LUT University

BEARING FAULT SIMULATION AND TRANSFER LEARNING

Lappeenranta-Lahti University of Technology LUT

Master’s Programme in Mechanical Engineering, Master’s thesis

2023

Presley Aduwenye

Examiners: Professor Jussi Sopanen

D.Sc. (Tech.) Charles Nutakor

ABSTRACT

Lappeenranta-Lahti Univeristy of Technology LUT

LUT School of Energy Systems

Mechanical Engineering

Presley Aduwenye

Bearing fault simulation and transfer learning

Master's thesis

2023

76 pages, 33 figures and 5 tables

Examiners: Professor Jussi Sopanen
D.Sc. (Tech) Charles Nutakor

Keywords: Bearing fault simulation, Convolution neural network, Transfer learning

Significant number of breakdowns in rotating equipment can be attributed to defective bearings. Thus, the close monitoring of bearing condition has seen increased attention in recent times. The difficulty in sample acquisition associated with defective bearings is one of the bottlenecks stifling the advancement of effective and reliable diagnosis model for real-time bearing condition monitoring.

This thesis aims to simulate the behavior of common bearing defects using a dynamic model that was previously proposed. It also demonstrates how bearing diagnosis knowledge from simulation can be applied in real-world settings, bridging the gap between the two domains and ultimately resolving the pervasive sample acquisition problem.

A simplified CNN architecture was used in developing a pretrained model using simulation data, thereafter, several transfer learning techniques were tested and evaluated on the basis of parameter modification. The results showed remarkable performance in three out of four techniques investigated.

Since the performance and generalizability of data-driven models are tied to the quality and scope of the data used for their development, further studies were suggested in that regard. Overall, a diagnosis framework has been studied and presented which would serve as a base for future research.

ACKNOWLEDGMENTS

Sincerest gratitude to both of my supervisors who have, in no small measures, steered this journey to a logical ending. First, to Professor Jussi Sopanen, my first supervisor, who acted as an umpire by overseeing the work, crosschecking the results and making corrections as and when due. And to Dr. Charles Nutakor, my second supervisor, whose timely and invaluable suggestions stemming from fruitful weekly discussions have guided the trajectory of this work, and ultimately the results presented. Not forgetting Dr. Tuhin Choudhury who did the initial call-up, without whom the route might have been different.

The completion of this thesis would have been a bumpy ride were it not for the love and support from family and friends. Their pieces of advice, even in the dead of night, have overtime and in the course of this work instilled in me the drive to tackle insurmountable problems with sheer alacrity. You are all appreciated!

SYMBOLS AND ABBREVIATIONS

Roman characters

a	Output of filter
C	Damping
c_b	Damping coefficient
c_d	Clearance
D	Raceway diameter
d	Diameter
E'	Modulus of elasticity
e	Eccentricity
F	Force
fs	Shaft's frequency
h	Height
h_0	Lubricant film thickness
I	Rotational inertia
K	Stiffness
K_c	Contact stiffness coefficient
$\bar{k}_e$	Approximated ellipticity parameter
L	Length
M	Total number of categories
m	Mass

N	Number of balls
n_s	Number of samples
n	Rotation speed
P	Output of the specific filter
p	Pooling output
q	Step
R	Curvature sum
R_d	Curvature difference
R_x	Radii of curvature in x-plane
R_y	Radii of curvature in y-plane
r	Radius
s	Output of layer
T	Torque
t	Neuron number
u	Bias
W	Width of pool
w	Weights
x	X-direction
y	Y-direction
Z	Sample
z	Z-direction
$z(f)$	Output of f-th neuron

Greek characters

α	Angle occupied
β	Attitude angle (azimuth angle)
ζ	Second ellipticity integral
δ	Deformation
ξ	First ellipticity integral
η	Viscosity parameter
μ	Viscosity parameter
υ	Poisson's ratio
γ	Misalignment
ϕ	Contact angle
π	Pi
Φ	Position angle

Subscripts

ax	Solid a, x-direction
ay	Solid a, y-direction
B	Bore
b	Ball
bd	Bearing distance
bx	Solid b, x-direction
by	Solid b, y-direction
d	Defect
fric	Friction
h	Housing
in	Inner race
j	jth ball
k	k-th filter
m	Pitch
out	Outer race
rd	Raceways distance
v	Vibration signal
x	x-direction
y	y-direction
z	z-direction

Superscripts

a	Aial
in	Inner
k	k-th filter
l	layer
out	Outer
r	Radial
tot	Total

Abbreviations

1D One-dimensional

2D Two-dimensional

AE Autoencoder

AI Artificial Intelligence

ANN Artificial Neural Network

BPFI Ball Pass Frequency of the Inner Race

BPFO Ball Pass Frequency of the Outer Race

CBM Condition-based Maintenance

CNN Convolution Neural Network

DBN Deep Belief Network

DL Deep Learning

DOF Degree of Freedom

EHL Elastohydrodynamic Lubrication

EMD Empirical Mode Decomposition

FDI Fault Detection and Isolation

FFT Fast Fourier Transform

FTF Fundamental Train Frequency

GAN Generative Adversarial Network

I4.0 Industry 4.0

IFD Intelligent Fault Diagnosis

KNN K - Nearest Neighbor

ML Machine Learning

PCA Principal Component Analysis

PdM Predictive Maintenance

REBs Rolling Element Bearings

ReLU Rectified Linear Units

RF Random Forest

SVD Singular Value Decomposition

SVM Support Vector Machines

TL Transfer Learning

TPU Tensor Processing Unit

VMD Variational Mode Decomposition

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In maintenance and reliability engineering, there are essentially four types of maintenance that are concomitant to the industrial revolutions (Poór, Ženíšek, and Basl 2019, p. 23): reactive maintenance (industry 1.0), preventive maintenance (industry 2.0), proactive maintenance (industry 3.0) and predictive maintenance (industry 4.0). However, there have been several classifications of maintenance strategies over the years. For instance, Shin and Jun (2015), classified maintenance as follows: corrective maintenance, condition-based maintenance (CBM), and preventive maintenance.

Jardine, Lin, and Banjevic (2006, p. 1483) classified maintenance into reactive, scheduled, condition-based and predictive maintenance. Other works have added prescriptive maintenance to the four highlighted in Poór et al. (2019, p. 23). A general classification of maintenance comprising corrective, preventive, condition-based, and predictive maintenance is presented in Figure 1. According to Prajapati, Bechtel, and Ganesan (2012, p. 384), CBM was first known as predictive maintenance (PdM) when the Rio Grande Railway Company first conceived it during the 1940s. There is a connection between CBM and PdM strategies, and unlike others are regarded to be a part of modern smart manufacturing or industry 4.0 (I4.0).

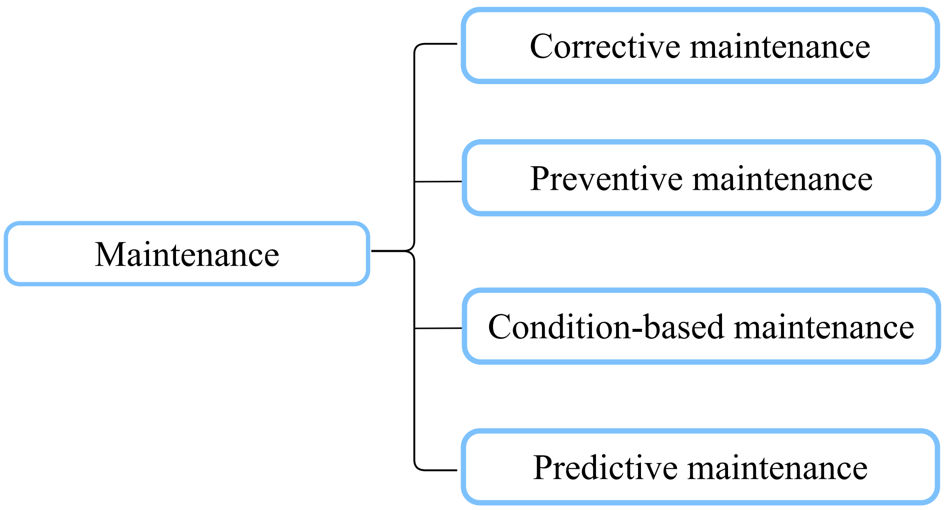


Figure 1: Maintenance strategies.

Condition monitoring, which is the main theme of this study and the driver of CBM can either be used for the prognosis of machine components (e.g., predicting a gear's remaining useful life) or diagnosis of machine components (e.g., detection and identification of incipient faults). Inaccurate diagnosis as well as late detection has proved to be costly and since the complexity of industrial systems and processes has increased, the challenges and consequences are enormous. The application of artificial intelligence (AI) to several fields has made a significant impact, and thus there has been a considerable shift in research related to condition monitoring in machines from traditional techniques which are often human and operating characteristics dependent to AI based methods.

This research focuses on how data associated with rotor-bearing faults from earlier developed dynamic models can be utilized in building models capable of intelligent fault detection and diagnosis. It begins with a comprehensive literature review of possible bearing faults, followed by an assessment of the contributions of AI including transfer learning (TL), for bearing diagnosis. Next, a review of the dynamic model earlier developed for each of these possible faults, followed by the TL model used for this study. Finally, case studies are presented.

1.1 Background

Bearings are machine components that are designed to support another machine element that is either rotating or oscillating (Chaudhari, Thakare, Patil, and Gunjal 2015, p. 1674). The basic functions of bearings are to reduce friction between moving parts, support radial and axial loads, and transfer forces to support structure. They typically comprises several balls or rollers, enclosed between a pair of rings, also known as races, that roll or rotate in response to the applied load. (Yakout and Elkhatib 2013, p. 3288.) The structure of Rolling Element Bearings (REBs) is dependent on its application as well as its type. An exploded view of the general components contained in a bearing is shown in Figure 2 below. REBs also allow for high-speed operation and are capable of efficient power transmission, which makes them ideal for different uses such as pumps, motors, gearboxes, and other high-speed rotating machinery.

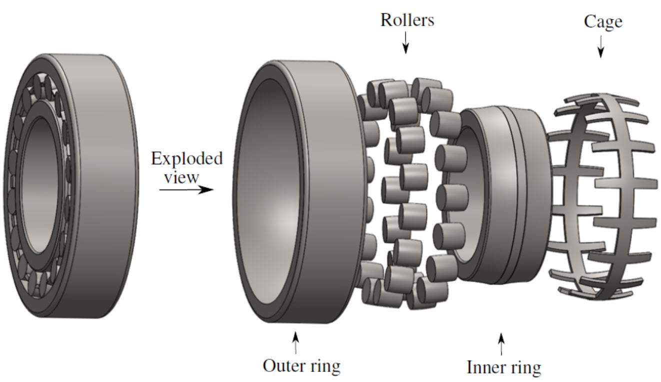


Figure 2: Spherical roller bearing (Ghalamchi 2014, p. 17).

Bearings are typically designed with accuracy and precision because their service performance is directly related to these (He, Zhang, Geng, Wang, Luo, Zhang, and Ren 2022, p. 3493). However, while in service, defects are induced in them due to several factors such as misalignment, poor lubrication, improper installation, fatigue, and contamination. Defects in bearing are classified as either localized or distributed defects. Localized defects refer to specific, isolated areas of damage or wear within the bearing whereas distributed defects refer to more widespread areas of damage or wear that are not limited to a specific area.

Many techniques are available for monitoring the health of bearings, however, vibration monitoring is widely regarded as a superior technique due to its dependability and exceptional responsiveness to the magnitude of defects (Patidar and Soni 2013, p. 1804). All vibrations are associated with a certain frequency, enabling analysts to use the frequency spectrum associated with a given system for identifying deviations. The raceways and roller faults give off high amplitude at unique frequencies which are referred to as the defect frequencies. The frequencies are obtained using the pitch diameter and rotational speed of the balls. In particular, the number of balls is a significant factor in detecting faults in raceways. (Saruhan, Sandemir, Çiçek, and Uygur 2014, pp. 384 - 385.)

Generally, there is a predictable course to failure of bearings which can be detected by vibration analysis (Alsalaet 2020). The four stages are:

1. Early signs of problems in bearings occur in the ultrasonic range between 5 Hz – 50 kHz and are measured using spike and shock pulse instruments.

2. During the second stage, small defects become apparent, aligning with the bearing's natural frequencies which typically fall within the 500 Hz to 2 kHz range.

3. The third stage shows the bearing defect frequencies on the spectrum as well as their harmonics which is an indication of the severity of the wear. These characteristic frequencies include the rollerpass frequency of the inner and outer races commonly referred to as (BPFI) and (BPFO), and the fundamental train frequency associated with the cage.

4. This is the final stage and damage is exponential during this stage.

While vibration analyses can show these faults, it is challenging in practice especially when they are early and there is low signal-to-noise ratio. In lieu of this, motor current signature analysis is utilized, but it faces practical issues such as variation in stator currents at different operating conditions. Furthermore, data needs to be taken from a healthy motor while it is operating under varying loads and speeds, and this needs a thorough and well-organized setup phase. This procedure, outlined in US5726905 patent (Yazici and Kliman 1998), is a learning phase; nevertheless, it is time consuming and expensive since it needs to be repeated at every unique power rating (S. Zhang, Zhang, Wang, and Habetler 2020, p. 29858).

The major drawback of these methods above is that they depend on the fault characteristic frequencies for bearing fault identification. Practically, the signals extracted are usually complex and a combination of faults can quickly change the fault features which might make human interpretation prone to errors. Although a professional with expert knowledge may be able to diagnose the faults, the application of AI for intelligent fault diagnosis can greatly increase the accuracy and speed thereby reducing the maintenance cycle.

1.2 Motivation

Bearings are often points of manifestations of problems developing or have already occurred in rotating machines. This is because they are subjected to the transmission of loads thereby playing an essential role in the system. (De Silva 2007, p. 213.) These problems could be due to unbalance, misalignment, high dynamic loads, or contaminants. An estimated 76.2% of breakdowns in the wind turbine drivetrains can be attributed to defective bearings (Sheng 2017, p. 3). Therefore, bearing faults detection and isolation is crucial if machine health is to maintain good working conditions. To achieve this, various faults prevalent in REBs need to be understood as well as the underlying physical phenomena contributing to the fault(s).

Studies in this regard would typically need a thorough analysis of the historical data related to a specific fault. But since the performance of REBs are generally operating condition dependent, it makes the whole process challenging. The safety risk a faulty equipment poses make data acquisition difficult, and the multi-physics nature of the problem further increases the complexity. Suppose there are difficulties in obtaining sufficient historical data for training a model, then, one approach to solve this would require one to use artificial (or simulated) data to make the diagnosis model, and subsequently for forecasting the condition of real equipment. Studies have revealed that this will underperform. The model cannot perform satisfactorily on the data associated with real equipment because its domain possesses features that makes it different from that of real equipment. This is a domain adaptation problem that may be regarded as a subset of TL. Since I4.0, AI has gained traction in recent years as it has proved superior in research in numerous fields. This thesis investigates using AI models to parse faulty data, learn from it, and then by means of transfer-learning identifying faults at the target.

1.3 Scope and objectives

This investigation focuses exclusively on simulating faults that may occur in specific components of bearings, namely, the outer and inner races. A deep groove ball bearing dynamic model earlier developed (Sopanen and Mikkola 2003a) and implemented in a multibody dynamic software, MSc Adams (Sopanen and Mikkola 2003b) will be further extended to the RoBeDyn toolbox. RoBeDyn is a rotor-bearing dynamic toolbox in Matlab developed at the Machine Dynamics Laboratory at LUT University. This study will further demonstrate how transfer learning can be implemented by exploiting the information contained in the source (simulated data) for fault identification at the target (real data). Transfer learning will be implemented using Google's TensorFlow framework, a leading machine learning framework used in the development of the Tensor Processing Unit (TPU), which enables fast AI model training.

Given the challenges stated in earlier sections, the objective is outlined as follows:

1. Identification of possible faults in a rolling element bearing
2. Extending the functionality of RoBeDyn for defective bearing simulation
3. Develop a DL model capable of fault recognition at the source (simulated data) level
4. Transfer of diagnosis knowledge to target (real data) level considering insufficient samples and domain discrepancies.

1.4 Structure of the thesis

The thesis' remaining sections are organized as follows:

Chapter 2 is dedicated to a review of literature pertinent to the study. It starts with an explanation of the concept of faults and the approaches that have been taken to model fault indicators. Next, a short but detailed review of dynamic bearing models that have been developed in the past. Finally, a review of the applications of some ML, DL, and TL theories.

Chapter 3 focuses on the ball bearing model and highlights key details from the previous study conducted. The significance of characteristic frequencies for bearing condition monitoring is also discussed in this chapter.

Chapter 4 introduces transfer learning, a powerful machine learning approach. It begins with a concise introduction to Convolutional Neural Networks (CNN) and describes some of the typical layers found in these networks. The chapter then presents the specific CNN architecture utilized in this study, along with an explanation of the different transfer strategies employed.

Chapter 5 is devoted to the case studies that were undertaken. The chapter presents two specific cases, each addressing different aspects of the study. The first case revolves around bearing defect detection utilizing the Fast Fourier Transform (FFT) technique. The simulation setup involves using the same equipment with a defective bearing as in the previous study, enabling a meaningful comparison and contrast of the results.

The second case explores fault recognition, specifically emphasizing the application of transfer learning. The simulation environment is meant to model a real-life situation, demonstrating how diagnostic information may be transferred from one setting to another. This case study demonstrates the effectiveness and applicability of transfer learning in practical contexts.

Chapter 6 presents a comprehensive analysis of the study's findings in relation to its set objectives and previous studies. It outlines the implications and reliability of the results, while also highlighting the contributions made by the study as well as its limitations.

Chapter 7 marks the concluding chapter of this study, providing a summary of the research conducted. It presents the key findings and conclusions drawn from the study. Additionally, the chapter offers insightful suggestions for future research directions and potential enhancements to further improve the obtained results.

2 LITERATURE REVIEW

This chapter examines available literature on bearing faults and the methods that have been developed to identify them. The AI section of this chapter delves into the many machine learning approaches used for defect detection and identification, highlighting the most important research in this area. Furthermore, the chapter examines the application of transfer learning in intelligent fault diagnostics, reviewing relevant research and its implications in the field.

2.1 Concept of faults

When a system's defining attribute deviates from its norm, we say that the system has a fault. Faults, if not detected and isolated, could lead to failure. Failure is an interruption of a system from functioning as desired. Faults can generally occur in three areas and are subsequently categorized as sensor faults, actuator faults, and plant faults (Larson, Parker, and Clark 2002, p. 4215). In order to maintain the functionality of systems, efforts have been made to identify faults in their early stages, isolate them, and sometimes, reconfigure them.

Over time, three primary methods for developing fault-sensitive indicators have emerged: data-driven, hybrid, and model-based strategy (H. Li and Xiao 2011; Tidriri, Chatti, Verron, and Tiplica 2016) as shown in Figure 3 below.

2.1.1 Model-based approach

This approach is divided into quantitative and qualitative methods, the former of which is accomplished via the use of mathematical relationships to connect the input and output, and the latter of which is accomplished through the utilization of a qualitative function. Understanding the process and the system at hand, generally from first principles, is usually necessary for success when using model-based approaches. The method relies on residuals, and system inconsistencies, to indicate faults.

In general, studies using model-based quantitative approaches can be grouped into observers (Du, Hou, and Wang 2022; Trajin, Regnier, and Faucher 2008; Ye, Zhang, and Yang 2022), frequency domain methods (Abboud, Sieg-Zieba, and Eltabach 2017; Mba and Reuben 2014), parameter estimation (Attaran and Ghanbarzadeh 2015; C. Li, Ledo, Delgado, Cerrada, Pacheco, Cabrera, Sánchez, and Valente de Oliveira 2017) and so on. Its counterpart, qualitative model-based approaches can be grouped into fault trees, digraphs, etc.

While the key advantage of the model-based approach is its cost-effectiveness since the implementation of fault detection requires no hardware, there is a challenge in modelling reliable and accurate models.

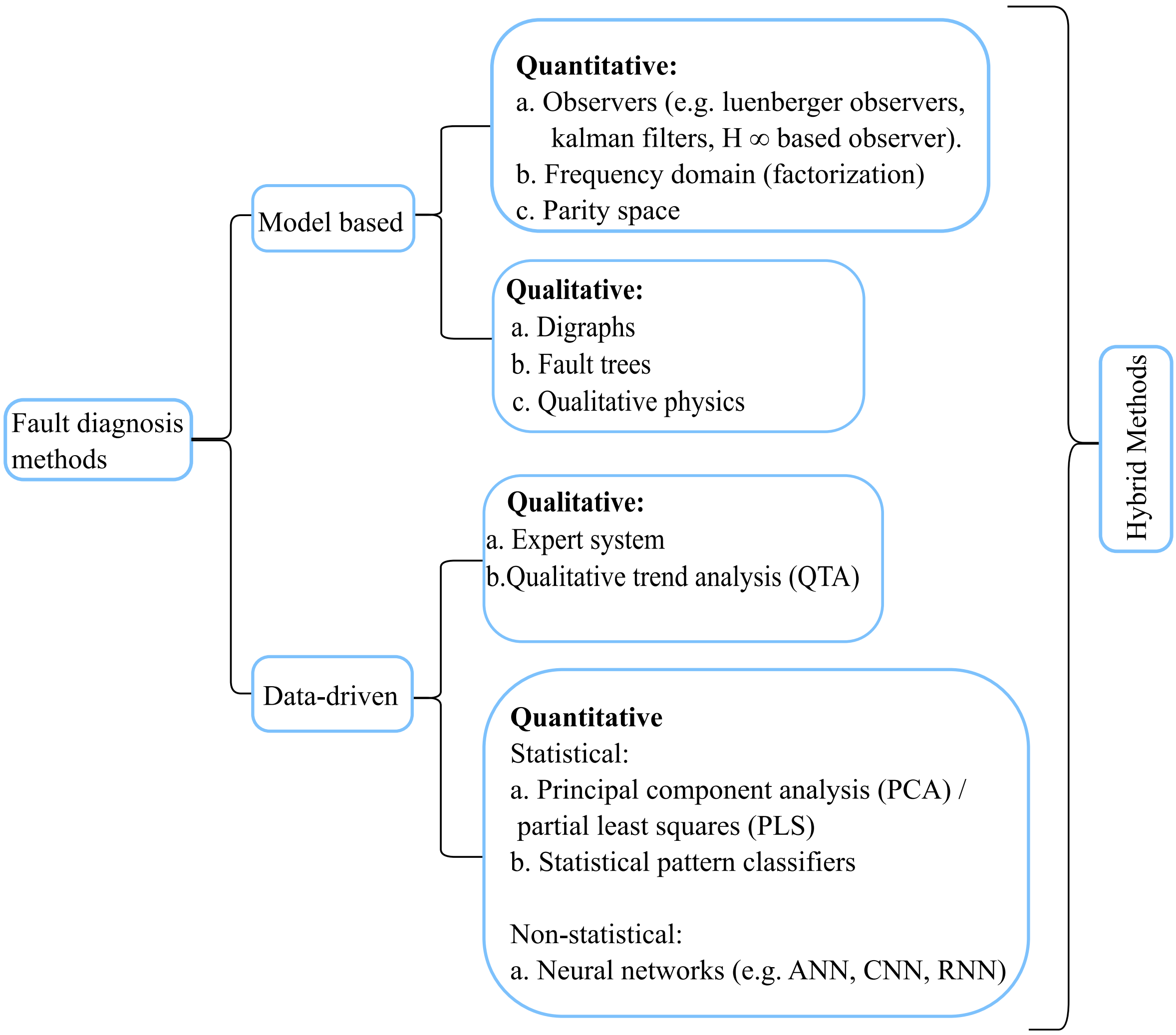


Figure 3: Fault diagnosis methods.

2.1.2 Data-driven approach

This approach is dependent on historical data, usually a large data set. A typical diagnostic system using data is essentially a detection and classification system. Feature extractions are performed on the data to aid diagnosis. Comparable to the model-based strategy, it is segmented into quantitative and qualitative sub-strategies, with the former including supervised and unsupervised learning approaches.

Most works in this area use methods such as artificial neural networks (Samanta and Al-Balushi 2003; Pandya, Upadhyay, and Harsha 2012), partial least squares (Cui, Hong, Qiao, and Yin 2017; F. Q. Li, Fan, and Zhang 2014), principal component analysis (De Moura, Souto, Silva, and Irmao 2011; Shi, Guo, Yuan, Liu, Hou, and Sun 2020) and so on. Qualitative works employing a data-driven approach employ the use of expert systems (Berredjem and Benidir 2018; T. I. Liu, Singonahalli, and Iyer 1996) and trend analysis.

Since I4.0, there has been a considerable shift towards AI methods and more research in literature is exploring their use in FDI. The latter portion of this chapter provides a more in-depth analysis of recent developments. However, this method relies on having sufficient data.

2.1.3 Hybrid approach

The need for a robust and reliable fault detection system especially for large and complex systems capable of realizing all faults has necessitated research into the possibility of integrating both approaches.

One of the early efforts was made by Mylaraswamy and Venkatasubramanian (1997), and it consisted of employing a combination of signed digraphs, qualitative trend analysis, and a statistical classifier based on probability density functions. During the test, the model was capable of detecting and recognizing the faults occurring in the plant. This laid the groundwork for subsequent studies that made use of the signed digraph, some of which were particularly concerned with isolating a particular type of residual generator within the data. Using bearings as an example, K. Li, Zhang, and Li (2012) showed that Kalman filtering and Support Vector Machines (SVM) may be combined to group defects with an accuracy of 96.6% or higher.

Overall, there are limited works that have provided a general framework for the fusion of a data-driven approach and a model-based approach. Although the benefits are enormous, it is still an emerging field.

2.2 Rolling element bearing models

Before the 1960s, studies on REBs were carried out experimentally. Since then, it has progressed from empirical formulations which were static or quasi-static to theoretical formulations which apply to transient conditions such as time-dependent increase and decrease of rotation speed. Initially, models were developed to describe the normal behaviour of bearing and rotor-bearing systems, the sensitivity of rotating equipment to faulty bearings has necessitated several research in the modelling of defective bearings.

Quasi-static models are often used in bearing design because of their ease of use and ability to accurately describe the stiffness, load distribution, and fatigue life. The solutions from the static equilibrium are utilized as the initial conditions to the differential equations of motion. Richard Stribeck is credited with having carried out one of the earliest studies in this area. Jacobson (2003, p. 784) in ”The Stribeck Memorial Lecture” paper, it is stated that R. Stribeck released his well-known publication, ”Kugellager für beliebige Belastugen” in 1901 where he showed that in radially loaded bearing, the load zone is less than 90°, and the most loaded ball is 4.37 times higher than the average ball load when there is no internal clearance. This was determined using Hertz’s contact theory, and Palmgren subsequently determined that the constant should be increased to five times its original value.

While Stribeck considered only a simple radial load, Rumbarger worked on a model for the axial load. Sjovall went on to show that radial and axial load affect the maximum roller load and bearing load. (Hong and V.-C. Tong 2016, p. 1730.) The Jones model, which Harris modified into what is now known as the Jones-Harris model (Harris and Crecelius 1986), is a model with 5-degrees of freedom (DOF) created to analyze deflections and loads associated with bearings. The bearing model is suitable for situations where there is misalignment or a high level of loading. Currently, the Jones-Harris model is popularly used to calculate bearing deformation.

The previously mentioned quasi-static bearing models can only be used in steady-state situations. When there are changes in rotational speed, such as acceleration or deceleration, or in applications where the speed is high, and unstable motion might lead to the bearing failing sooner than anticipated like in turbines, a dynamic model is necessary. High rotational speeds can also result in failure types that are not typical. Thus, differential equations derived from Newton’s second law replace the quasi-static model’s equilibrium equations. (Hong and V.-C. Tong 2016, p. 1739.)

Dynamic modeling of bearings sparked a lot of interest since Walters (1971) first attempt, and this has culminated in several dynamic models for ideal and defective bearings. Bearing-induced vibration is mostly caused by faults in the bearing’s component parts. These defects are categorized as localized and distributed, depending on the relative area covered by the damage. A pulse is produced when a localized defect passes its corresponding surface which coincides with the defect frequency given that the rotational speed is constant. (Tandon and Choudhury 1999, p. 470.)

Gupta’s 6-DOF model (P. K. Gupta 1984) laid the foundation for numerous types of research for dynamic modelling of bearings. He also studied how out-of-roundness would affect bearing performance and found that only large deviations from circularity will influence it (P. Gupta 2008). On the basis of Gupta’s work, a dynamic high-speed bearing model consisting of defect on the raceways was constructed by Niu, Cao, He, and Li (2014) so that he could investigate the bearing's reaction to vibrations. Likewise, X. Liu, Deng, and Teng (2011) made a dynamic cage model to examine its stability in high-speed applications.

Nutakor, Talbot, and Kahraman (2019) have shown that depending on operating conditions, the lubrication at rolling contacts dictates the rolling-sliding contacts' friction. Hamrock and Dowson (1981) had investigated the elastohydrodynamic lubrication (EHL) associated with lubricant film thickness leading to closed-form equations for finding the minimum or central film thickness for smooth surfaces. After researching the damping capacity of bearings, Dietl, Wensing, and Van Nijen (2000) revealed that the lubricated contact that exists at the roller-raceways has a major influence on the damping.

Sopanen and Mikkola (2003a) created a 6-DOF ball bearing dynamic model which was subsequently implemented in MSC Adams - a commercial multi-body system software application (Sopanen and Mikkola 2003b). Not only did the study consider both localised and distributed defects, but they also studied the effect of diametral clearance which had a significant effect on vibration level, and low-order waviness which caused vibrations whose frequencies are equal to the product of the rotation speed and waviness order. The excitation function for the piecewise displacement was derived using the step as well as tangent functions related to the defect's size unlike Patil, Mathew, Rajendrakumar, and Desai (2010) who employed a half-sine function for modelling the bearing's defect.

Arslan and Aktürk (2008) investigated the response of angular contact bearings under normal and defective conditions using a shaft-bearing model which was regarded as a mass-spring system. To better forecast the dynamics of rotor-spherical roller bearing systems, Ghalamchi, Sopanen, and Mikkola (2015) introduced a novel 3-DOF spherical roller bearing model. The model may be used in both steady-state and transient investigations since it accounts for imperfections such as localized defects and waviness. The model was validated through numerical calculations and showed that it accurately predicts changes in contact forces, bearing displacement, elastic deformation, and diametral clearance with varying radial loads. In their subsequent paper, Ghalamchi, Sopanen, and Mikkola (2016) shed more light on spherical roller bearings with localized defects where they presented analytical formulations for the defect depth and stiffness calculations.

Studies on high-speed applications, and other forms of defects such as waviness have also seen increased attention since the latter was first investigated by H. Perret, and subsequently E. Meldau (Harsha and Kankar 2004, p. 1058). Hou, Wang, Xie, and Peng (2021) recently revised a ball bearing model to assess outer race waviness-induced vibration at high speeds and axial loads. Kurvinen, Sopanen, and Mikkola (2015) improved the work by Sopanen and Mikkola (2003a) by taking into account centrifugal forces and gyroscopic moments.

Y. Wang, Wang, Zhang, and Zhao (2015) created a dynamic model for analyzing skidding's influence in angular contact ball bearings. This model takes into consideration the interactions that occur between the components of the bearing and the lubricant. According to the findings of the research, the axial load has a significant impact on the behavior of skidding. This is due to differences in the load and the rotational and orbital speed of the ball that occur as a consequence of different operating conditions. Also, J. Liu, Xu, and Pan (2021), suggested a model that accounts for differential sliding friction torques, EHL, and elastic hysteresis. The research compares the new model to older ones and investigates how bearing speed, load, and defect size affect the bearing's acoustic properties.

2.3 Bearing fault diagnosis

Detecting and isolating faults as well as other processes that go beyond just determining the existence and cause of a defect are all included in the concept of fault diagnostics. After locating and isolating the issue, the system is examined to ascertain its underlying cause and gauge its effect on the system's operation. A diagnosis is given based on this analysis, which identifies the issue, its fundamental cause, and any required remedial measures.

Study conducted by Peng, Bi, Xue, Zhang, and Wan (2022) showed that the methods associated with fault diagnosis of REBs can be classified as fault recognition and detection. The focus of this study, fault recognition, is concerned with the classification of the faults using feature extraction techniques or automatically learning features from the data, whereas fault detection methods rely on the defect frequencies earlier discussed. The diagnostic process is shown in the flowchart presented in Figure 4.

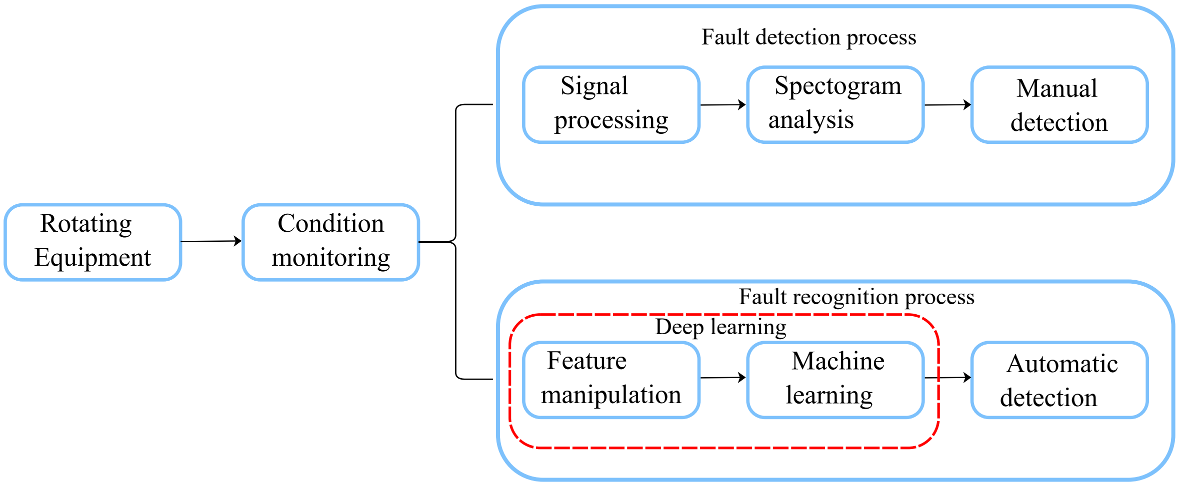


Figure 4: General flowchart of the diagnosis process.

2.3.1 Fault detection

Vibration signals are often indirectly obtained using a transducer mounted on the bearing support, which leads to the introduction of noise into the data. In order to reduce this interference, magnify the signals associated with the fault, expedite the fault detection process, and boost its effectiveness, a number of fault detection approaches based on vibration analysis are available in literature. Figure 5 below illustrates how Peng et al. (2022, p. 347) classifies these techniques into four broad categories: morphological transformation techniques, filter techniques, decomposition techniques, and deconvolution techniques.

Morphological transformation-based methods are a category of signal processing techniques that use mathematical transformations to extract or enhance features in a signal for fault detection. The mathematical morphology theory, a concept concerned with the study of forms and structures, was first used as a denoiser in image processing by Matheron, and subsequently in signal processing by Maragos (Matheron 1975; Randall and Antoni 2011). After that, Nikolaou and Antoniadis (2003), used it for demodulating the defective signal to obtain an envelope from it. Some common examples under this category that are used in defective bearing studies include composite morphological filters, morphological gradient-based methods, and multi-scale morphological filters.

The filter-based methods are a category of signal processing techniques that use filters to extract features and can effectively identify and remove noise to improve fault diagnosis accuracy. Despite their effectiveness, they are limited by the selection of filter parameters and design, which can complex and laborious. The term ”optimal wavelet filter” refers to a class of wavelet filters that aim to optimize the identification of important signal characteristics while limiting the detection of noise. Some of the filters are morlet, symlets, and coiflets.

With decomposition-based methods, the signal is broken down into several basis functions that correspond to the signal's various frequency components. It is then possible to separate the fault-related components and eliminate the noise and other non-fault-related components. Well-known techniques used for REBs fault detection in this category include the local mean decomposition (LMD), empirical mode decomposition (EMD), varitional mode decomposition (VMD) and singular value decomposition (SVD).

Deconvolution-based techniques are used to identify faults by eliminating the transmission path's impacts from the measured signal. By using an inverse filter to simulate the transmission path and account for its impacts on the observed signal, it is possible to estimate the fault-related signal.

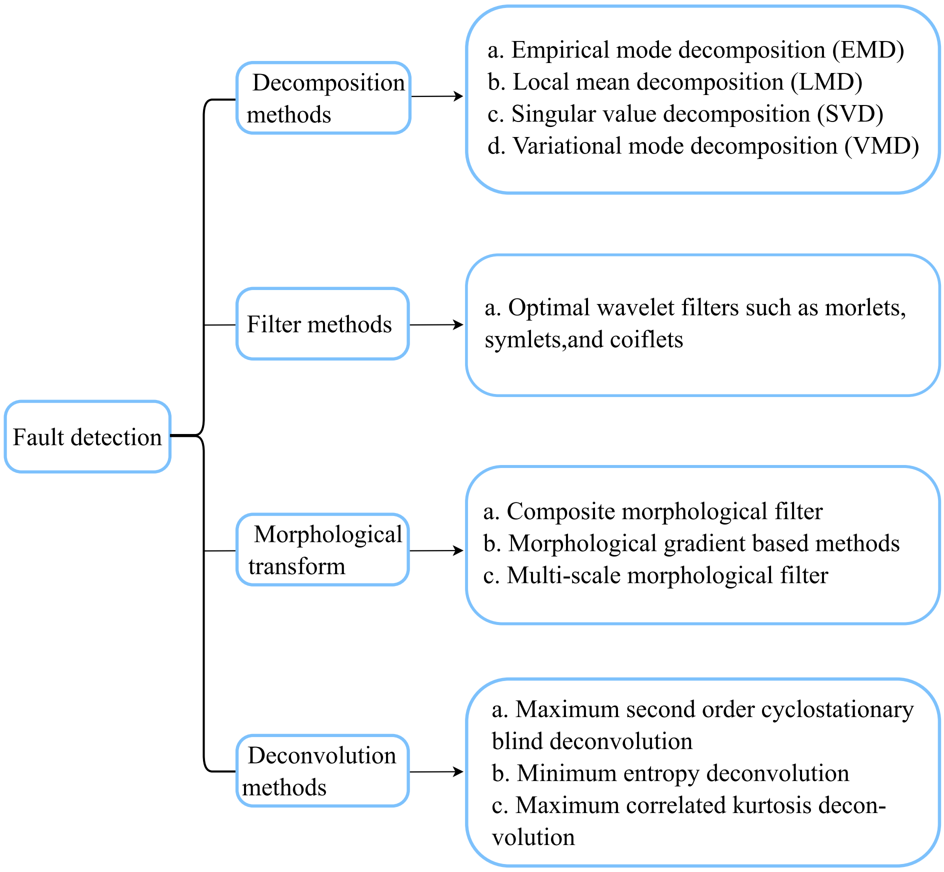


Figure 5: Fault detection methods.

2.3.2 Fault recognition

In contrast to manual detection methods that require observation of the spectral lines associated with the defects on the spectrum, the use of artificial intelligence enables knowledge extraction from the given data, learning features and classification which facilitates the recognition of bearing faults (Lei, Yang, Jiang, Jia, Li, and Nandi 2020, p. 2). Figure 6 shows that fault recognition may be split into deep learning and traditional approaches.

In traditional approach, important features are extracted to create a feature set. This includes calculating statistical descriptive measures like root mean square, skewness, and root mean square frequency in the time or frequency-domain, to extract properties from vibration signals. The feature set is then decreased in dimension while keeping the essential details required for classification during the subsequent feature reduction stage. According to Basu and Andrews (2013), this is crucial since having too many features might cause the model to overfit and lose some of its generalizability. Techniques like principal component analysis may be used for achieving this. In the final phase, classification, the input data is divided into several categories, such as normal and defective, using the reduced feature set. This is often accomplished using either an unsupervised learning technique, where the model de-

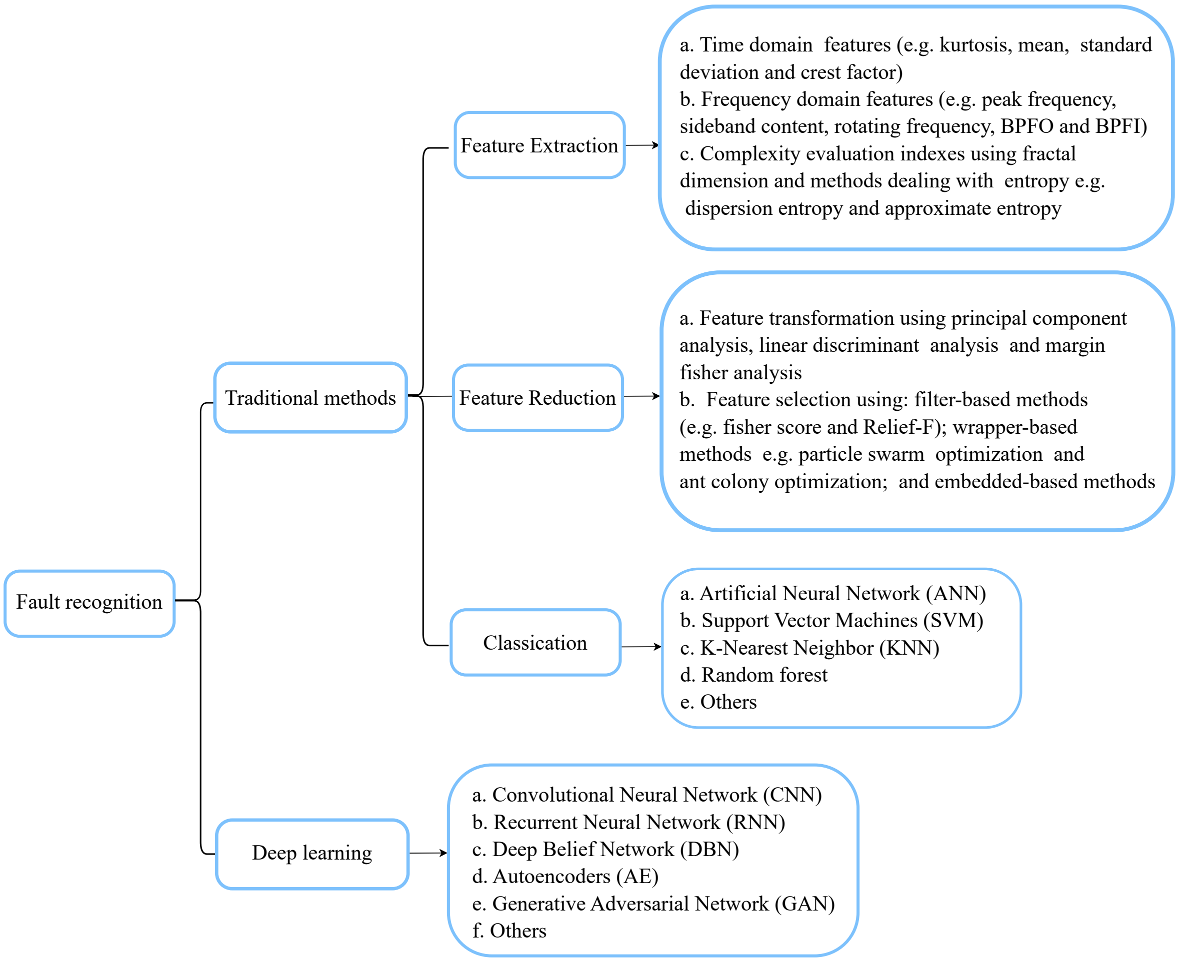


Figure 6: Fault recognition methods.

tects patterns in the data without previous labeling, or a supervised learning approach, where the training is conducted using labeled data. SVM, ANN, and decision trees are common classification techniques used.

Deep learning approaches to bearing fault recognition involve training deep neural networks to learn and automatically extract relevant features from vibration signals. They are designed with multiple layers that can learn abstract representations of the data, allowing them to understand complex relationships between the input data and the fault types (R. Liu, Yang, Zio, and Chen 2018, p. 36). Deep learning has demonstrated promising results in bearing defect detection, with some studies showing high classification accuracy. However, training these methods requires a lot of data and computational power, which may restrict their use in certain applications. Furthermore, it may be challenging to interpret the learned features and the underlying fault mechanisms due to their black-box nature (Camilleri and Prescott 2017, p. 87). These techniques and those discussed in the traditional methods are further explicated in the next section.

2.4 Application of artificial intelligence to bearing fault diagnosis

The term ”artificial intelligence” refers to computational methods that simulate human cognitive processes such as reasoning, classification, and decision-making (Voda and Radu 2019, p. 199). According to Lei et al. (2020, p. 2), IFD refers to the use of AI for machine fault diagnostics. The purpose of IFD is to create diagnostic models that can learn from data and then use that information to determine the equipment's health with little or no human involvement.

ML and DL are two subsets of AI that has been applied for IFD as depicted in Figure 6. The application of AI to bearing fault diagnosis has been studied as a classification problem over the years where the data associated with the fault labels are used for training, validation, and testing.

2.4.1 Application of machine learning

According to El Bouchefry and Souza (2020, p. 225), ML is a subfield of AI that enables systems with the capacity to infer meaning from data and improve their operation over time with no further programming. Traditional approaches in IFD include the use of ML algorithms like ANN and SVM. They typically require manual extraction and subsequent manipulation of the features associated with the collected data. A short review of their application to bearing faults is further presented below.

ANN was designed using the human central nervous system as a base, and is regarded as one of the oldest AI paradigm (Makokha 2021, p. 14). One of the first investigations of bearing diagnostics in induction motors was conducted by Chow, Mangum, and Yee (1991) using neural networks. Despite a detection accuracy of 94%, the study could be improved by using a speed encoder which was not readily available for small motors. Fault detection methods such as EMD have also been used in conjunction with ANN for effective diagnosis (Yu, Yu, and Junsheng 2006). ANN has slow convergence, and interpretability is challenging due to its black-box nature.

SVM falls into the supervised learning category, useful for classification and regression analysis. Its algorithm works by finding the ideal hyperplane that categorizes the data points into various groups or forecasts the value of the target variable in regression problems. From literature, not only has SVM proven to be superior to ANN in classification (Rojas and A. K. Nandi 2005), but also current studies are primarily concerned with improving its diagnostic accuracy by either modifying the SVM (Sugumaran, Muralidharan, and Ramachandran 2007; F. Chen, Tang, Song, and Li 2014) or the optimization of it (Jack and A. Nandi 2002; Samanta, Al-Balushi, and Al-Araimi 2003; Zhu, Song, and Xue 2014; Y. Li, Miao, Zhang, Chen, Liu, and Jiang 2019).

The ”k” in the KNN algorithm denotes how many closest neighbors will be taken into account when making a prediction. First, the difference between each training data point and the input data is calculated. Thereafter, k nearest neighbors are selected based on this distance, and subsequently assigned to a class (for classification) or value (for regression) that is most common among these neighbors as the predicted value. Mechefske and Mathew (1992) conducted the first study using KNN in low-speed bearings to ascertain whether short vibration signals may be useful in identifying rolling element bearing defects. The study showed that the method was effective in providing diagnostic results, without the need for an experienced analyst. Reduction methods involving principal component analysis (PCA), are commonly used in conjunction with k-NN. Generally, k-NN based models are easy to implement and the parameter ”k” should be carefully chosen because it is a crucial factor that significantly impacts algorithm performance.

In 2001, Breiman suggested the use of Random Forest, useful for classification and regression (Breiman 2001). With this method of ensemble learning, the results of many different decision trees are pooled together to get a decision. When random forest receives the test data, each decision tree in the ensemble independently classifies the data. The final classification is then determined by aggregating the votes from all the trees, with the most frequently predicted class being selected as the final output. In other words, the outcome is dependent on the predictions of most trees. RF has also been used with methods like genetic algorithm (B.-S. Yang, Di, and Han 2008) and VMD (T. Han and Jiang 2016) for effective fault diagnosis.

2.4.2 Application of deep learning

DL techniques do not explicitly require feature manipulation prior to being fed into the algorithm (or network) unlike ML methods. They are capable of handling the raw data, learning from them, and subsequently able to recognize the condition of the associated machine(s). DL models are structured to form a hierarchy, and abstracted features are learned on a layer-by-layer basis. For multi-class problems, a deep neural network’s last layer (output) will include several neurons as there are classes or labels in the input, or only one for a binary problem.

The human visual cortex served as inspiration for CNN, which has found widespread application in the areas of image and video processing, identification, and classification. It is the most used deep learning method for bearing diagnostics. A CNN may receive input in the form of one-dimensional (1D) vibration signals as well as two-dimensional (2D) images. Networks may be built using numerous layers of convolution and pooling layers, which are connected to flattening, and fully connected nodes, which would ultimately be passed to a softmax or sigmoid function.

Janssens, Slavkovikj, Vervisch, Stockman, Loccufier, Verstockt, Van de Walle, and Van Hoecke (2016) published the first paper showing the use of CNN for bearing fault diagnosis. Two uni-axis accelerometers were used for the collection of vibration data, which was pre-processed using scaled discrete Fourier transform enabling a CNN to independently learn bearing defect detection characteristics. The work not only showed the significance of CNN above classical ML methods for bearing fault diagnosis but also demonstrated its relevance for bearing prognosis. Eren (2017) employed a 1D CNN using raw vibration data to illustrate the usefulness of CNN after removing the computational complexity of conventional approaches, achieving over 97% accuracy. Due to CNN's superior performance with 2D data, the 1D signals are often reshaped (S. Li, Liu, Tang, Lu, and Hu 2017; Fuan, Hongkai, Haidong, Wenjing, and Shuaipeng 2017) or pre-processed (Verstraete, Ferrada, Droguett, Meruane, Modarres, et al. 2017) to enable conversion time to time-frequency domain.

AE is a kind of neural network often employed for pretraining, it is suitable for data compression and reconstruction. It consists of an encoder and a decoder, where the former transforms the input data into a representation in a lower dimension, and the latter returns it to its original form. Its goal is to reduce the sum of squared deviations between the original data and the reconstructed version. There are numerous studies in bearing fault diagnosis associated with several types of AE, each with a different architecture and objective. Two prominent ones are the stacked autoencoder (Xiang, Zhang, Zhang, and Xia 2019), the denoising autoencoder(Gu, Cao, Song, and Yao 2021; X. Guo, Shen, and Chen 2016)) or even a combination of both (Lu, Wang, Qin, and Ma 2017). Overall, autoencoders can be adapted and modified to suit different input and learning objectives.

The architecture of DBN is comprised of stacked layers of Restricted Boltzmann machines (RBMs). An RBM, commonly used in dimensionality reduction and feature learning, has a visible layer representing the input and the features learned are represented by a hdden layer. Literature review showed that DBN is used with other techniques to improve effectiveness. Z. Chen and W. Li (2017) published the first study using DBN for fault diagnosis of bearings where AE was used to fuse the features consisting of time and frequency extracted from several sensors signals, which was then used to train the DBN. The combination of DBN and SVM employing a teager-kaiser energy operator for obtaining the demodulation signal was also suggested by D. Han, Zhao, and Shi (2017). In addition, a new technique that includes compressed sensing (CS) was developed by Shao, Jiang, Zhang, Duan, Liang, and Wu (2018), which outperformed the conventional methods for bearing defect identification.

GAN has changed the game since Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio (2020) introduced it in 2014. The "adversarial" in "adversarial network" refers to the fact that the goals of the generator and discriminator networks in the minimax game are at odds with one another. The discriminator aims to accurately identify real and synthetic bearing data, while the generator strives to generate synthetic bearing data that is comparable to the actual bearing data and capable of tricking the discriminator. Both networks learn from each other as they continually exchange feedback. Not only has GAN been employed as the primary framework for classification (Lee, Jo, and Hwang 2017; Xie and T. Zhang 2018b), but it has also been used for data augmentation to remedy the class imbalance issue caused by small samples in one class compared to the other class(es). Overall, GAN have shown promising results for IFD, and they have the potential for increased reliability and performance of machines by enabling more accurate and efficient fault diagnosis.

2.4.3 Use of transfer learning

One reason for the increased use of transfer learning is related to computation time and cost associated with training a DL model. Unlike ML models that are designed for linear problems and have less dependency on large datasets, DL models deal with non-linear problems requiring large datasets. Apart from the learning costs of training a DL model, there are other challenges such as the need for expert knowledge, and the safety risk a faulty equipment poses cannot be dismissed. TL addresses these issues by leveraging knowledge and learned features from a pre-trained model, and transferring that knowledge to a new, related task. This enables the model to make use of the characteristics learned by the pre-trained model, eliminating the need to retrain the model from scratch. TL can either be used on identical or different machines. Figure 7 shows that the approaches used for TL can be divided into four groups, viz: feature-based, GAN-based, parameter-based and instance-based.

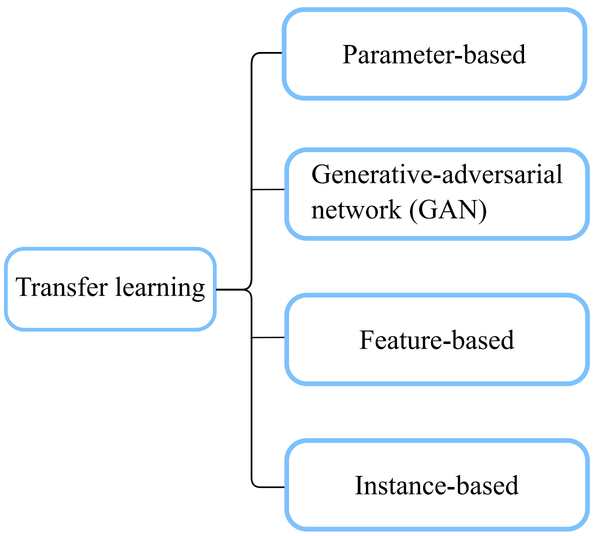


Figure 7: Transfer learning methods.

The parameter-based technique assumes that the target domain contains few labeled samples for diagnosis. Using limited labelled data from the target task, pre-trained model parameters are reused and retrained. VGG-16 network (Mao, Ding, Tian, and Liang 2020; Ullah, Lodhi, and Hur 2020), ResNet (Wen, Li, and Gao 2020) and AlexNet (J. Wang, Mo, Zhang, and Miao 2019) are popular pre-trained models for bearing diagnostics. The trained diagnostic models' parameters are fine-tuned for IFD at the target, using fewer computing resources while processing massive data. The efficiency of this method, however, is strongly influenced by the number of labeled samples in the target.

In the same vein as parameter-based methods, the instance-based methods make an assumption of insufficient data for real diagnosis work. Relevant instances or examples are transferred from the source to the target domain to enhance the diagnostic model. From literature, the Adaboost algorithm has been used to achieve this by adjusting the importance of instances in both domains by assigning weights to them. If samples contained in the source domain are mislabeled, their weight will be weakened. Shen, Chen, Yan, and Gao (2015) used a combination of SVD and TrAdaboost, an extension of Adaboost, for training a kNN-based diagnosis model which proved effective for classification despite the limited target data available. A major advantage of instance-based methods is their ease of implementation.

GAN technique can also be used for TL. It has been used effectively for the transfer of diagnostic knowledge between identical machines as well as different machines. L. Guo, Lei, Xing, Yan, and Li (2018) carried out one of the earliest GAN-based transfer learning between different types of bearings including railway locomotive bearings which fall under the category of different machines. The research found that the proposed method outperformed other well-known methods classification-wise. T. Han, C. Liu, Yang, and Jiang (2019) and Xie and T. Zhang (2018a) have also carried out similar studies, although for identical machines.

In feature-based transfer learning, the focus is primarily on the feature representation rather than the entire model. The idea is to exploit the high-level abstract features learned by a pre-trained model, which captures generic patterns and structures that are transferable across domains. By using these pre-trained features, the model can benefit from the knowledge gained from a large-scale, labeled dataset without requiring extensive training on the target domain. This technique work by mapping certain features into a common feature space and subsequently the distance between features are minimized using an optimization strategy. Typical studies using this technique especially transfers on different machines, were conducted by X. Li, Zhang, and Ding (2018) and Z. Tong, Li, Zhang, Jiang, and Zhou (2018).

3 BALL BEARING MODEL INCLUDING FAULTS

In the study conducted by Sopanen and Mikkola (2003a), a deep groove ball bearing model that constrains relative motion, and aids the support and transfer of loads from the rotor to the supporting structure was developed. The general purpose model proposed consists of 6 degrees of freedom, it uses the Hertz contact theory to describe the deformation and also accounts for the film thickness between components of the bearing.

Non-idealities such as defective inner and outer race were modelled as a dent using a step function while the waviness was presented using a Fourier cosine series. The important details of the analytical model are presented in appropriate headings below.

3.1 Elliptical contact

When a bearing is loaded, the balls or rollers rotate following an elliptical path along the raceway. This is not unconnected to manufacturing tolerances, bearing misalignment, and external stresses. As a result, the point at which the ball or roller makes contact with the raceway is not fixed but rather changes.

For two solids in contact subjected to a given force *F*, Figure 8 depicts the elliptical contact area, such that a positive or negative radius of contact curvature is dependent on the surface's convexity or concavity.

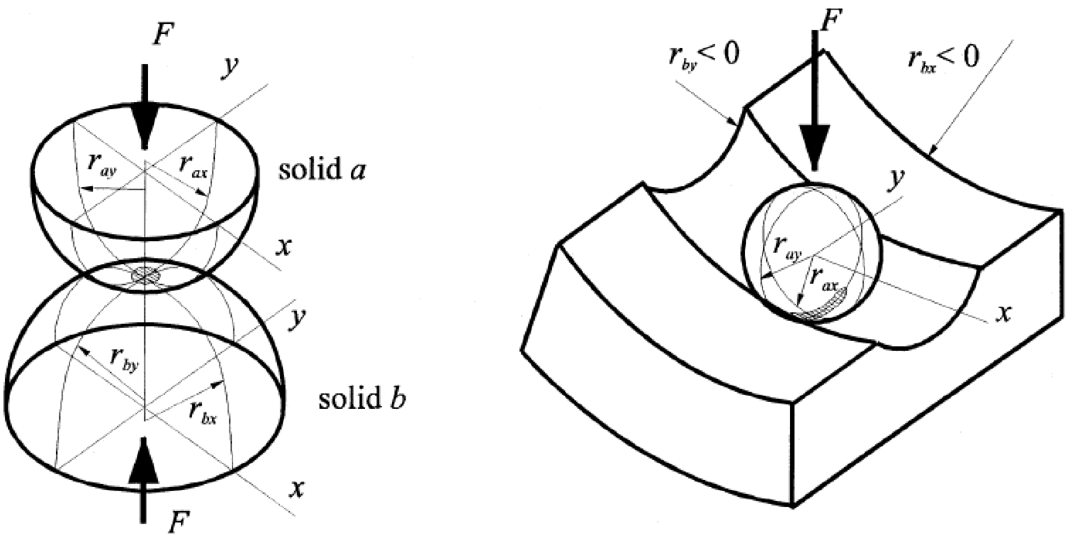


Figure 8: Elliptical contact (Sopanen and Mikkola 2003a, p. 22).

The sum of curvature, denoted by *R*, and the difference between the curvatures, denoted by R_d, are used to describe the deformation as follows:

$$\begin{aligned}&\tfrac{1}{R}=\tfrac{1}{R_x}+\tfrac{1}{R_y}\\&R_{\mathrm{d}}=R\left(\tfrac{1}{R_x}-\tfrac{1}{R_y}\right),\\&\text{where}\\&\tfrac{1}{R_x}=\tfrac{1}{r_{\mathrm{a}x}}+\tfrac{1}{r_{\mathrm{b}x}},\ \tfrac{1}{R_y}=\tfrac{1}{r_{\mathrm{a}y}}+\tfrac{1}{r_{\mathrm{b}y}}\end{aligned} \tag{3.1}$$

The subscripts ax and bx represent the radii in solids a and b in the x direction, while ay and by indicate the radii in the y direction. The effective radii of curvatures are denoted by R_x and R_y. The following formula may be used to determine the contact stiffness:

$$K_c = \pi \bar{k}_e E' \sqrt{\frac{R\bar{\xi}}{4.5\zeta^3}} \tag{3.2}$$

where the ellipticity parameter is denoted as $\bar{k}_e$, the curvature sum as *R*, ξ the elliptical integral of the first kind, *E'* the effective modulus of elasticity and ζ the elliptical integral of the second kind.

3.2 Ball bearing geometry and elastic deformation

Figure 9 illustrates the structure of a ball bearing, whereby each component is labeled using its respective dimension. Specifically, r_{out} and r_{in} refer to the outer and inner groove radius, d_b represents the diameter of the balls, the diametral clearance is denoted as c_d, the bearing housing diameter as D_h, D_i and D_o corresponds to the inner and outer raceway diameter, d_m is the pitch diameter and the bore diameter is denoted as d_B.

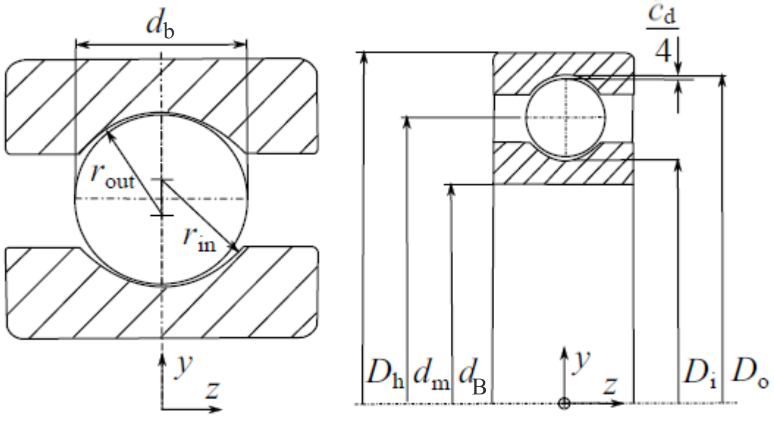


Figure 9: Bearing dimensions.

The contact deflections arising from the ball and raceways may be added together to provide the total deflection, which can be written as:

$$\delta_0 = (\frac{F}{K_c})^{\frac{2}{3}} \tag{3.3}$$

where the normal load is denoted by F, and the contact stiffness is K_c. However, K_c^{tot} which represents the total stiffness coefficient can be expressed as:

$$K_c^{tot} = \frac{1}{(\frac{1}{K_c^{in}})^{\frac{2}{3}} + (\frac{1}{K_c^{out}})^{\frac{2}{3}}} \tag{3.4}$$

where K_c^{in} and K_c^{out} refer to the respective ball-race contact stiffnesses that are affected by the material properties and curvatures of the surfaces in contact.

3.3 Bearing forces and moments

The A-A plane in Figure 10 depicts the ball's axial and transverse directions. The notation e_x and e_y describe the eccentric distance in the global directions, x and y, assuming that the shaft (or inner race) centre point is not coincident with the centre of mass. The axial and radial displacement denoted by e_j^a and e_j^r, at ball i can be expressed as:

$$\begin{aligned} e_j^r &= e_x cos\beta_j + e_y sin\beta_j \\ e_j^a &= e_z - (-\gamma_x sin\beta_j + \gamma_y cos\beta_j)(R_{in} + r_{in}) \end{aligned} \tag{3.5}$$

where γ_y and γ_x are the misalignments between the raceways. The attitude and contact angle of ball *j* can be calculated using the equations below.

$$\begin{aligned} \beta_j &= \beta_j^0 + \tfrac{\theta_{in} R_{in} + \theta_{out} R_{out}}{2(r + R_{in})} \\ \phi_j &= tan^{-1}(\tfrac{e_j^a}{R_{in} + r_{in} + e_j^r - R_{out} + r_{out}}) \end{aligned} \tag{3.6}$$

In Equation 3.6, β_j^0 represents the initial angle of the jth ball and rotation of the raceways is represented as θ_{in} and θ_{out}, with the inner race rotating at speed equal to the shaft and the outer race is fixed. Total elastic deformation for a bearing with a perfect geometry can be expressed as:

$$\delta_j^{tot} = 2r + h_0^{in} + h_0^{out} - L_{rd} \tag{3.7}$$

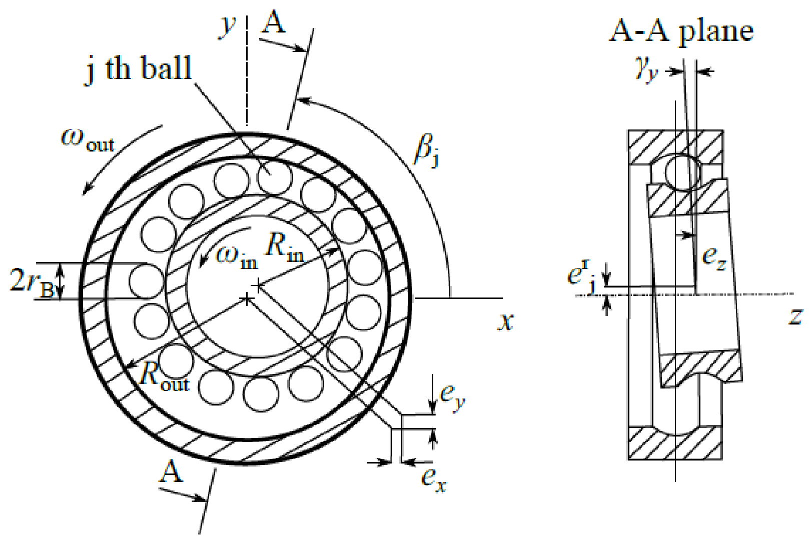


Figure 10: Axial and transverse cross-section (Kurvinen, Sopanen, and Mikkola 2015).

where L_{rd} is the race distance along the contact line as stated below, and h_0^{in} and h_0^{out} represent the lubricant film thicknesses.

$$L_{rd} = r_{out} + r_{in} - \frac{R_{in} + r_{in} + e_j^r - R_{out} + r_{out}}{cos\phi_j} \tag{3.8}$$

Therefore, the force on ball *j* may be expressed as:

$$F_j = K_c^{tot}(\delta_j^{tot})^{1.5} \tag{3.9}$$

Thus, the forces and moments on the shaft in the x, y, and z directions are:

$$\begin{aligned}
F_x &= -\textstyle\sum_{j=1}^{N} F_j cos\phi_j cos\beta_j \\
F_y &= -\textstyle\sum_{j=1}^{N} F_j cos\phi_j sin\beta_j \\
F_z &= -\textstyle\sum_{j=1}^{N} F_j sin\phi_j \\
T_x &= -\textstyle\sum_{j=1}^{N} F_j (R_{in} + r) sin\phi_j sin\beta_j \\
T_y &= -\textstyle\sum_{j=1}^{N} F_j (R_{in} + r) sin\phi_j (-cos\beta_j) \\
T_z &= T_{fric}
\end{aligned} \tag{3.10}$$

where the total friction torque, T_{fric}, is a sum of load-dependent friction torque, viscous friction torque, and seal friction as stated in Sopanen and Mikkola (2003a, pp. 12 - 13).

3.4 Modelling of bearing defects

Geometric imperfections leading to bearing-induced vibrations are (i) localised defects and (ii) distributed defects, where the former has to do with spalls, pits, and dents whereas the latter relates to surface roughness, waviness, misalignments, and so on.

Whenever a ball travels over the surface of the bearing defect, a sequence of contact vibrations is excited. Localised faults defects usually occur in either one of the components or multiple components of a bearing in case of a complex fault. When a ball enters a dent region, it suddenly loses contact with the surface, creating a large impulse force that disrupts the normal motion of bearing components, resulting in changes in the Hertzian contact coefficients.

3.4.1 Defect frequencies

These defect frequencies, also called fault characteristic frequencies, refer to specific frequencies that are generated due to the interaction of the bearing parts and the defects present on their surfaces. In bearing analysis, BPFO and BPFI are two frequently used defect characteristic frequencies.

BPFO describes the frequency with which the rolling element traverses an outer race defect. The BPFO is usually 1 to 5 times that of the bearing's rotational speed. The BPFO can be calculated as follows:

$$BPFO = N\frac{fs}{2}(1 - \frac{d_b}{d_m}cos\phi) \tag{3.11}$$

where N refers to the total number of balls, d_b and d_m refers to the ball and pitch diameter respectively, ϕ is the contact angle, and fs is the frequency of the shaft.

BPFI refers to the frequency with which the rolling element passes an inner race defect. A defect in the inner ring, however, can cause the frequency spectrum to produce extra harmonics. These frequencies are sidebands around the different components of the BPFI, and this is a result of the modulation process. When the rolling elements move over the defect, a modulation process takes place. This results in an irregular but consistent change in the amplitude of the vibration signal resulting in the formation of sidebands around the BPFI at harmonics related to the shaft's frequency. The BPFI can be calculated as follows:

$$BPFI = N\frac{fs}{2}(1 + \frac{d_b}{d_m}cos\phi) \tag{3.12}$$

There is also the fundamental train frequency related to the cage and the ball spin frequency. Bearing component failures also generate these frequencies, making them useful diagnostic tools.

3.4.2 Localised defects

The defect associated with the outer and inner race was modelled as a dent and can be described by its length, L_d, and height, h_d. The total ball deformation is altered in the suggested model whenever the ball and defect come in contact.

To determine changes in total deformation, the properties of the tangent function are utilised in a step function having cubic polynomials and continuous for first derivatives. The step function takes five inputs, and these can be calculated as stated in the equations below (Sopanen and Mikkola 2003b, pp. 16 - 18):

$$g = -|tan[\frac{1}{2}(\beta_j - \theta_{in} - \Phi_{def})]| \tag{3.13}$$

$$g_0 = -|tan(\frac{L_{def}}{4R_{in}})| \tag{3.14}$$

$$h_0 = 0 \tag{3.15}$$

$$g_1 = -|tan(\frac{L_{def}}{8R_{in}})| \tag{3.16}$$

$$h_1 = -h_{def} \tag{3.17}$$

The step function, stated below, takes an independent variable g, as well as initial and final values h_0 and h_1 corresponding to the function and g_0 and g_1 for the step.

$$STEP(g, g_0, h_0, g_1, h_1)$$

$$= \begin{cases} h_0 & g \leq g_0 \\ h_0 + \Delta h \Delta g^2(3 - 2\triangle g) & g_0 < g < g_1 \\ h_1 & g \geq g_1 \end{cases} \tag{3.18}$$

The definition of Δh and Δg can be stated as:

$$\Delta h = h_1 - h_0 \tag{3.19}$$

$$\Delta g = \frac{g - g_0}{g_1 - g_0} \tag{3.20}$$

The ring's rotation angle increases as the experiment progresses, as does the attitude angle of the ball j. It is possible to determine the angle that the defect occupies on the ring from its length and the ring's radius. This is illustrated in Figure 11, where α_1 and α_2 can be computed from the equation below.

$$\begin{aligned} \alpha_1 &= \frac{L_{defect}}{R_{in}}(1 - \frac{\beta_j}{\theta_{in}}) \\ \alpha_2 &= \frac{1}{2}\alpha_1 \end{aligned} \tag{3.21}$$

As seen in the figure, α_1 and α_2 denote the sizes of the angle occupied by the defect.

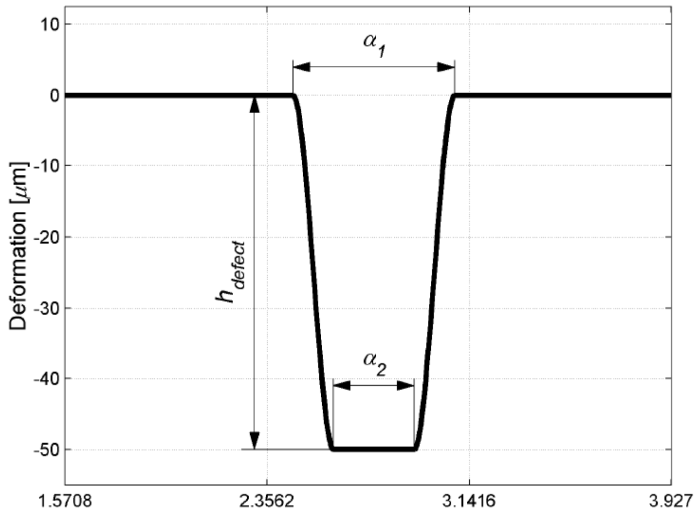


Figure 11: Cage Rotation Angle (Sopanen and Mikkola 2003a, p. 25).

The inner ring rotation, θ_{in}, in equation 3.13 should be replaced with θ_{out} to account for the defective outer ring, and all other equations are left unchanged.

4 TRANSFER LEARNING FRAMEWORK FOR FAULT DIAGNOSIS

In a general supervised machine learning context, it is expected that the distribution and feature space of the data used for training, validation (optional), and testing are the same. But what happens if this assumption fails? Furthermore, deep learning models require sufficient data as a prerequisite and lack thereof will lead to overfitting making the new model ungeneralizable to unseen data. These two concerns are the main motivation for the use and development of transfer learning over the years.

As stated in an earlier section, a previously trained model is the foundation for the implementation of TL. Studies have shown that the target dataset may or may not be related to the source dataset used for training the pre-trained model. A saved model that has been previously trained using a large dataset for a specific task is referred to as a pre-trained model. For the parameter-based approach, TL aims to modify some layers of the pre-trained model so it is applicable to a new, sometimes unrelated, task. The degree of modification depends on how the source dataset is related to the target dataset. Literature suggests that, for computer vision and other image-recognition tasks, CNNs are the most used deep learning models while parameter-based and adversarial methods form the largest share of studies for TL.

Contrary to most TL studies (mechanical engineering and otherwise) associated with CNN that use well-known pre-trained models such as Inception V3, VGG-16, Res-Net, Alex-Net, and MobileNet V2 that have been trained on 100,000s of images, this study uses bearing simulation data to develop a pre-trained model which may be better suited as a foundation for transferring bearing diagnostic knowledge. Thereafter, small samples were extracted from a real experiment to demonstrate the transfer learning process. Simply put, the goal is to use the source model and a small sample of the target to create a new target model that would be difficult to achieve with just the small sample. This is illustrated in Figure 12 below.

Dong, Li, Zheng, Wang, and Xu (2022, p. 336) stressed that, despite having similar features, the distributions of simulation data and real data are not the same for a number of reasons: (1) modeling cannot entirely resolve uncertainty and randomness caused by the environment associated with a real system. (2) It is impossible for a dynamic model to possess all of the features of the real system; instead, the equipment is simplified, with just the most important elements taken into consideration. (3) quality of real signals collected is reliant on sensor type, fixation technique, drift, and structural transmission properties, while vibration signals in dynamic models are simply the solutions of dynamic differential equations. Therefore,

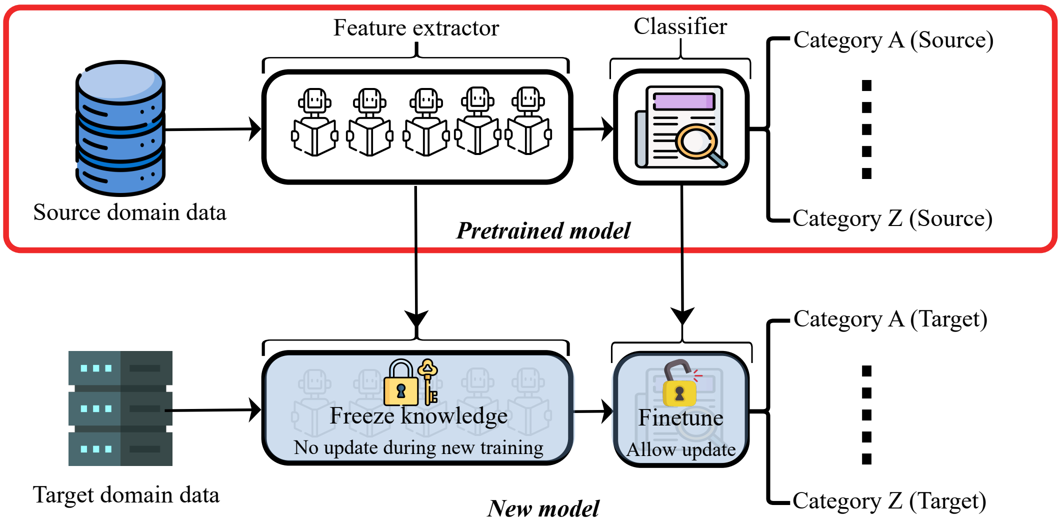


Figure 12: Transfer learning showing fr_ft strategy.

the use of a model trained with simulation data for fault recognition on real data will grossly underperform. This is one of the problems TL aims to resolve.

For clarity, definitions of relevant terms are presented below:

1. Domain: A domain, denoted by $\mathcal{D}$, is comprised of a marginal distribution and feature space denoted by $\mathcal{X}$ and $P(X)$ respectively. Mathematically, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, X represents a set of instances, and $\mathbf{x}_i$, a member of $\mathcal{X}$ denotes each instance . The definition of this instance set is $X = \{\mathbf{x} \mid \mathbf{x}_i \in \mathcal{X}, i = 1, \ldots, n\}$

2. Task: A task, which is denoted by $\mathcal{T}$, has a decision function, f, and label space, $\mathcal{Y}$. Mathematically, $\mathcal{T} = \{\mathcal{Y}, f\}$ where $\mathcal{Y}$ represents possible labels for the selected sample whereas f is a function that needs to be learned from the sample, and subsequently useful for making predictions.

3. Transfer learning: This strategy is useful for enhancing the predictive function f_T $(\cdot)$ associated with a target in its domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$ by leveraging the knowledge derived from $\mathcal{T}_S$ and $\mathcal{D}_T$ which denotes the source task and domain respectively. It is particularly useful when both domains or tasks are not the same, denoted by $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. Also, there may be other constraints such as insufficient or no labels on the target domain (Pan and Q. Yang 2010, p. 1348).

Recall that a general supervised learning context can be stated as $\mathcal{D}_S = \mathcal{D}_T$ and $\mathcal{T}_S = \mathcal{T}_T$. For a transfer learning problem of $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$, we can have two cases for each problem. For this study, the problem falls into $\mathcal{D}_S \neq \mathcal{D}_T$ whereby the probability distribution is not the samae $(P(X)_S \neq P(X)_T)$ unlike the feature space $(\mathcal{X}_S = \mathcal{X}_T)$. This study, therefore, tackles two problems: (1) sample acquisition and (2) domain adaptation, which are common in realistic scenarios.

4.1 Introduction to convolutional neural network

CNNs are widely used for image recognition and classification tasks in several fields not limited to computer vision and natural language processing. They have even surpassed humans in this subject, and showed exceptional performance in certain tasks like ImageNet Large Scale Visual Recognition Challenge (ILSVRC), defeating other AI models (Iman, Arabnia, and Branchinst 2021, p. 82).

For bearing fault identification, there are two primary CNN models: the 2D and the 1D model. Vibration data which are initially 1D can be converted to 2D using several time-frequency methods including wavelet transform, and spectral kurtosis, or can be reshaped to form a 2D matrix. However, this study uses 1D CNN which has shown remarkable performance and has also been applied in similar studies. The primary components of a CNN are the input layer, convolutional layers that convolves with the input, pooling layers that downsamples, and fully connected layers, along with an output that serves as the classifier as illustrated in the figure below.

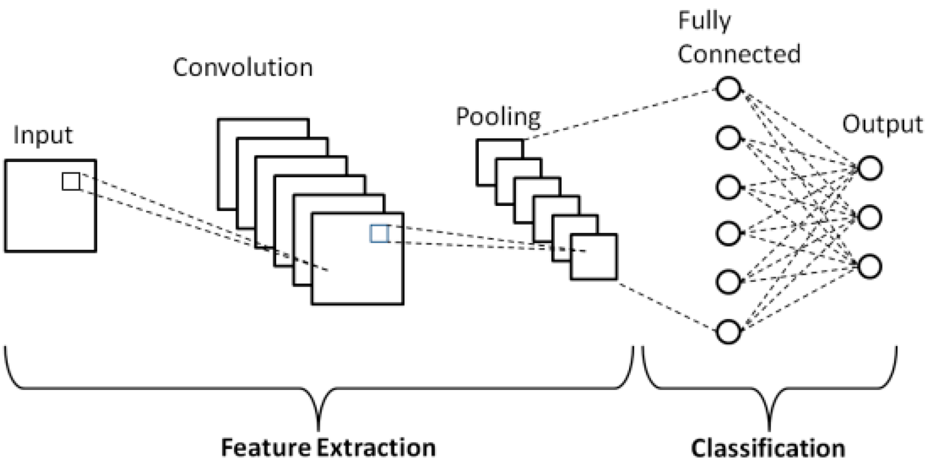


Figure 13: Components of a CNN (Phung and Rhee 2019, p. 3).

In order to create feature maps using the input, the convolution layer(s) apply filter(s) to the input local areas. By utilizing the same weights and biases across all filters, each one captures distinct properties associated with the inputs, this is parameter sharing (Bobylev, Choudhury, Miettinen, Viitala, Kurvinen, and Sopanen 2021, p. 120654). A convolutional layer can employ many filters to extract different local features, which are subsequently activated by the commonly used Rectified Linear Units (ReLU) function which enables increased capability of feature representation and enhanced convergence of CNNs. According to C. Chen, Liu, Yang, Wu, and Ye (2020, p. 3), by using the back-propagation technique for parameter adjustment, ReLU ensures that weights in the shallow layer become more trainable. The following expression may be used to describe a convolutional layer's output:

$$a_k^l = s^{l-1} \cdot w_k^l + b_k^l \tag{4.1}$$

where the weights is represented as w_k^l and b_k^l the bias of the *k-th* filter in layer *l*. The output of *l-1th* layer is represented as s^{l-1}, a_k^l denotes the output of *k-th* filter while $\cdot$ indicates the convolution between the filter and local input regions. Thereafter, the ReLU activation function, which outputs 1 for values greater than 0 and 0 for otherwise, is applied. The expression is stated below:

$$f(x) = \max\{0, y_k^l\} \tag{4.2}$$

By applying the convolutional operation successively across multiple layers, it is possible to grasp the underlying features related to input samples and represent them as feature maps in the output.

The pooling layer's primary function is for dimensionality reduction of the feature maps by summarizing the information in the feature maps while maintaining the important information. It does this by dividing the feature maps into small regions, called pooling regions or pooling windows, and then applying a pooling function, such as max pooling or average pooling, to each region. By contrast, average pooling strikes a balance by averaging the values from each pool, while max pooling takes the highest. The pooling layer's output is a downsampled version of the input feature map, with reduced dimensionality, making it easier to process. Max pooling can be stated generally as:

$$P_k^l(p) = \max_{(p-1)W+1 \le t \le pW} (p_k^{l-1}(t)) \tag{4.3}$$

where W denotes pool's width, $p_k^{l-1}(t)$ is the *t-th* neuron's output in *l-1* layer, $t \in |(p-1)W + 1, jW|$ and $P_k^l(p)$ represents *p-th* output of the pooling operation.

In order to combat overfitting, regularization techniques are used, this happens when the model is overtrained on the training data and cannot perform on unseen data. Ying (2019, pp. 1 - 4) stated that this can happen when the model has too many parameters, making it overly complex and prone to memorizing the training data instead of learning general patterns. Regularization techniques aim to prevent this by adding additional constraints or penalties to the loss function during training. A common method is the use of a dropout which randomly eliminates set percentages from the model. The preceding layer's (typically convolution or pooling layer) flattened output is sent into the fully connected layer. The network typically ends with a softmax activation function, which gives a probability distribution across the labels. This probability distribution can be used to classify the input into one of several categories/labels. Mathematically, this function is:

$$softmax(z(f)) = \frac{e^{z(f)}}{\Sigma_{k=1}^{M} e^{z(f)}} \tag{4.4}$$

where $z(f)$ refers to the *f-th* neuron's output in the output layer, while the total categories are denoted as *M*.

4.2 Pre-train: Using simulated dataset

The method used to generate samples from a simulated dataset, and the basic CNN architecture associated with the pre-trained model is presented in this section. After training and testing, the resulting model is saved, and referred to as a pre-trained network, for reuse in TL.

4.2.1 Sliding frame

The sliding frame method has been used in several studies (R. Zhang, Tao, Wu, and Guan 2017; Hasan, Sohaib, and Kim 2019) to generate samples from vibration data. Frames are employed to capture samples by sliding along the vibration data since they are 1D and often lengthy. This is illustrated in Figure 14 below.

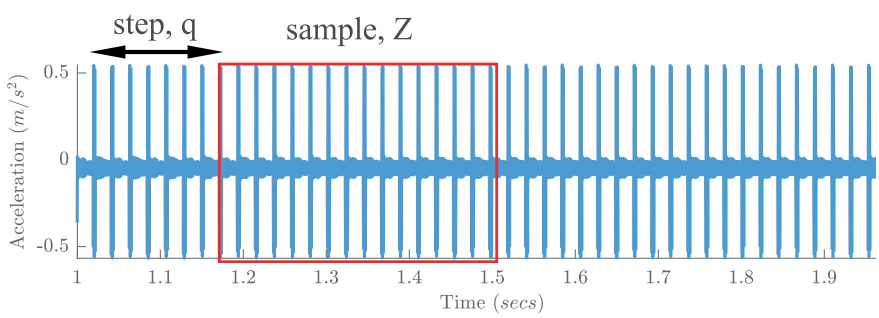


Figure 14: Sliding frame.

The frame's length is represented as Z and corresponds to the total points contained in a sample. The step size, q, is set to be smaller than Z to generate as many samples as needed. The number of samples, n_s, may be calculated using the length of the vibration signal, L_v, as stated below:

$$ns = \frac{L_v - Z}{q} + 1 \tag{4.5}$$

The frame length and step size must be, however, chosen carefully since they have a significant effect on the network's performance. A short frame length may not sufficiently capture data points associated with low frequencies, and a long frame decreases computation speed. On the other hand, small steps may generate identical samples, while insufficient samples may be obtained if large steps are used.

4.2.2 Proposed architecture

The architecture of the CNN is illustrated in Figure 15 and is comprised of one input layer, two convolution and two pooling layers, a flatten layer, one fully connected layer as well as the output layer. To prevent internal co-variate shift and gradient dispersion among neurons within each layer, and stabilization during training, the inputs are first normalized.

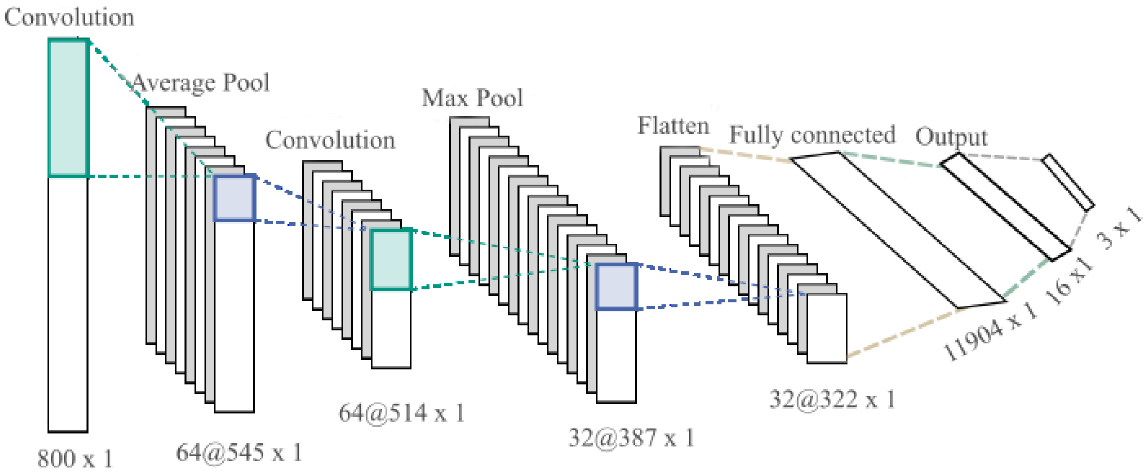


Figure 15: Architecture of the CNN.

The study conducted by C. Chen et al. (2020) revealed that the performance of a CNN increases when the convolution kernel's size decreases progressively. Hence, a reduction of the kernels from 64 to 32. The green area depicts the convolution operation on a sample size of 800 x 1. The convolution layers have a stride of 1 by default, this was not altered. After the first convolution layer, an average pooling layer was added, and then a maximum pooling layer followed the next convolution layer, with a stride of 1, resulting in the feature map size being slightly different from its input. The last pooling layer has a size of 322×1 associated with 32 filters, which is then flattened into a 1D vector before being connected to the output. Further details of the network are presented in the specific case study section.

4.3 Transfer: Parameter-based approach

Model-based or parameter-based approaches are commonly used because they have the capability to address domain adaptation issues arising from the target and source datasets by adjusting the network or model (Iman, Arabnia, and Rasheed 2023, p. 40). In order to minimise the domain distributions between both domains and address the sample acquisition problem, parameters (weights and biases) associated with a pre-trained model are modified, thereby using knowledge from the simulation domain to aid real-world events diagnosis. Parameter-based methods use a variety of strategies, combining pre-trained layers and any other methods such as freezing, fine-tuning, and/or adding new layers (randomization) to make a new model. These strategies are often applied to either the feature extractor, classifier, or both sections of the model. In some cases, the feature extractor's final layers are changed alongside the classifier part, followed by global fine-tuning. Decisions on specific approaches may be subjective inasmuch as the pre-trained parameters are utilized.

A general way of conducting the TL, introduced by Google in 2016, involves freezing the feature extractor part and adding new layers for classification. The initial training is conducted with a moderate learning rate. In the next step, the feature extractor part is unfrozen, and global fine-tuning is conducted using a low learning rate to make the final model. (ibid., p. 40.) However, Dong et al. (2022) recently studied several strategies using different combinations. The influence of some of these strategies on the TL process is explored in this study. These transfer strategies are explained below:

1. Fr: It means freezing the layers of the network it is applied on. Therefore, the parameters of the pre-trained layers associated with the pre-trained models are unchanged in the new network during training.

2. Ft: This means fine-tuning layers under consideration. That is, parameters of the specific pre-trained layers are used for initialization. Consequently, during the training of the new network, the parameters are modified.

3. Ran: Means randomizing the layers of the network. This suggests that the specific layers use random parameters as the starting point. They converge to new values during training.

Since the CNN model comprises two sections viz: (1) feature extractor and (2) classifier, by combining the three different techniques, it is possible to obtain different strategies. A typical example is fr_ran which entails freezing the feature extractor section and randomizing the classifier part. However, the strategies selected for this study are ft_ft, ran_fr, fr_ran, and ft_fr.

5 CASE STUDIES

Two in-depth case studies focusing on the implementation of a bearing dynamic model and analysis of a motor equipped with the bearing, and real-world scenario diagnosis via transfer learning are presented. These case studies aim to provide practical insights and empirical evidence regarding the application and effectiveness of the proposed approaches.

5.1 Case 1: Electric motor with ball bearings

The behavior of defective ball bearing associated with an electric motor has been investigated and implemented in MSC Adams, a commercial multibody software application (Sopanen and Mikkola 2003a; Sopanen and Mikkola 2003b). In the aforementioned study, the software's mathematical properties were utilized through modification of the user interface with the aid of a macro script. However, this study extends the functionality of RoBeDyn to account for the defective conditions associated with the bearings. Thereafter, the results are compared and contrasted with those from MSC Adams.

5.1.1 Structure under investigation

The structure being investigated is shown in Figure 16. The electric motor consists of two ball bearings with model number 6010. The bearing housings served as connection points for the outer rings of Bearings B1 and B2, while both inner rings were attached to the shaft. Spring dampers were used to link the bearing housings to the ground in the X and Y directions, which correspond to the DOF of the housings.

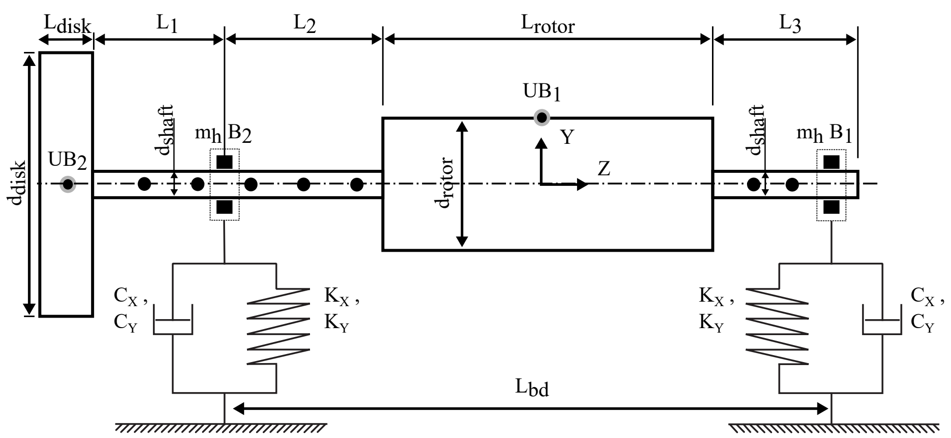


Figure 16: Electric motor under investigation.

In Table 1, detailed descriptions of the dimensions of the electric motor are shown. Translation and rotation in the rotor bearing system axial direction is restricted, while movements in the Y and Z direction are unrestricted.

Table 1: Parameters and dimensions of the motor (Sopanen and Mikkola 2003b, p. 4).

Diameter of the shaft	d_{shaft}	50.0 mm
Diameter of the rotor	d_{rotor}	175.0 mm
Diameter of the disk	d_{disk}	300.0 mm
Length of shaft 1	L_1	105.0 mm
Length of shaft 2	L_2	100.0 mm
Length of shaft 3	L_3	125.0 mm
Length of the rotor	L_{rotor}	300.0 mm
Width of the disk	L_{disk}	45.0 mm
Bearing distance	L_{bd}	500.0 mm
Horizontal support stiffness	K_X	100 kN/mm
Vertical support stiffness	K_Y	100 kN/mm
Horizontal support damping	C_X	10.0 Ns/mm
Vertical support damping	C_Y	10.0 Ns/mm
Mass of the bearing housing	m_h	3.17 kg
Mass of the disk	m_{disk}	20.0 kg
Rotational inertias of disk	I_{xx}, I_{yy}, I_{zz}	{1.1, 1.1, 2.0} kg
Gravity constant (-Y)	-	9.80665 m/s^2

There are a total of 60 DOFs in the system, distributed between 16 nodes and their associated four DOFs. The steel used to develop the rotor and the shaft has a Young's modulus, E, of 206000 MPa and a Poisson's ratio, υ, of 0.3 . The rotor rotates on the Z-axis, and the model implementation in RoBeDyn is presented in Figure 17.

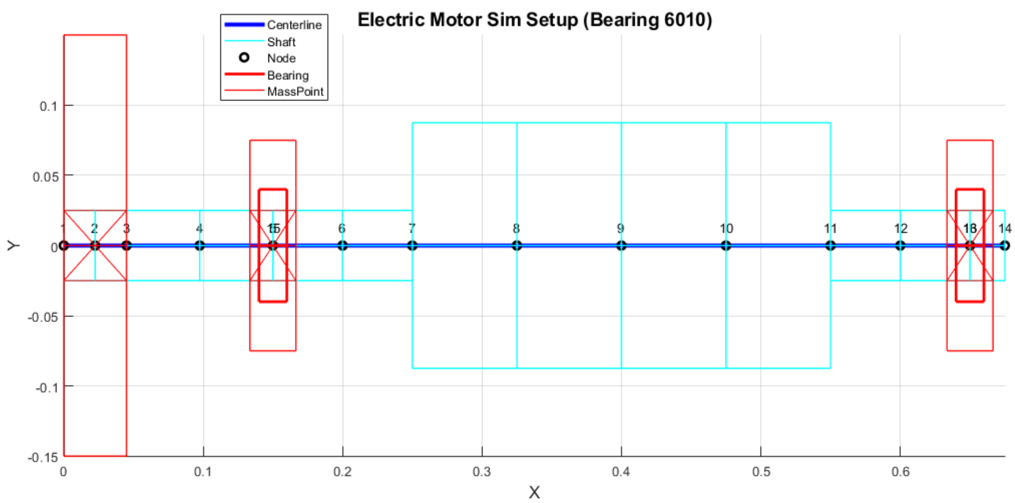


Figure 17: 2D Wireframe plot of electric motor, RoBeDyn.

The bearing housing was modeled as mass points at respective locations. Its inertia is set to zero on all three axes. Table 2 summarizes the bearing's properties. Only the diametral clearance, C2, corresponding to 5.5 µm was used for the study of localized defects.

Table 2: Dimensions of ball bearing 6010 (Sopanen and Mikkola 2003b, p. 4).

Bore diameter	d_B	50.0 mm
Outer diameter	d_o	80.0 mm
Pitch diameter	d_m	65.0 mm
Ball diameter	d_b	8.73 mm
Number of balls	N	14
Diametral clearance, C2	c_d	5.5 µm
Diametral clearance, C5	c_d	59.0 µm
Bearing damping coefficient	c_b	0.55 Ns/mm
Inner and outer race conformity	R_r	0.52
Static load rating	C_0	16000 N
Modulus of elasticity	E	206000 MPa
Poisson's ratio	υ	0.3
Viscosity parameter	μ	0.023 mm^2/N
Viscosity parameter	η	0.04e-6 Ns/mm^2

Similar to the earlier study, the simulation was run at a speed of 900 rpm and 2 kHz sampling frequency for 9 secs. Defects of length 2 mm and height 50 µm having position angle, Φ_{def}, of $0°$ and $270°$ corresponding to the outer and inner races were simulated. Unlike the previous study that used the Gear stiff numerical integrator, this one uses the Ode23 time integrator scheme available within the Matlab environment with a timestep of 2e-4 secs and an error tolerance of 1e-4. Thereafter, signal processing techniques such as Hanning window, FFT, and Hanning correction were applied to the last 8.192 seconds of the resulting displacements. The initial transient vibration associated with the first 0.808 secs was not considered.

5.1.2 Results and analyses

The predicted vertical and horizontal displacements are presented in Figures 18 and 19 respectively. In order to expose the components of the respective displacements, a short duration was extracted from the 9 secs vibration. Figure 18 shows that the healthy bearing has a low amplitude. The absolute peak-to-peak amplitude of the healthy bearing is 2 µm whereas the amplitude of the defective counterpart is about 15 µm which is 8 times the healthy condition. Furthermore, the plot shows that the defective outer race has more components if contrasted with the defective inner race. More components or peaks indicate that the defective outer race is generating more pronounced and frequent vibrations or disturbances in the bearing system. Generally, this could be manifestations of defect size, the severity of the defect, defect position, or different dynamics associated with the specific race.

Figure 19 shows the horizontal displacements of all three conditions - healthy and thedefective races. The average amplitude is lower than the vibrational response on the vertical side. Similar to the vertical displacements, the healthy bearing shows distinct features from the defective counterpart.

a.
b.
c.

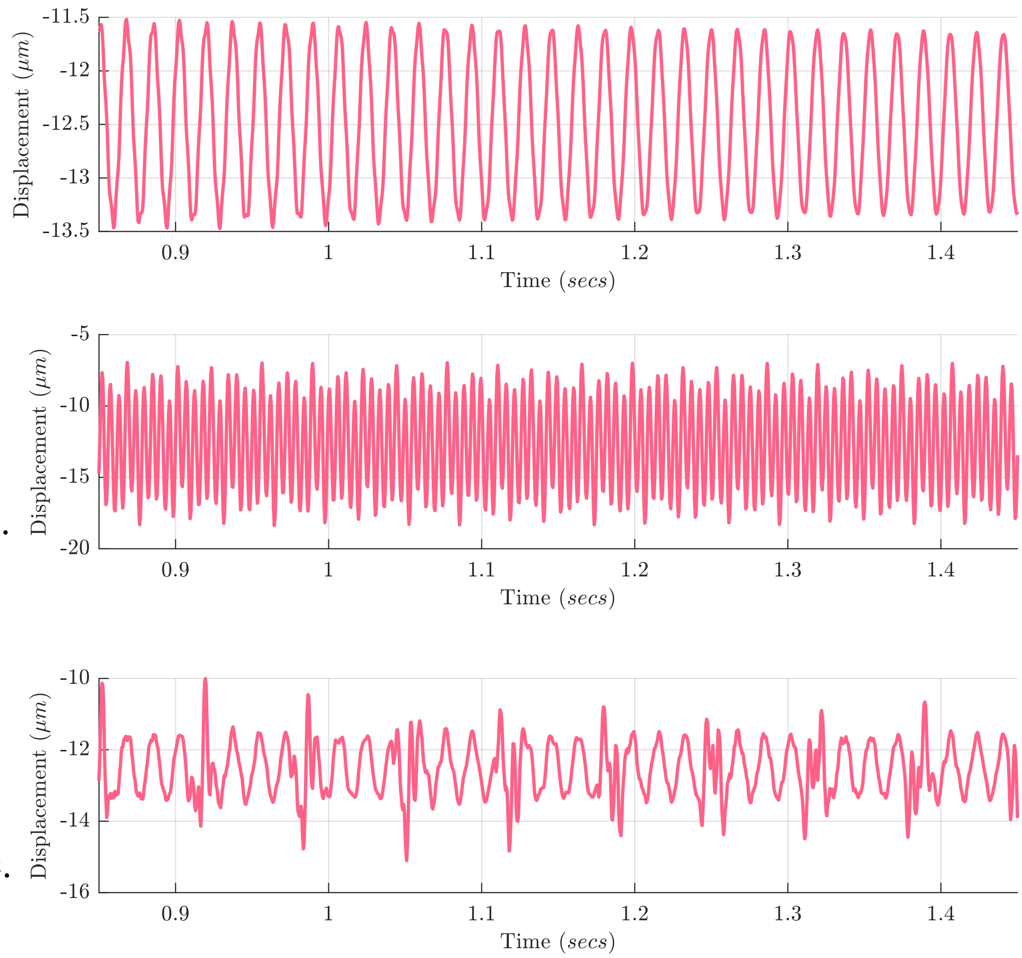


Figure 18: Vertical displacements: a. healthy; b. defective outer race; c. defective inner race.

Unlike the vertical displacements associated with defective races which are different on observation, the horizontal displacements have prominent similar features. Also, it may not be easy at first glance to distinguish which of them has more vibrational response components.

The frequency response, which exposes important system characteristics such as the rotating frequency, FTF, BPFI, and BPFO is further presented. These frequencies relevant for bearing condition monitoring are (1) shaft frequency = 15 Hz; (2) cage frequency = 40.79 Hz; (3) ball pass of the outer race = 90.90 Hz; and (4) ball pass of the inner race = 119.10 Hz.

Figures 20 - 23 show the vertical and horizontal displacements associated with the defective outer and inner race from MSC Adams and RoBeDyn placed side by side.

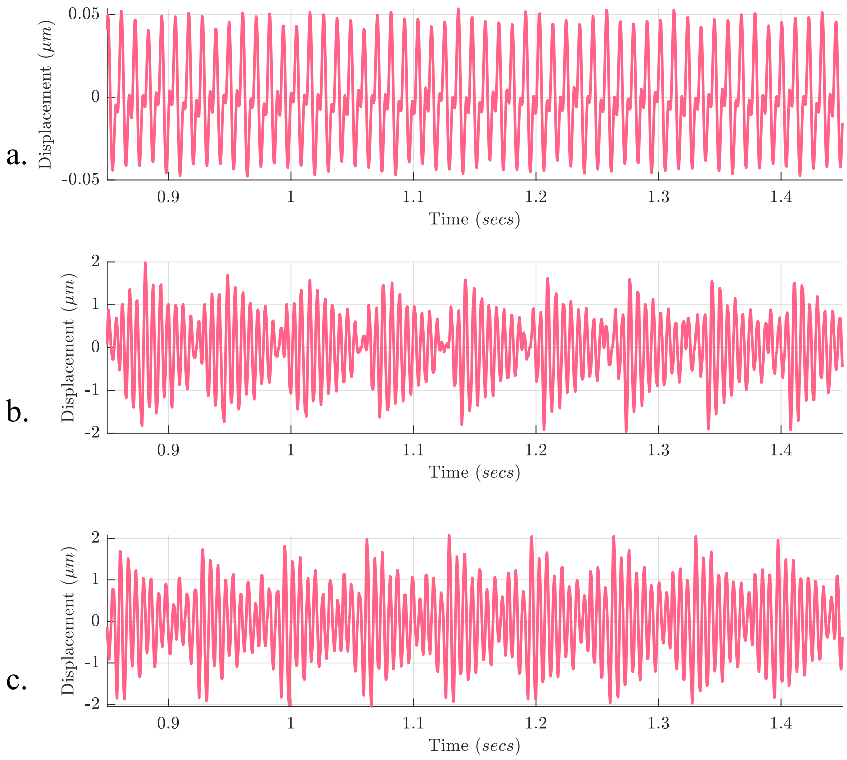
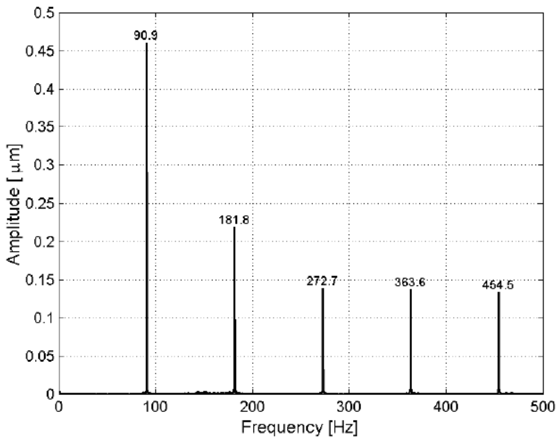
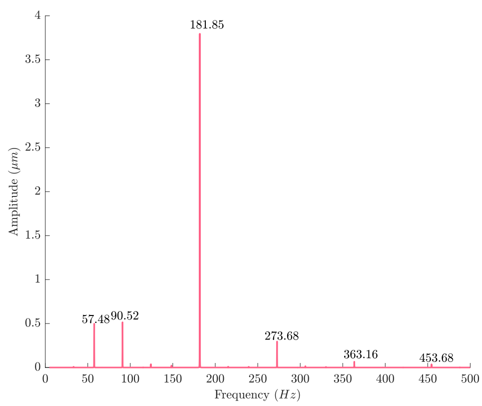


Figure 19: Horizontal displacements: a. healthy; b. defective outer race; c. defective inner race.



(a) Vertical displacements, Adams (Sopanen and Mikkola 2003b).

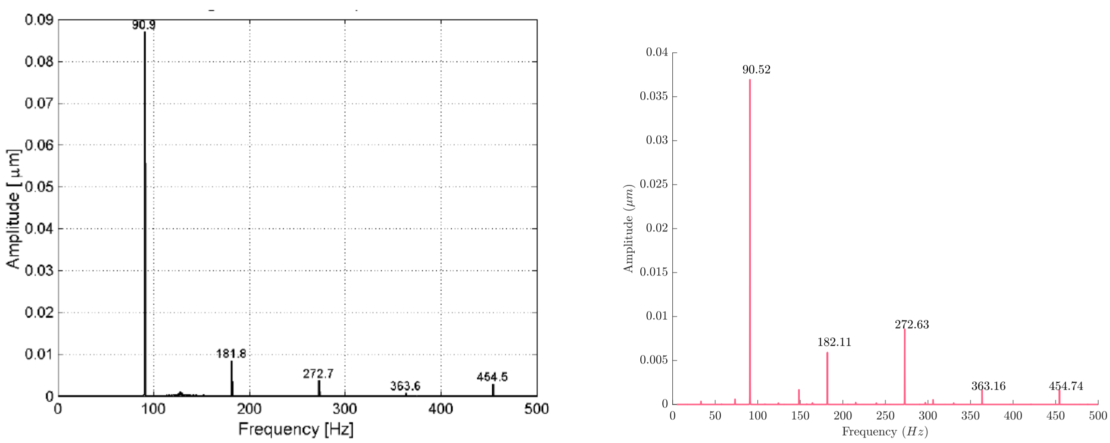


(b) Vertical displacements, RoBeDyn.

Figure 20: Spectra of defective outer race - vertical displacements.

The disparity in the amplitudes obtained from MSC Adams and RoBeDyn is evident at first glance which may not be unconnected with the different solvers used for the simulation.

The outer race spectra show the BPFO as well as its harmonics which are often observed in experimental studies. However, the second harmonic in the vertical displacement is the

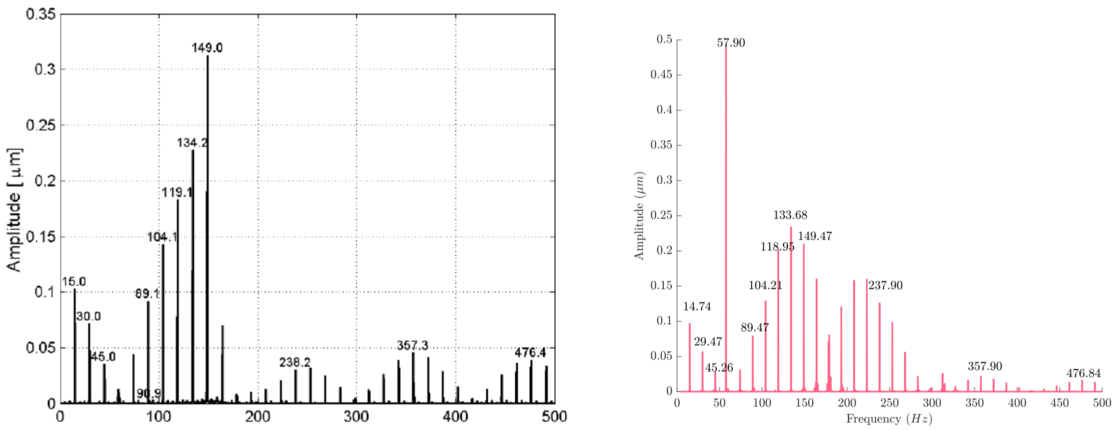


(a) Horizontal displacements, Adams (Sopanen and Mikkola 2003b).

(b) Horizontal displacements, RoBeDyn.

Figure 21: Spectra of defective outer race - horizontal displacements.

most prominent, unlike its counterpart from Adams where the first harmonic had the highest amplitude. The conclusions drawn by Sopanen and Mikkola (2003b) are observable, such that the sideband frequencies are not represented as well as negligible excitations emanating from the horizontal response. Also, the vertical displacement plot shows a frequency that is 4 times the frequency of rotation, which may be due to a natural frequency.

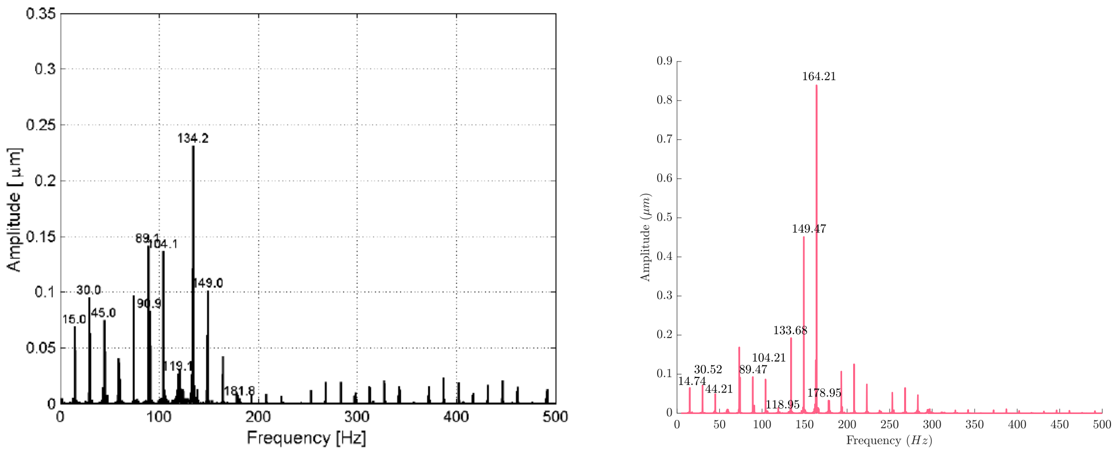


(a) Vertical displacements, Adams (Sopanen and Mikkola 2003b).

(b) Vertical displacements, RoBeDyn.

Figure 22: Spectra of defective inner race - vertical displacements.

The spectra associated with the inner race clearly show the rotation speed, BPFI, and their respective multiples on the results from Adams and RoBeDyn. Again, the frequency corresponding to 4 times the rotating frequency having a max amplitude of $0.83\,\mu\mathrm{m}$ can be observed in the vertical displacement plot. Unlike Adams, where the maximum amplitude was obtained at 134 Hz, the maximum amplitude in the horizontal displacement plot was obtained at 164 Hz. Overall, the result from RoBeDyn appears to agree with those from Adams despite a few inconsistencies not unconnected to the use of different integration schemes.



(a) Horizontal displacement Adams (Sopanen and Mikkola 2003b).

(b) Horizontal displacement RoBeDyn.

Figure 23: Spectra of defective inner race - horizontal displacements.

5.2 Case 2: Diagnosis scenario from simulation to real under similar condition

In this section, the transfer of diagnosis knowledge from simulation to real given an insufficient sample is demonstrated. Data from both domains are related but not identical, however, there is some basis for the transfer. Firstly, the simulated data are derived from the dynamic model whose formulations are established according to the principles governing the real-world scenario. Furthermore, geometrical parameters such as sampling frequency, shaft speed, and defect description are the same. Several transfer strategies are considered in order to effectively transfer diagnosis knowledge thereby eliminating the differences in the domain.

5.2.1 Dataset description

Real dataset: Paderborn dataset was provided freely online by the Faculty of Mechanical Engineering, Paderborn to foster research and enable collaboration in bearing condition monitoring (Lessmeier, Kimotho, Zimmer, and Sextro 2016). The dataset is comprised of 32 deep groove bearings of type 6203 which can be grouped into 12 artificially damaged, 6 undamaged (healthy), and 14 real damages resulting from accelerated lifetime tests. The test rig used in conducting the experiment and presented in Figure 24 had current and vibration signals sampled from it at 64 kHz. Although these bearings were operated under varying conditions, the specific conditions selected for this study are (1) 0.7 Nm load torque (T), (2) 900 rpm speed, (n), and (3) 1000 N radial force (F). The respective folders (bearing names) associated with the 32 bearing datasets contain 20 Matlab files each. The folders (bearing names) used for this study are K001, KA01, and KI01 corresponding to the normal condition, defective outer and inner races respectively.

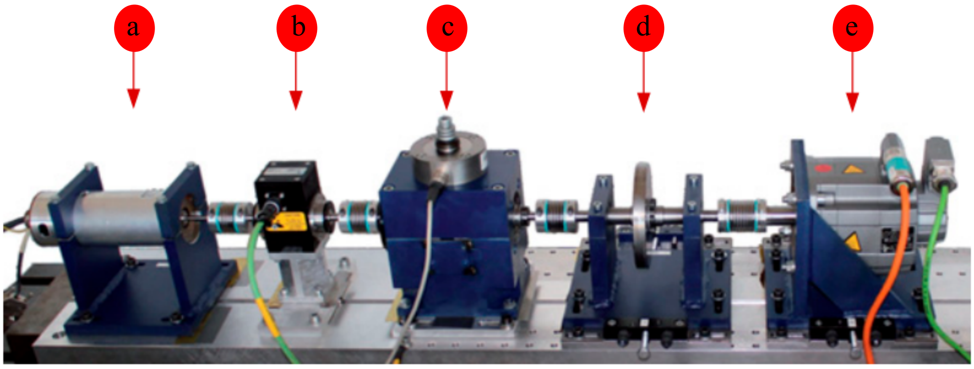


Figure 24: Paderborn experimental test rig setup: (a) test motor; (b) measuring shaft; (c) bearing module; (d) flywheel; (e) load motor.

Electric discharge machining (EDM), manual electric engraving, and drilling were the three techniques employed to create artificial damage. Based on the fault description, defects caused by EDM on balls with a depth of 1-2 mm and length of 0.25 mm were selected. While the total samples provided by Paderborn University may be sufficient for training a diagnosis model, small samples have been purposely extracted to demonstrate the TL process. Also, during the implementation, the sampling frequency was downsampled to about 12 kHz because simulating the source dataset at the initial 64 kHz would reduce computation speed, and generate large file sizes.

Simulated dataset: For diagnosis scenarios under similar working conditions, the simulated dataset needs to be modelled after the real dataset. The setup in RoBeDyn is presented in Figure 25 below. The setup in RoBeDyn does not include the flywheel used during the experimental study.

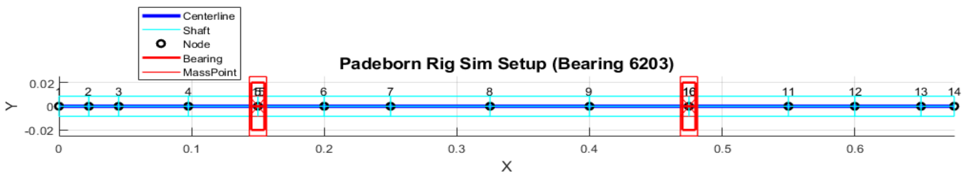


Figure 25: Paderborn rig setup in RoBeDyn.

The simulation was conducted using the parameters presented in Table 3, at 12 kHz sampling frequency for a timeframe of 200 secs. The three bearing conditions (i.e., normal and defective races) were simulated using fault description as the real dataset.

The displacements obtained from RoBeDyn were differentiated accordingly to acquire the acceleration signals for all conditions.

Table 3: Dimensions and parameters for Paderborn rig simulation

Bore diameter	d_B	17.0 mm
Outer diameter	d_o	40.0 mm
Pitch diameter	d_m	29.05 mm
Ball diameter	d_b	6.75 mm
Number of balls	N	8
Diametral clearance, C2*	c_d	5.5 µm
Bearing damping coefficient*	c_b	0.55 Ns/mm
Inner and outer race conformity	R_r	0.52
Static load rating	C_0	16000 N
Modulus of elasticity	E	206000 MPa
Poisson's ratio	υ	0.3
Viscosity parameter	μ	0.023 mm^2/N
Viscosity parameter	η	0.04e-6 Ns/mm^2
Width of the fault	L_d	0.25 mm
Depth of the fault	h_d	1 mm
Gravity constant (-Y)	-	9.80665 m/s^2
Horizontal support stiffness*	K_X	100 N/mm
Vertical support stiffness*	K_Y	100 N/mm
Horizontal support damping*	C_X	0.1 Ns/mm
Vertical support damping*	C_Y	0.1 Ns/mm
Mass of the bearing housing*	m_h	1.17 kg
Length of shaft*	L	675 mm
Distance between bearings*	L	345 mm

Note: Dimensions and parameters marked with * are assumed values.

5.2.2 Implementation of the CNN - pre-train, and transfer learning

The detailed architecture of the network used in making the pre-trained model, which was subsequently modified in the TL cases is presented in Table 4. All CNN experiments were implemented using the TensorFlow framework running on NVIDIA Quadro K2200.

Table 4: Details of the CNN architecture

Layer(type)	Filters	Kernel/Pool	Stride	Output shape
Conv1D	64	256 x 1	1	(545, 64)
AveragePooling1D	-	32 x 1	1	(514, 64)
Conv1D	32	128 x 1	1	(387, 32)
MaxPooling1D	-	16	1	(322, 32)
Flatten	-	-	-	(None, 11904)
Dense (fully connected)	-	-	-	(None, 16)
Dense (Classifier)	-	-	-	(None, 3)

Figures 26 through 28 illustrate similarities and differences between simulation and real signals for three distinct conditions: normal, defective inner race, and defective outer race.

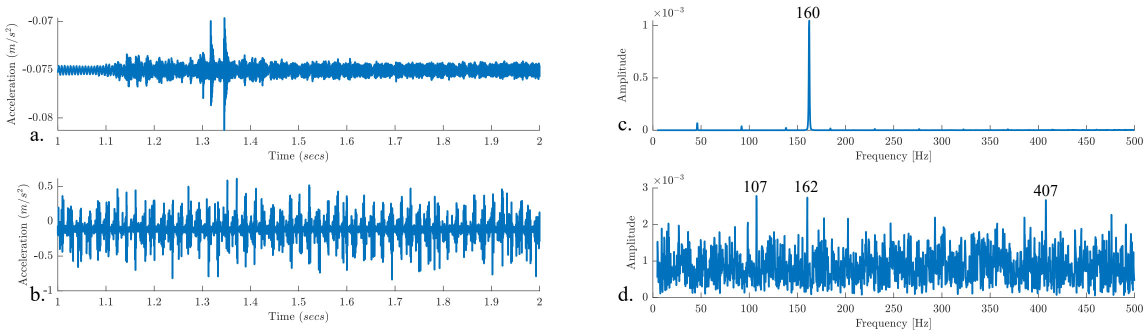


Figure 26: Simulation and real signal under the normal bearing condition: (a) simulated signal's time domain; (b) real signal's time domain; (c) simulated signal's spectrum; (d) real signal's spectrum.

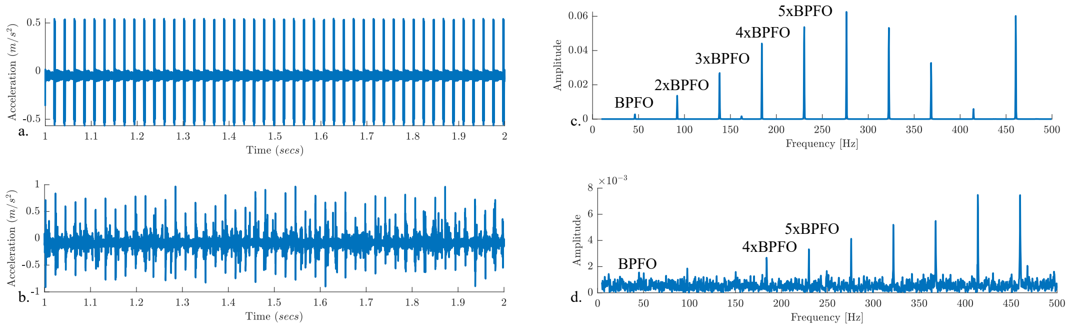


Figure 27: Simulation and real signal under defective outer race condition: (a) simulated signal's time domain; (b) real signal's time domain; (c) simulated signal's spectrum; (d) real signal's spectrum.

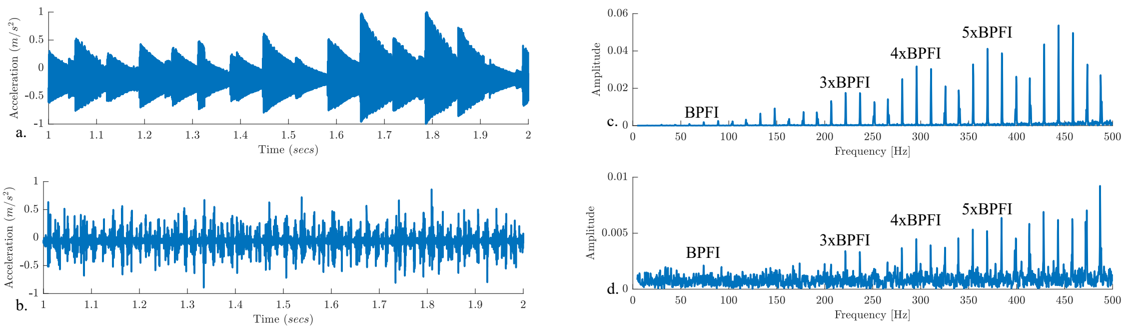


Figure 28: Simulation and real signal under defective inner race condition: (a) simulated signal's time domain; (b) real signal's time domain; (c) simulated signal's spectrum; (d) real signal's spectrum.

Signals corresponding to 1 second and < 500 Hz frequency spectrum range are presented in the figures to ensure clarity in the comparison. For reasons earlier stated in previous sections, the distribution of the data is not the same. Therefore, the model associated with one cannot deliver guaranteed performance on the other without TL. The normalized plots show that the real signals have higher amplitudes when contrasted with the simulation signals. Furthermore, the influence of noise, the effect of signal extraction techniques, and other environmental and motor factors are clearly visible in the real signals which are often not considered in dynamic models. In some studies, gaussian noise has been added to the dynamic models to account for such an effect. The spectra clearly show fundamental frequencies and their harmonics when both domains are compared.

A trained model using the simulated dataset, referred to as a pre-trained model, was first developed using a sample size of 800 and a step of 400. For a total of 3333 samples, the training, validation, and testing datasets were obtained by splitting the total samples in a ratio of 7:2:1. The parameters of the aforementioned network were randomly initialised, and the necessary hyperparameters were set as follows: (1) batch size of 64; (2) optimizer = stochastic gradient descent with a momentum equal to 0.9; (3) a learning rate of 0.001; and (4) epochs = 30. The categorical cross-entropy loss function was utilized since it is a classification task. After successful confirmation of the effectiveness of the pre-trained model by using the test set, the TL was conducted. The small samples obtained from the real dataset, which is often the case in real life, would not be sufficient for making a diagnosis model. Depending on the hyperparameter settings, the consequence is overfitting, and the learning curve may appear haphazard as shown in Figure 29 where the highest accuracy obtained was 50%.

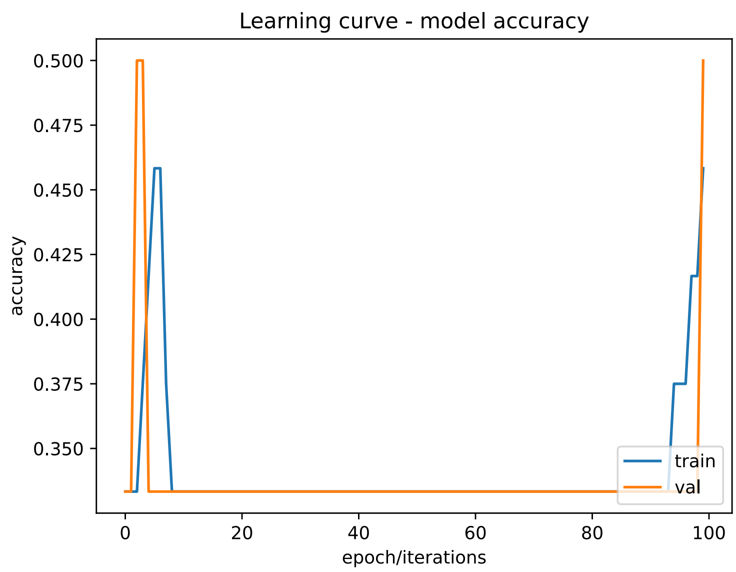


Figure 29: Accuracy curve of bad model.

Presented with such a problem, TL attempts to use the knowledge gained in the pre-train as leverage for real scenario (small sample) diagnosis. Several transfer strategies were conducted, and Table 5 gives a summary of the trainable and non-trainable parameters associated with the pretrain and TL strategies.

Table 5: Summary of the parameters associated with the CNN

CNN mode	Trainable parameters	Non-trainable parameters	Total parameters
Pretrain	469,155	0	469,155
Fr_Ran	190,531	278,624	469,155
Ft_Ft	469,155	0	469,155
Ft_Fr	278,624	190,531	469,155
Ran_Fr	278,624	190,531	469,155

During the pre-train, all parameters (weights and biases) were trained from scratch, however, only certain parameters are trained during the transfer-learning process. For this, the hyper-parameter settings were adjusted as follows: (1) batch size = 32 (default value); (2) optimizer = stochastic gradient descent, momentum equal to 0.9; (3) learning rate = 0.001; (4) epochs = 100. The default batch size was used in the pre-train since the sample is small and no improvement was observed for smaller sizes. In accordance with the corresponding transfer techniques, the feature extractor and classifier were changed.

5.2.3 Diagnosis results and analyses

Learning curves: A typical learning curve displays the model's performance (accuracy or error) on the vertical axis while the horizontal axis shows the number of times the model has run over the whole training dataset (referred to as epochs). The model is first exposed to the training dataset during the training phase, and then its performance is tested using a separate dataset (the validation dataset) that is not exposed to the model during training. This evaluation is typically done after each epoch or a certain number of iterations. The validation dataset serves as a proxy for unseen data and helps estimates the model's generalizability to new, unseen data. The learning curve associated with the pre-train is presented in Figure 30.

The learning curve helps to diagnose conditions of overfit, underfit, or a good fit, with the latter being desirable. Simply put, overfit suggests that the model has memorised the training dataset too well instead of understanding the underlying patterns leading to poor performance during testing, while an underfit is used to represent a model that can neither model the train data nor perform well on test data due to simplicity of the model or small observations. A good fit refers to a model that adequately captures patterns in the training data, and also recognize them in unseen data. The train and validation accuracy rose steadily until reaching 100

The model's accuracy as it progresses through training is shown by the accuracy curve. This tells us how effectively the model is categorizing or predicting the target variable. Higher values on the accuracy curve indicate better predictive performance, while a plateau or fluctuations may indicate the need for adjustments, such as model complexity or hyperparameter tuning. On the other hand, loss quantifies the error or disparity between the forecasted and actual values. During training, the objective is to reduce the loss, as lower values indicate better alignment between predicted and actual values.

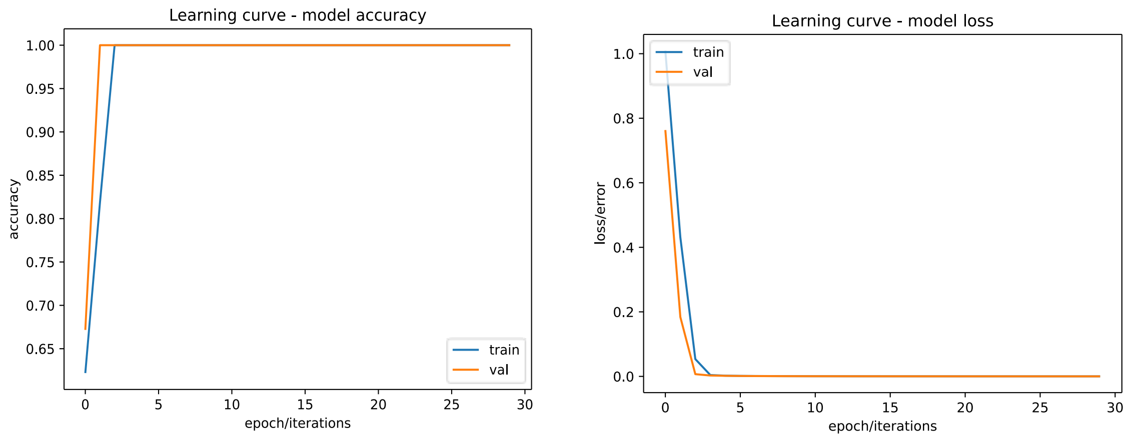


Figure 30: Pretrained model - learning curves.

In the same vein, the effectiveness of considered transfer strategies is presented in Figure 31. In terms of accuracy, Ran_Fr had a smooth and faster rise to 100% accuracy than all other strategies. Fr_Ran had 84% accuracy which is the lowest, and can consequently predict faults associated with the target at this accuracy. Both Ft_Fr and Ft_Ft had unclear movements in the beginning before converging to 100% which may be attributed to the technique applied to the feature extractor part, unlike the other two strategies.

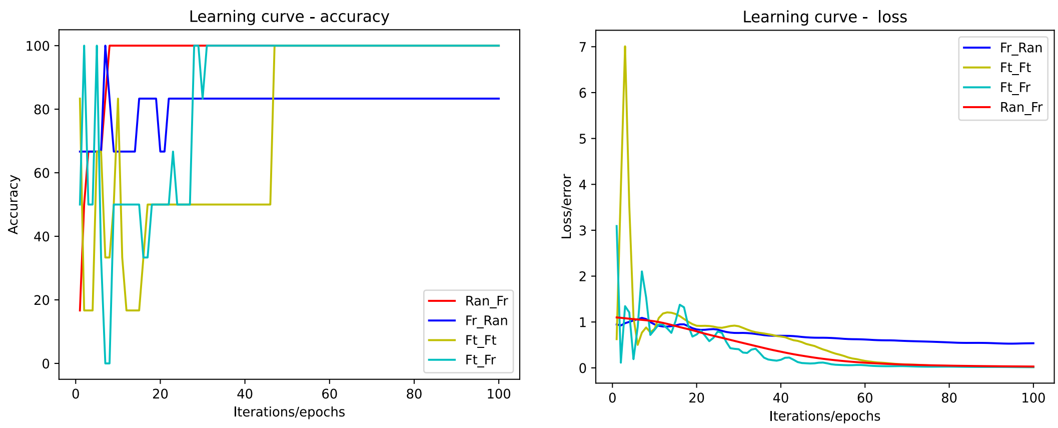


Figure 31: Transfer strategies - learning curves.

On losses, Ft_Fr and Ft_Ft had the highest error during the start of the training before reduction to 0. It can be seen that the loss curve associated with Fr_Ran does not tend to 0 which will eventually affect its generalizability during testing. After 60 epochs, all other strategies had attained zero error which can be traced to 100% accuracy in the accuracy curves.

Confusion matrix: The confusion matrix provides a comprehensive summary of the predictions made by a model, enabling a detailed analysis of the classification results. It comprises a tabular representation that contrasts the test data's actual labels with the predicted labels produced by a model. It provides valuable insights into the model's ability to correctly classify instances into different classes or categories. The strengths and drawbacks of the model may be better understood by an analysis of the confusion matrix's generated performance indicators.

The result obtained when the pre-trained model was tested on the test data is given in Figure 32 where the model attained 100% prediction accuracy of the bearing condition associated with the simulated dataset.

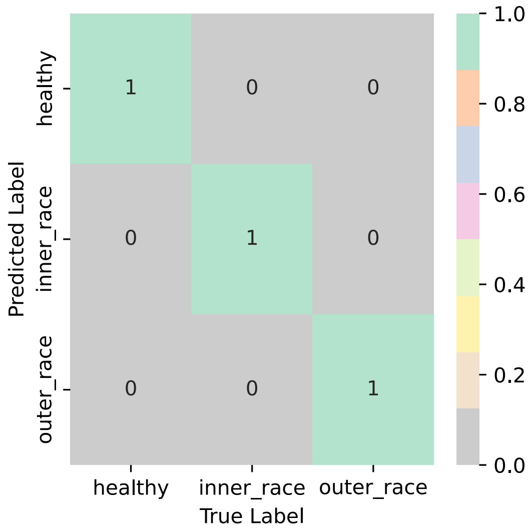


Figure 32: Confusion matrix of pretrained model.

A confusion matrix takes its form according to the number of categories being used in the analysis. It comprises rows and columns, where each row represents the predicted classes and each column the actual. The cells of the confusion matrix contain the counts or frequencies of instances that belong to specific combinations of predicted and true classes. The accuracy of the prediction is shown in the diagonal of the matrix. Important performance indicators including accuracy, recall, and F1-score may also be generated from the confusion matrix. These metrics provide valuable insights into the model's overall classification accuracy, its ability to correctly identify positive or negative instances and the trade-off required between precision and recall.

Furthermore, the pre-trained model was used as a base to develop a new model for a real-world case given insufficient samples. The figure below shows how well the model worked with data from the real world.

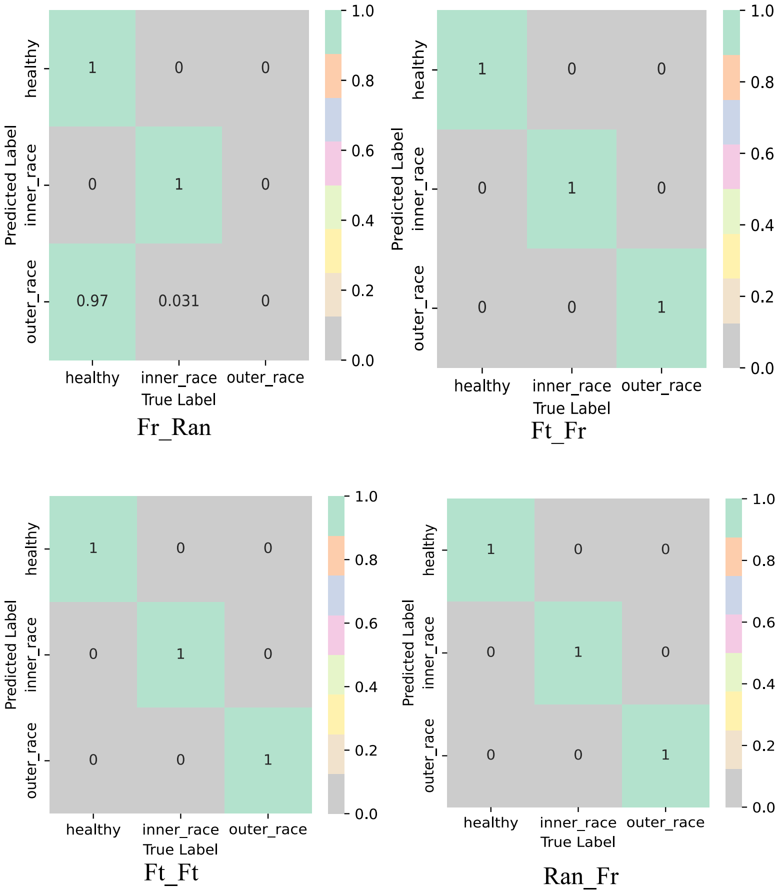


Figure 33: Confusion matrix of several transfer learning models.

The prediction result shows a high misclassification in the case of Fr_Ran where 97% of the outer race test dataset was predicted as healthy and 3% as defective inner race. This is undesirable and can be attributed to the size of the samples used for training. Other strategies showed an excellent prediction of the labels/categories the test dataset belongs.

6 DISCUSSION

The results and analyses have been presented in the case study chapter under respective sub-headings. In this chapter, they are put into perspective by relating them to the objectives of the study, literature reviewed and, to a large extent, the field of IFD thereby acknowledging the research study's contributions and limitations .

6.1 Overview

This study used a simulated dataset from a rotor-bearing model to develop a diagnosis model, and subsequently transfer the knowledge learned to a target model using limited samples. We found that the results, for case one, are consistent with experimental results and those from previous studies. Furthermore, case two's results revealed that the model can accurately recognize the conditions of the bearings for at least three transfer strategies evaluated. Practically, this means that three of the models can be deployed for real-time bearing condition monitoring without (or reduced) the need for experts or domain-specific knowledge.

Contrary to common TL using a parameter transfer routine that starts with Fr_Ran followed by Ft_Ft, this study splits the strategies to evaluate their effectiveness leading to three positive transfers and one negative transfer. Since the model can intelligently classify the bearing conditions with 100% accuracy at the source level (see Figure 32), a negative transfer in this context can be stated as a reduction in the performance of the new model, whereas a positive transfer at the target leads to an equal or increased accuracy. The negative transfer of Fr_Ran which shows a high misclassification in diagnosing outer race fault may imply low adaptation, a need to tweak the architecture or hyperparameters, or insufficient target samples for generalization. From the literature, an arbitrary amount of samples are used for the target, so the minimum samples for effective transfer are still unclear. Also, the plateau observed in the learning curves may suggest that the model has reached a saturation point or a case of the plateau phenomenon where the model may appear to have reached saturation. Although the confusion matrix's result may not suggest this phenomenon, it may be worthwhile to train for more epochs to ascertain the accuracy.

Furthermore, selecting the ideal hyperparameter settings for training is crucial for the model performance as the learning curves and resulting prediction can quickly change with a slight variation of the settings. Suitable hyperparameter settings can be obtained by either a manual trial and error process or a systematic and automated process using available libraries (e.g. using GridSearchCv available in Python's SciKitLearn). The hyperparameters used were manually selected which might suggest the need for further investigation. The default batch size of 32 was used in the transfer learning process since lower values did not show significant performance. Typically, the main purpose of setting the batch size lower than the available samples is to increase the speed of convergence. Therefore, for small samples, speed may not be of paramount importance.

The results from the TL model should be evaluated and compared with the results obtained from Linear Regression, ANN, SVM and so on. The purpose is to assess the performance and effectiveness of the model against these established methods.

6.2 Contributions

It is pertinent to state that the methodology for carrying out CNN-based transfer learning associated with supervised learning (i.e., labelled datasets in both domains) for bearings can take three forms, viz: (1) Using a pre-trained model (e.g., VGG-16, ResNet, and Alexet) and a target dataset; (2) Using a source dataset from real equipment to train a model (later referred to as pre-trained model) and a target dataset from another real equipment; and (3) Using a simulated dataset from a model similar to real equipment to train a model (later referred to as pre-trained model) and a target dataset from real equipment.

Most of the studies from our review show that the first form is the most studied, and the third form which is the main theme of this study is the least, in fact, less than three papers were identified. The dearth of simulation-based transfer learning using CNN justifies the importance of this research study. The key contribution of this study is the provision of a framework for intelligent fault diagnosis of bearings in scenarios where there are few samples. If the size of the target samples used in this study is juxtaposed with samples from most CNN-based transfer learning studies, the relevance and findings of this research can be properly substantiated.

6.3 Limitations

There are two main limitations of the research study. First, the diagnosis model was trained using vibration data from ball bearings running at 900 revs/min subjected under a single load. Therefore, the performance of the model cannot show satisfactory results if deployed for monitoring bearings installed on different equipment or bearings operating at different working conditions. If the model is trained using data from several equipment and various working conditions, then, the effectiveness of the model for health recognition tasks would increase. However, the proposed framework is a case study for bearing diagnosis, it can be applied to other machines including gearboxes.

Furthermore, if the architecture used is compared to popular pre-trained models like ResNet, VGG-16 or MobileNet, the simplicity is obvious. These pre-trained models use several convolution and pooling layers in the feature extractor section resulting in a more robust model. Since the convolution layers at the beginning of a CNN learn generic features while increasing the specificity as the network deepens, the structure of the architecture could be improved.

In spite of these limitations, which provide opportunities for researchers to build upon, the study gives a strong argument that supports the transfer of diagnosis knowledge from simulation to the physical scenario.

7 CONCLUSIONS

In this master's thesis, a diagnosis framework whose base is reliant on a dynamic model and transfer learning, capable of resolving the difficulty associated with sample acquisition in the course of model training as well as domain discrepancy common in real-world scenarios is presented. This work uses a dynamic model earlier proposed for simulating the vibration signals. The implementation of the model had earlier been achieved in MSC Adams, however, this study extended the functionality of a rotor-bearing library in Matlab (RoBeDyn) developed at LUT University to account for the localized defects presented in the aforementioned study. Furthermore, the parameters of the bearing and other necessary parameters in case 1 were modified to match an experimental test rig setup for case 2. Thereafter, simulation data obtained was used to demonstrate the transfer of knowledge to the real-world when either common challenges such as sample acquisition may arise or the need for less computation time by leveraging the knowledge already learned in the pre-trained.

The results from case 1, except for some disparity, are in agreement with those obtained from Adams. This may be due to the different solvers used for both studies. Also, unlike the study conducted by Sopanen and Mikkola (2003b), this study did not obtain the results of both large and small defects which may provide a better viewpoint for comparing and contrasting results from both studies. Only defect size corresponding to large defects was considered in this study. The study of how unbalance and clearance affect the system could give more information about how the RoBeDyn rotor-bearing system works.

Diagnosis results from case 2 showed that the resulting model is effective for predicting the condition of the bearing despite some differences in the data distribution from both domains. Since some essential parameters such as damping and masses of bearing housing used were not provided along with the Paderborn dataset, the differences seen in both frequency and time-domain figures may not be a complete reflection of the results. Generally, the resulting diagnosis model cannot be a silver bullet for all bearing working conditions and equipment because their applicability and performance are dependent on the dataset used for training. Therefore, the use of data corresponding to several working conditions and equipment would make a more complete model. Also, the sample size used is small when contrasted with similar CNN-based transfer learning studies, the ideal sample size to use for successful generalizability of the new model is still unclear from the literature reviewed. Hence, a study conducted in this regard will be beneficial for future studies. It should be noted that the results of models are very sensitive to hyperparameters and to some extent computing resources used in the training.

The objective of the study stated in the introduction section which cuts across defective bearing diagnosis from simulation to the physical domain has been demonstrated. The results obtained can be useful for future research and development in this domain as it provides insights into the behaviour of bearing conditions as well as fault recognition.

REFERENCES

Abboud, D., S. Sieg-Zieba, and M. Eltabach (Feb. 2017). "Envelope analysis of rotating machine vibrations in variable speed conditions: A comprehensive treatment". In: *Mechanical Systems and Signal Processing* 84, pp. 200–226. DOI: `10.1016/j.ymssp.2016.06.033`.

Alsalaet, J. (Feb. 2020). "Detecting Rolling Elements Bearings Faults". In: Unpublished.

Arslan, H. and N. Aktürk (2008). "An investigation of rolling element vibrations caused by local defects". In: *ASME. J. Tribol* 130.4, p. 041101.

Attaran, B. and A. Ghanbarzadeh (Dec. 2015). "Bearing Fault Detection Based on Maximum Likelihood Estimation and Optimized ANN Using the Bees Algorithm". In: *J. Appl. Comput. Mech.* 1, pp. 35–43. DOI: `10.22055/jacm.2014.10547`.

Basu, S. and J. Andrews (2013). "Complexity in mathematical models of public health policies: a guide for consumers of models". In: *PLoS medicine* 10.10, e1001540.

Berredjem, T. and M. Benidir (Oct. 2018). "Bearing faults diagnosis using fuzzy expert system relying on an Improved Range Overlaps and Similarity method". In: *Expert Systems with Applications* 108, pp. 134–142. ISSN: 0957-4174. URL: `https://www.sciencedirect.com/science/article/pii/S0957417418302653`.

Bobylev, D., T. Choudhury, J. O. Miettinen, R. Viitala, E. Kurvinen, and J. Sopanen (2021). "Simulation-based transfer learning for support stiffness identification". In: *IEEE Access* 9, pp. 120652–120664.

Breiman, L. (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Camilleri, D. and T. Prescott (2017). "Analysing the limitations of deep learning for developmental robotics". In: *Biomimetic and Biohybrid Systems: 6th International Conference, Living Machines 2017, Stanford, CA, USA, July 26–28, 2017, Proceedings 6*. Springer, pp. 86–94.

Chaudhari, C. P., B. B. Thakare, S. R. Patil, and S. U. Gunjal (2015). "A study of bearing and its types". In: *Int J Adv Sci* 4.Special Issue 1, pp. 1674–1681.

Chen, C., Z. Liu, G. Yang, C.-C. Wu, and Q. Ye (2020). "An improved fault diagnosis using 1d-convolutional neural network model". In: *Electronics* 10.1, pp. 1–19.

Chen, F., B. Tang, T. Song, and L. Li (2014). "Multi-fault diagnosis study on roller bearing based on multi-kernel support vector machine with chaotic particle swarm optimization". In: *IEEE Transactions on Instrumentation and Measurement* 47, pp. 576–590.

Chen, Z. and W. Li (2017). "Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network". In: *IEEE Transactions on Instrumentation and Measurement* 66.7, pp. 1693–1702.

Chow, M.-Y., P. M. Mangum, and S. O. Yee (1991). "A neural network approach to real-time condition monitoring of induction motors". In: *IEEE Transactions on industrial electronics* 38.6, pp. 448–453.

Cui, H., M. Hong, Y. Qiao, and Y. Yin (2017). "Application of VPMCD method based on PLS for rolling bearing fault diagnosis". In: *Journal of Vibroengineering* 19.1, pp. 160–175.

De Moura, E., C. Souto, A. Silva, and M. Irmao (2011). "Evaluation of principal component analysis and neural network performance for bearing fault diagnosis from vibration signal processed by RS and DF analyses". In: *Mechanical Systems and Signal Processing* 25.5, pp. 1765–1772.

De Silva, C. W. (2007). *Vibration : fundamentals and practice*. 2nd ed. Boca Raton (FL): CRC Press/Taylor & Francis, p. 213. ISBN: 0-8493-1987-0.

Dietl, P., J. Wensing, and G. Van Nijen (2000). "Rolling bearing damping for dynamic analysis of multi-body systems—experimental and theoretical results". In: *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics* 214.1, pp. 33–43.

Dong, Y., Y. Li, H. Zheng, R. Wang, and M. Xu (2022). "A new dynamic model and transfer learning based intelligent fault diagnosis framework for rolling element bearings race faults: Solving the small sample problem". In: *ISA transactions* 121, pp. 327–348.

Du, W., X. Hou, and H. Wang (2022). *Time-Varying Degradation Model for Remaining Useful Life Prediction of Rolling Bearings under Variable Rotational Speed*. DOI: 10.3390/app12084044. URL: https://doi.org/10.3390/app12084044.

El Bouchefry, K. and R. S. de Souza (2020). "Learning in big data: Introduction to machine learning". In: *Knowledge discovery in big data from astronomy and earth observation*. Elsevier, pp. 225–249.

Eren, L. (2017). "Bearing fault detection by one-dimensional convolutional neural networks". In: *Mathematical Problems in Engineering* 2017, pp. 1–9.

Fuan, W., J. Hongkai, S. Haidong, D. Wenjing, and W. Shuaipeng (2017). "An adaptive deep convolutional neural network for rolling bearing fault diagnosis". In: *Measurement Science and Technology* 28.9, p. 095005.

Ghalamchi, B. (2014). "Dynamic Analysis Model of Spherical Roller Bearings with Defects". PhD thesis, p. 17.

Ghalamchi, B., J. Sopanen, and A. Mikkola (Jan. 2015). "Spherical Roller Bearing Simulation Model with Localized Defects". In: *Mechanisms and Machine Science* 21, pp. 1899–1909. DOI: 10.1007/978-3-319-06590-8_157.

– (2016). "Modeling and dynamic analysis of spherical roller bearing with localized defects: analytical formulation to calculate defect depth and stiffness". In: *Shock and Vibration* 2016.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2020). "Generative adversarial networks". In: *Communications of the ACM* 63.11, pp. 139–144.

Gu, Y., J. Cao, X. Song, and J. Yao (2021). "A denoising autoencoder-based bearing fault diagnosis system for time-domain vibration signals". In: *Wireless Communications and Mobile Computing* 2021, pp. 1–7.

Guo, L., Y. Lei, S. Xing, T. Yan, and N. Li (2018). "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data". In: *IEEE Transactions on Industrial Electronics* 66.9, pp. 7316–7325.

Guo, X., C. Shen, and L. Chen (2016). "Deep fault recognizer: An integrated model to denoise and extract features for fault diagnosis in rotating machinery". In: *Applied Sciences* 7.1, p. 41.

Gupta, P. (2008). "On the geometrical imperfections in ball bearings". In: *ASME. J. Tribol* 110.1, pp. 19–25.

Gupta, P. K. (1984). *Advanced dynamics of rolling elements*. Berlin: Springer. ISBN: 0-387-96031-7.

Hamrock, B. J. and D. Dowson (1981). "Ball bearing lubrication: the elastohydrodynamics of elliptical contacts". In.

Han, D., N. Zhao, and P. Shi (2017). "A new fault diagnosis method based on deep belief network and support vector machine with Teager–Kaiser energy operator for bearings". In: *Advances in Mechanical Engineering* 9.12, p. 1687814017743113.

Han, T. and D. Jiang (2016). "Rolling bearing fault diagnostic method based on VMD-AR model and random forest classifier". In: *Shock and Vibration* 2016.

Han, T., C. Liu, W. Yang, and D. Jiang (2019). "A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults". In: *Knowledge-based systems* 165, pp. 474–487.

Harris, T. A. and W. Crecelius (1986). "Rolling bearing analysis". In: *ASME. J. Tribol* 108.1, pp. 149–150.

Harsha, S. P. and P. Kankar (2004). "Stability analysis of a rotor bearing system due to surface waviness and number of balls". In: *International Journal of Mechanical Sciences* 46.7, pp. 1057–1081.

Hasan, M. J., M. Sohaib, and J.-M. Kim (2019). "1D CNN-based transfer learning model for bearing fault diagnosis under variable working conditions". In: *Computational Intelligence in Information Systems: Proceedings of the Computational Intelligence in Information Systems Conference (CIIS 2018) 3*. Springer, pp. 13–23.

He, C., J. Zhang, K. Geng, S. Wang, M. Luo, X. Zhang, and C. Ren (2022). "Advances in ultra-precision machining of bearing rolling elements". In: *The International Journal of Advanced Manufacturing Technology* 122.9-10, pp. 3493–3524.

Hong, S.-W. and V.-C. Tong (2016). “Rolling-element bearing modeling: A review”. In: *International Journal of Precision Engineering and Manufacturing* 17, pp. 1729–1749.

Hou, P., L. Wang, Z. Xie, and Q. Peng (2021). “An improved model for investigating the vibration characteristics of ball bearing owing to outer race waviness under high speed”. In: *Noise Control Engineering Journal* 69.2, pp. 89–101.

Iman, M., H. R. Arabnia, and R. M. Branchinst (2021). “Pathways to artificial general intelligence: a brief overview of developments and ethical issues via artificial intelligence, machine learning, deep learning, and data science”. In: *Advances in Artificial Intelligence and Applied Cognitive Computing: Proceedings from ICAI’20 and ACC’20*, pp. 73–87.

Iman, M., H. R. Arabnia, and K. Rasheed (2023). “A review of deep transfer learning and recent advancements”. In: *Technologies* 11.2, p. 40.

Jack, L. and A. Nandi (2002). “Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms”. In: *Mechanical systems and signal processing* 16.2-3, pp. 373–390.

Jacobson, B. (2003). “The Stribeck memorial lecture”. In: *Tribology International* 36.11, pp. 781–789.

Janssens, O., V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke (2016). “Convolutional neural network based fault detection for rotating machinery”. In: *Journal of Sound and Vibration* 377, pp. 331–345.

Jardine, A. K., D. Lin, and D. Banjevic (2006). “A review on machinery diagnostics and prognostics implementing condition-based maintenance”. In: *Mechanical systems and signal processing* 20.7. pp. 1483–1510, pp. 1483–1510.

Kurvinen, E., J. Sopanen, and A. Mikkola (2015). “Ball bearing model performance on various sized rotors with and without centrifugal and gyroscopic forces”. In: *Mechanism and Machine Theory* 90, pp. 240–260.

Larson, E. C., B. E. Parker, and B. R. Clark (2002). “Model-based sensor and actuator fault detection and isolation”. In: *Proceedings of the 2002 American Control Conference (IEEE cat. no. ch37301)*. Vol. 5. IEEE, pp. 4215–4219.

Lee, Y. O., J. Jo, and J. Hwang (2017). “Application of deep neural network and generative adversarial network to industrial maintenance: A case study of induction motor fault detection”. In: *2017 IEEE international conference on big data (big data)*. IEEE, pp. 3248–3253.

Lei, Y., B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi (2020). “Applications of machine learning to machine fault diagnosis: A review and roadmap”. In: *Mechanical Systems and Signal Processing* 138, pp. 1–39.

Lessmeier, C., J. K. Kimotho, D. Zimmer, and W. Sextro (2016). “Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification". In: *PHM Society European Conference*. Vol. 3. 1.

Li, C., L. Ledo, M. Delgado, M. Cerrada, F. Pacheco, D. Cabrera, R.-V. Sánchez, and J. Valente de Oliveira (Aug. 2017). "A Bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification". In: *Knowledge-Based Systems* 129, pp. 39–60. ISSN: 0950-7051. URL: https://www.sciencedirect.com/science/article/pii/S0950705117302186.

Li, F. Q., Y. G. Fan, and Y. X. Zhang (2014). "Rolling bearing fault detection method based on wavelet packet energy spectrum-PLS". In: *Advanced Materials Research*. Vol. 971. Trans Tech Publ, pp. 697–700.

Li, H. and D.-y. Xiao (2011). "Fault diagnosis of Tennessee Eastman process using signal geometry matching technique". In: *EURASIP Journal on Advances in Signal Processing* 2011.1, pp. 1–19.

Li, K., Y. L. Zhang, and Z. X. Li (2012). "Application research of Kalman filter and SVM applied to condition monitoring and fault diagnosis". In: *Applied Mechanics and Materials*. Vol. 121. Trans Tech Publ, pp. 268–272.

Li, S., G. Liu, X. Tang, J. Lu, and J. Hu (2017). "An ensemble deep convolutional neural network model with improved DS evidence fusion for bearing fault diagnosis". In: *Sensors* 17.8, p. 1729.

Li, X., W. Zhang, and Q. Ding (2018). "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning". In: *Neurocomputing* 310, pp. 77–95.

Li, Y., B. Miao, W. Zhang, P. Chen, J. Liu, and X. Jiang (2019). "Refined composite multi-scale fuzzy entropy: Localized defect detection of rolling element bearing". In: *Journal of Mechanical Science and Technology* 33, pp. 109–120.

Liu, J., Y. Xu, and G. Pan (2021). "A combined acoustic and dynamic model of a defective ball bearing". In: *Journal of Sound and Vibration* 501, p. 116029.

Liu, R., B. Yang, E. Zio, and X. Chen (2018). "Artificial intelligence for fault diagnosis of rotating machinery: A review". In: *Mechanical Systems and Signal Processing* 108, pp. 33–47.

Liu, T. I., J. H. Singonahalli, and N. R. Iyer (Sept. 1996). "DETECTION OF ROLLER BEARING DEFECTS USING EXPERT SYSTEM AND FUZZY LOGIC". In: *Mechanical Systems and Signal Processing* 10.5, pp. 595–614. ISSN: 0888-3270. URL: https://www.sciencedirect.com/science/article/pii/S0888327096900412.

Liu, X., S. Deng, and H. Teng (2011). "Dynamic stability analysis of cages in high-speed oil-lubricated angular contact ball bearings". In: *Transactions of Tianjin University* 17.1, pp. 20–27.

Lu, C., Z.-Y. Wang, W.-L. Qin, and J. Ma (2017). "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification". In: *Signal Processing* 130, pp. 377–388.

Makokha, J. M. (2021). "Artificial Intelligence Paradigms and the Future of Learning: What a Partial Review of Half a Century of AI Conceptualization Suggests". In: *2021 ASEE Virtual Annual Conference Content Access*, pp. 1–22.

Mao, W., L. Ding, S. Tian, and X. Liang (2020). "Online detection for bearing incipient fault based on deep transfer learning". In: *Measurement* 152, p. 107278.

Matheron, G. (1975). "Random Sets and Integral Geometry, Jonn Wiley & Sons". In: *Inc., New York*.

Mba, D. and L. Reuben (Mar. 2014). "Bearing Time-to-Failure Estimation using Spectral Analysis Features". In: *Structural Health Monitoring* 13, pp. 219–230. DOI: `10.1177/1475921713520029`.

Mechefske, C. and J. Mathew (1992). "Fault detection and diagnosis in low speed rolling element bearings Part II: The use of nearest neighbour classification". In: *Mechanical Systems and Signal Processing* 6.4, pp. 309–316.

Mylaraswamy, D. and V. Venkatasubramanian (1997). "A hybrid framework for large scale process fault diagnosis". In: *Computers & chemical engineering* 21, S935–S940.

Nikolaou, N. and I. Antoniadis (2003). "Application of morphological operators as envelope extractors for impulsive-type periodic signals". In: *Mechanical Systems and Signal Processing* 17.6, pp. 1147–1162.

Niu, L., H. Cao, Z. He, and Y. Li (2014). "Dynamic modeling and vibration response simulation for high speed rolling ball bearings with localized surface defects in raceways". In: *Journal of Manufacturing Science and Engineering* 136.4.

Nutakor, C., D. Talbot, and A. Kahraman (2019). "An experimental characterization of the friction coefficient of a wind turbine gearbox lubricant". In: *Wind Energy* 22.4, pp. 509–522.

Pan, S. J. and Q. Yang (2010). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.

Pandya, D., S. Upadhyay, and S. Harsha (2012). "ANN based fault diagnosis of rolling element bearing using time-frequency domain feature". In: *International Journal of Engineering Science and Technology* 4.6, pp. 2878–2886.

Patidar, S. and P. K. Soni (2013). "An overview on vibration analysis techniques for the diagnosis of rolling element bearing faults". In: *International Journal of Engineering Trends and Technology (IJETT)* 4.5, pp. 1804–1809.

Patil, M., J. Mathew, P. Rajendrakumar, and S. Desai (2010). "A theoretical model to predict the effect of localized defect on vibrations associated with ball bearing". In: *International Journal of Mechanical Sciences* 52.9, pp. 1193–1201.

Peng, B., Y. Bi, B. Xue, M. Zhang, and S. Wan (2022). “A Survey on Fault Diagnosis of Rolling Bearings”. In: *Algorithms* 15.10, p. 347.

Phung, V. H. and E. J. Rhee (2019). “A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets”. In: *Applied Sciences* 9.21, pp. 1–16.

Poór, P., D. Ženíšek, and J. Basl (2019). “Historical overview of maintenance management strategies: Development from breakdown maintenance to predictive maintenance in accordance with four industrial revolutions”. In: *Proceedings of the international conference on industrial engineering and operations management, Pilsen, Czech Republic*, pp. 23–26.

Prajapati, A., J. Bechtel, and S. Ganesan (Jan. 2012). “Condition based maintenance: a survey”. In: *Journal of Quality in Maintenance Engineering* 18.4, pp. 384–400. ISSN: 1355-2511. DOI: `10.1108/13552511211281552`. URL: `https://doi.org/10.1108/13552511211281552`.

Randall, R. B. and J. Antoni (2011). “Rolling element bearing diagnostics—A tutorial”. In: *Mechanical systems and signal processing* 25.2, pp. 485–520.

Rojas, A. and A. K. Nandi (2005). “Detection and classification of rolling-element bearing faults using support vector machines”. In: *2005 IEEE Workshop on Machine Learning for Signal Processing*. IEEE, P. 153–158.

Samanta, B. and K. Al-Balushi (Mar. 2003). “ARTIFICIAL NEURAL NETWORK BASED FAULT DIAGNOSTICS OF ROLLING ELEMENT BEARINGS USING TIME-DOMAIN FEATURES”. In: *Mechanical Systems and Signal Processing* 17.2, pp. 317–328. ISSN: 0888-3270. URL: `https://www.sciencedirect.com/science/article/pii/S0888327001914621`.

Samanta, B., K. Al-Balushi, and S. Al-Araimi (2003). “Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection”. In: *Engineering applications of artificial intelligence* 16.7-8, pp. 657–665.

Saruhan, H., S. Sandemir, A. Çiçek, and I. Uygur (June 2014). “Vibration Analysis of Rolling Element Bearings Defects”. In: *Journal of Applied Research and Technology* 12.3, pp. 384–395. ISSN: 1665-6423. URL: `https://www.sciencedirect.com/science/article/pii/S1665642314716207`.

Shao, H., H. Jiang, H. Zhang, W. Duan, T. Liang, and S. Wu (2018). “Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing”. In: *Mechanical Systems and Signal Processing* 100, pp. 743–765.

Shen, F., C. Chen, R. Yan, and R. X. Gao (2015). “Bearing fault diagnosis based on SVD feature extraction and transfer learning classification”. In: *2015 prognostics and system health management conference (PHM)*. IEEE, pp. 1–6.

Sheng, S. (2017). *Wind turbine gearbox reliability database, operation and maintenance research update*. Tech. rep. National Renewable Energy Lab.(NREL), Golden, CO (United States), p. 3.

Shi, H., J. Guo, Z. Yuan, Z. Liu, M. Hou, and J. Sun (2020). "Incipient fault detection of rolling element bearings based on deep EMD-PCA algorithm". In: *Shock and Vibration* 2020, pp. 1–17.

Shin, J.-H. and H.-B. Jun (Apr. 2015). "On condition based maintenance policy". In: *Journal of Computational Design and Engineering* 44. DOI: 10.1016/j.jcde.2014.12.006.

Sopanen, J. and A. Mikkola (2003a). "Dynamic model of a deep-groove ball bearing including localized and distributed defects. Part 2: Implementation and results". In: *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics* 217.3, pp. 213–223.

Sopanen, J. and A. Mikkola (Sept. 2003b). "Dynamic Model of a Deep Groove Ball Bearing Including Localized and Distributed Defects, Part 1: Theory". In: *Proceedings of the Institution of Mechanical Engineers Part K Journal of Multi-body Dynamics* 217(K), pp. 201–211. DOI: 10.1243/14644190360713551.

Sugumaran, V., V. Muralidharan, and K. Ramachandran (2007). "Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing". In: *Mechanical systems and signal processing* 21.2, pp. 930–942.

Tandon, N. and A. Choudhury (1999). "A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings". In: *Tribology international* 32.8, pp. 469–480.

Tidriri, K., N. Chatti, S. Verron, and T. Tiplica (2016). "Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges". In: *Annual Reviews in Control* 42, pp. 63–81.

Tong, Z., W. Li, B. Zhang, F. Jiang, and G. Zhou (2018). "Bearing fault diagnosis under variable working conditions based on domain adaptation using feature transfer learning". In: *IEEE Access* 6, pp. 76187–76197.

Trajin, B., J. Regnier, and J. Faucher (2008). "Detection of bearing faults in asynchronous motors using Luenberger speed observer". In: *2008 34th Annual Conference of IEEE Industrial Electronics*, pp. 3073–3078. DOI: 10.1109/IECON.2008.4758451.

Ullah, Z., B. A. Lodhi, and J. Hur (2020). "Detection and identification of demagnetization and bearing faults in PMSM using transfer learning-based VGG". In: *Energies* 13.15, p. 3834.

Verstraete, D., A. Ferrada, E. L. Droguett, V. Meruane, M. Modarres, et al. (2017). "Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings". In: *Shock and Vibration* 2017.

Voda, A. I. and L.-D. Radu (2019). "How can artificial intelligence respond to smart cities challenges?" In: *Smart Cities: Issues and Challenges*. Elsevier, pp. 199–216.

Walters, C. T. (1971). "The dynamics of ball bearings". In.

Wang, J., Z. Mo, H. Zhang, and Q. Miao (2019). "A deep learning method for bearing fault diagnosis based on time-frequency image". In: *IEEE Access* 7, pp. 42373–42383.

Wang, Y., W. Wang, S. Zhang, and Z. Zhao (2015). "Investigation of skidding in angular contact ball bearings under high speed". In: *Tribology International* 92, pp. 404–417.

Wen, L., X. Li, and L. Gao (2020). "A transfer convolutional neural network for fault diagnosis based on ResNet-50". In: *Neural Computing and Applications* 32, pp. 6111–6124.

Xiang, Z., X. Zhang, W. Zhang, and X. Xia (2019). "Fault diagnosis of rolling bearing under fluctuating speed and variable load based on TCO spectrum and stacking auto-encoder". In: *Measurement* 138, pp. 162–174.

Xie, Y. and T. Zhang (2018a). "A transfer learning strategy for rotation machinery fault diagnosis based on cycle-consistent generative adversarial networks". In: *2018 Chinese automation congress (CAC)*. IEEE, pp. 1309–1313.

– (2018b). "Imbalanced learning for fault diagnosis problem of rotating machinery based on generative adversarial networks". In: *2018 37th Chinese Control Conference (CCC)*. IEEE, pp. 6017–6022.

Yang, B.-S., X. Di, and T. Han (2008). "Random forests classifier for machine fault diagnosis". In: *Journal of mechanical science and technology* 22, pp. 1716–1725.

Yazici, B. and G. B. Kliman (Mar. 1998). *Adaptive, on line, statistical method and apparatus for motor bearing fault detection by passive motor current monitoring*. US Patent 5,726,905.

Ye, M., J. Zhang, and J. Yang (2022). *Bearing Fault Diagnosis under Time-Varying Speed and Load Conditions via Observer-Based Load Torque Analysis*.

Ying, X. (2019). "An overview of overfitting and its solutions". In: *Journal of physics: Conference series*. Vol. 1168. IOP Publishing, pp. 1–7.

Yu, Y., D. Yu, and C. Junsheng (2006). "A roller bearing fault diagnosis method based on EMD energy entropy and ANN". In: *Journal of sound and vibration* 294.1-2, pp. 269–277.

Zhang, R., H. Tao, L. Wu, and Y. Guan (2017). "Transfer learning with neural networks for bearing fault diagnosis in changing working conditions". In: *IEEE Access* 5, pp. 14347–14357.

Zhang, S., S. Zhang, B. Wang, and T. G. Habetler (2020). "Deep learning algorithms for bearing fault diagnostics—A comprehensive review". In: *IEEE Access* 8, pp. 29857–29881.

Zhu, K., X. Song, and D. Xue (2014). "A roller bearing fault diagnosis method based on hierarchical entropy and support vector machine with particle swarm optimization algorithm". In: *Measurement* 47, pp. 669–675.