene learning: A novel cross-domain adaptation model based on sparse filter for traction motor bearing fault diagnosis in high-speed EMU. *Advanced Engineering Informatics*, 60, 102536. <https://doi.org/10.1016/j.aei.2024.102536>
- Lu, H., Pavan Nemani, V., Barzegar, V., Allen, C., Hu, C., Laflamme, S., Sarkar, S., & Zimmerman, A. T. (2023). A physics-informed feature weighting method for bearing fault diagnostics. *Mechanical Systems and Signal Processing*, 191, 110171. <https://doi.org/10.1016/j.ymssp.2023.110171>
- Mahesh, T. R., Chandrasekaran, S., Ram, V. A., Kumar, V. V., Vivek, V., & Guluwadi, S. (2024). Data-Driven Intelligent Condition Adaptation of Feature Extraction for Bearing Fault Detection Using Deep Responsible Active Learning. *IEEE Access*, 12, 45381–45397. IEEE <https://doi.org/10.1109/ACCESS.2024.3380438>
- Mazhar, S. A. (2021). Methods of Data Collection: A Fundamental Tool of Research. *Journal of Integrated Community Health*, 10, 6–10. <https://doi.org/10.24321/2319.9113.202101>
- Muñoz-Terol, R., Reina Reina, A., Ziaei, S., & Gil, D. (2020). A Machine Learning Approach to Reduce Dimensional Space in Large Datasets. *IEEE Access*, 8, 1–1. <https://doi.org/10.1109/ACCESS.2020.3012836>
- Nagel, H. (2018, May 7). *Motor Speeds Explained: Diving into AC and DC Motors | Groschopp Blog*. Groschopp. <https://www.groschopp.com/motor-speeds-explained-ac-dc-motors/>
- Nasir, V., & Sassani, F. (2021). A review on deep learning in machining and tool monitoring: Methods, opportunities, and challenges. *The International Journal of Advanced Manufacturing Technology*, 115(9), 2683–2709. <https://doi.org/10.1007/s00170-021-07325-7>
- Neupane, D., & Seok, J. (2020). Bearing Fault Detection and Diagnosis Using Case Western Reserve University Dataset With Deep Learning Approaches: A Review. *IEEE Access*, 8, 93155–93178. IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2990528>
- Nyquist, H. (1924). Certain Factors Affecting Telegraph Speed. *Transactions of the American Institute of Electrical Engineers*, XLIII, 412–422. Transactions of the American Institute of Electrical Engineers. <https://doi.org/10.1109/T-AIEE.1924.5060996>
- Pham, M. T., Kim, J.-M., & Kim, C. H. (2020). Deep Learning-Based Bearing Fault Diagnosis Method for Embedded Systems. *Sensors (Basel, Switzerland)*, 20(23), 6886. <https://doi.org/10.3390/s20236886>
- Rahman, Md. S., Ghosh, T., Aurna, N. F., Kaiser, M. S., Anannya, M., & Hosen, A. S. M. S. (2023). Machine learning and internet of things in industry 4.0: A review. *Measurement: Sensors*, 28, 100822. <https://doi.org/10.1016/j.measen.2023.100822>
- Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing*, 25(2), 485–520. <https://doi.org/10.1016/j.ymssp.2010.07.017>
- Ren, J. (2012). ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging. *Knowledge-Based Systems*, 26, 144–153. <https://doi.org/10.1016/j.knosys.2011.07.016>
- Ruan, D., Wang, J., Yan, J., & Gühmann, C. (2023). CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis. *Advanced Engineering Informatics*, 55, 101877. <https://doi.org/10.1016/j.aei.2023.101877>

- Sacerdoti, D., Strozzi, M., & Secchi, C. (2023). A Comparison of Signal Analysis Techniques for the Diagnostics of the IMS Rolling Element Bearing Dataset. *Applied Sciences*, 13, 5977. <https://doi.org/10.3390/app13105977>
- Samanta, B., & Al-balushi, K. R. (2003). ARTIFICIAL NEURAL NETWORK BASED FAULT DIAGNOSTICS OF ROLLING ELEMENT BEARINGS USING TIME-DOMAIN FEATURES. *Mechanical Systems and Signal Processing*, 17(2), 317–328. <https://doi.org/10.1006/mssp.2001.1462>
- Sehri, M., & Dumond, P. (2023). *University of Ottawa Rolling-element Dataset – Vibration and Acoustic Faults under Constant Load and Speed conditions (UORED-VAFCLS)*. 5. <https://doi.org/10.17632/y2px5tg92h.5>
- Sehri, M., Dumond, P., & Bouchard, M. (2023a). University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets. *Data in Brief*, 49, 109327. <https://doi.org/10.1016/j.dib.2023.109327>
- Sehri, M., Dumond, P., & Bouchard, M. (2023b, November 21). *Design of a Data Fusion and Deep Transfer Learning Test Rig for Roller Bearings Diagnosis and Prognosis*. ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. <https://doi.org/10.1115/DETC2023-117264>
- Sehri, M., Hua, Z., Boldt, F. de A., & Dumond, P. (2025). Selective embedding for deep learning. *Knowledge-Based Systems*, 330, 114535. <https://doi.org/10.1016/j.knosys.2025.114535>
- Sehri, M., Varejão, I., Hua, Z., Bonella, V., Santos, A., Boldt, F. de A., Dumond, P., & Varejão, F. M. (2025). Towards a Universal Vibration Analysis Dataset: A Framework for Transfer Learning in Predictive Maintenance and Structural Health Monitoring. *International Journal of Prognostics and Health Management*, 16(3), Article 3. <https://doi.org/10.36001/ijphm.2025.v16i3.4239>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. The Bell System Technical Journal. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shao, H., Jiang, H., Lin, Y., & Li, X. (2018). A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders. *Mechanical Systems and Signal Processing*, 102, 278–297. <https://doi.org/10.1016/j.ymssp.2017.09.026>
- She, Y., Ai, M., Li, P., Wu, J., & Ma, J. (2024). Multiscale permutation entropy gray image coding method and its application in bearing fault diagnosis. *Engineering Research Express*, 6(1), 015092. <https://doi.org/10.1088/2631-8695/ad23c7>
- Skf Ball Bearing. (n.d.). Indiamart.Com. Retrieved August 16, 2023, from <https://www.indiamart.com/proddetail/skf-ball-bearing-2850651442855.html>
- Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, 64–65, 100–131. <https://doi.org/10.1016/j.ymssp.2015.04.021>
- Soomro, A. A., Muhammad, M. B., Mokhtar, A. A., Md Saad, M. H., Lashari, N., Hussain, M., Sarwar, U., & Palli, A. S. (2024). Insights into modern machine learning approaches for bearing fault classification: A systematic literature review. *Results in Engineering*, 23, 102700. <https://doi.org/10.1016/j.rineng.2024.102700>
- Sun, T., & Gao, J. (2024). New Fault Diagnosis Method for Rolling Bearings Based on Improved Residual Shrinkage Network Combined with Transfer Learning. *Sensors*, 24(17), Article 17. <https://doi.org/10.3390/s24175700>
- Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. *Reliability Engineering & System Safety*, 215, 107864. <https://doi.org/10.1016/j.res.2021.107864>
- Thuan, N. D., & Hong, H. S. (2023a). HUST bearing: A practical dataset for ball bearing fault diagnosis. *BMC Research Notes*, 16(1), 138. <https://doi.org/10.1186/s13104-023-06400-4>
- Thuan, N. D., & Hong, H. S. (2023b). *HUST bearing: A practical dataset for ball bearing fault diagnosis* (No. arXiv:2302.12533). arXiv. <https://doi.org/10.48550/arXiv.2302.12533>
- Unal, M., Onat, M., Demetgul, M., & Kucuk, H. (2014). Fault diagnosis of rolling bearings using a genetic algorithm optimized neural network. *Measurement*, 58, 187–196. <https://doi.org/10.1016/j.measurement.2014.08.041>
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- Vashishtha, G., Chauhan, S., Sehri, M., Zimroz, R., Dumond, P., Kumar, R., & Gupta, M. K. (2025). A roadmap to fault diagnosis of industrial machines via machine learning: A brief review. *Measurement*, 242, 116216. <https://doi.org/10.1016/j.measurement.2024.116216>
- Wang, Q., Yan, K., Li, H., & Yuan, M. (2009). Motor noise source identification based on frequency domain analysis. *2009 International Conference on Mechatronics and Automation*, 2373–2377. <https://doi.org/10.1109/ICMA.2009.5246240>
- Wang, X., Mao, D., & Li, X. (2021). Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. *Measurement*, 173, 108518. <https://doi.org/10.1016/j.measurement.2020.108518>

- Wang, Y., Yan, J., Sun, Q., Jiang, Q., & Zhou, Y. (2020). Bearing Intelligent Fault Diagnosis in the Industrial Internet of Things Context: A Lightweight Convolutional Neural Network. *IEEE Access*, 8, 87329–87340. IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2993010>
- Wei, H., Zhang, Q., Shang, M., & Gu, Y. (2021). Extreme learning Machine-based classifier for fault diagnosis of rotating Machinery using a residual network and continuous wavelet transform. *Measurement*, 183, 109864. <https://doi.org/10.1016/j.measurement.2021.109864>
- Wen, L., Li, X., & Gao, L. (2021). A New Reinforcement Learning Based Learning Rate Scheduler for Convolutional Neural Network in Fault Classification. *IEEE Transactions on Industrial Electronics*, 68(12), 12890–12900. IEEE Transactions on Industrial Electronics. <https://doi.org/10.1109/TIE.2020.3044808>
- Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, 17(1), 26–40. <https://doi.org/10.11989/JEST.1674-862X.80904120>
- Wu, J., Tang, T., Chen, M., Wang, Y., & Wang, K. (2020). A study on adaptation lightweight architecture based deep learning models for bearing fault diagnosis under varying working conditions. *Expert Systems with Applications*, 160, 113710. <https://doi.org/10.1016/j.eswa.2020.113710>
- Xie, F., Wang, L., Zhu, H., & Xie, S. (2023). Research on a Rolling Bearing Fault Diagnosis Method Based on Multi-Source Deep Sub-Domain Adaptation. *Applied Sciences*, 13(11), Article 11. <https://doi.org/10.3390/app13116800>
- Xie, W., Li, Z., Xu, Y., Gardoni, P., & Li, W. (2022). Evaluation of Different Bearing Fault Classifiers in Utilizing CNN Feature Extraction Ability. *Sensors (Basel, Switzerland)*, 22(9), 3314. <https://doi.org/10.3390/s22093314>
- Xu, J., Zhou, H., & Fang, Y. (2022). Intelligent Roller Bearing Fault Diagnosis in Industrial Internet of Things. *Wireless Communications and Mobile Computing*, 2022(1), 1860946. <https://doi.org/10.1155/2022/1860946>
- Xu, W., & Li, H. (2022). Bearing Fault Diagnosis Based on Hermitian Wavelet and One-Dimensional Convolutional Neural Network. In X. Li (Ed.), *Advances in Intelligent Automation and Soft Computing* (pp. 369–375). Springer International Publishing. https://doi.org/10.1007/978-3-030-81007-8_41
- Xu, Y., Jiang, Q., Li, S., Zhao, Z., Sun, B., & Huang, G. Q. (2024). Digital twin-driven discriminative graph learning networks for cross-domain bearing fault recognition. *Computers & Industrial Engineering*, 193, 110292. <https://doi.org/10.1016/j.cie.2024.110292>
- Yang, S., Yang, P., Yu, H., Bai, J., Feng, W., Su, Y., & Si, Y. (2022). A 2DCNN-RF Model for Offshore Wind Turbine High-Speed Bearing-Fault Diagnosis under Noisy Environment. *Energies*, 15, 3340. <https://doi.org/10.3390/en15093340>
- Yang, Y., Fu, P., & He, Y. (2018). Bearing Fault Automatic Classification Based on Deep Learning. *IEEE Access*, 6, 71540–71554. IEEE Access. <https://doi.org/10.1109/ACCESS.2018.2880990>
- Zhang, J., Sun, Y., Guo, L., Gao, H., Hong, X., & Song, H. (2020). A new bearing fault diagnosis method based on modified convolutional neural networks. *Chinese Journal of Aeronautics*, 33(2), 439–447. <https://doi.org/10.1016/j.cja.2019.07.011>
- Zhang, R., & Gu, Y. (2022). A Transfer Learning Framework with a One-Dimensional Deep Subdomain Adaptation Network for Bearing Fault Diagnosis under Different Working Conditions. *Sensors (Basel, Switzerland)*, 22(4), 1624. <https://doi.org/10.3390/s22041624>
- Zhang, S., Lv, Q., Zhang, S., & Shan, J. (2021). Multi-attribute quantitative bearing fault diagnosis based on convolutional neural network. *Cognitive Computation and Systems*, 3(4), 287–296. <https://doi.org/10.1049/ccs2.12016>
- Zhang, W., Peng, G., & Li, C. (2017). Bearings Fault Diagnosis Based on Convolutional Neural Networks with 2-D Representation of Vibration Signals as Input. *MATEC Web of Conferences*, 95, 13001. <https://doi.org/10.1051/mateconf/20179513001>
- Zhang, X., Hu, N., Cheng, Z., & Hu, L. (2012). Enhanced detection of rolling element bearing fault based on stochastic resonance. *Chinese Journal of Mechanical Engineering*, 25(6), 1287–1297. <https://doi.org/10.3901/CJME.2012.06.1287>
- Zhang, X., Zhao, B., & Lin, Y. (2021). Machine Learning Based Bearing Fault Diagnosis Using the Case Western Reserve University Data: A Review. *IEEE Access*, 9, 155598–155608. IEEE Access. <https://doi.org/10.1109/ACCESS.2021.3128669>
- Zhiwei, L. (2022). Bearing Fault Diagnosis of End-to-End Model Design Based on 1DCNN-GRU Network. *Discrete Dynamics in Nature and Society*, 2022(1), 7167821. <https://doi.org/10.1155/2022/7167821>
- Zilong, Z., & Wei, Q. (2018). Intelligent fault diagnosis of rolling bearing using one-dimensional multi-scale deep convolutional neural network based health state classification. *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, 1–6. <https://doi.org/10.1109/ICNSC.2018.8361296>