Department of Electrical, Electronic and Computer Engineering

Artificial Intelligence Techniques for Bearing Fault Diagnosis

MD JUNAYED HASAN

Doctor of Philosophy

2021

UNIVERSITY OF ULSAN, REPUBLIC OF KOREA

Artificial Intelligence Techniques for Bearing Fault Diagnosis

by

Md Junayed Hasan

A thesis submitted in partial fulfillment for the
Degree of Doctor of Philosophy

in the
Department of Electrical, Electronic and Computer Engineering

Supervisor: Jong-Myon Kim, Ph.D.

May 2021

Publication No. _____

DECLARATION OF AUTHORSHIP

I, Md Junayed Hasan, declare that this thesis titled "Artificial Intelligence Techniques for Bearing Fault Diagnosis" and the work presented herein are my own. I confirm that:

- This work was done completely while in candidature for a research degree at the University of Ulsan, Republic of Korea.
- Where I have consulted the published work of others, it has always been clearly attributed to its original sources, and I have acknowledged all the sources.
- Where I have quoted from the work of others, the source is always mentioned. Except for such quotations, this thesis is entirely my work.
- Each research chapter of this dissertation relates to one or more SCI(E) indexed journal articles, which have been published/accepted.

Md Junayed Hasan

UNIVERSITY OF ULSAN, REPUBLIC OF KOREA

May 2021

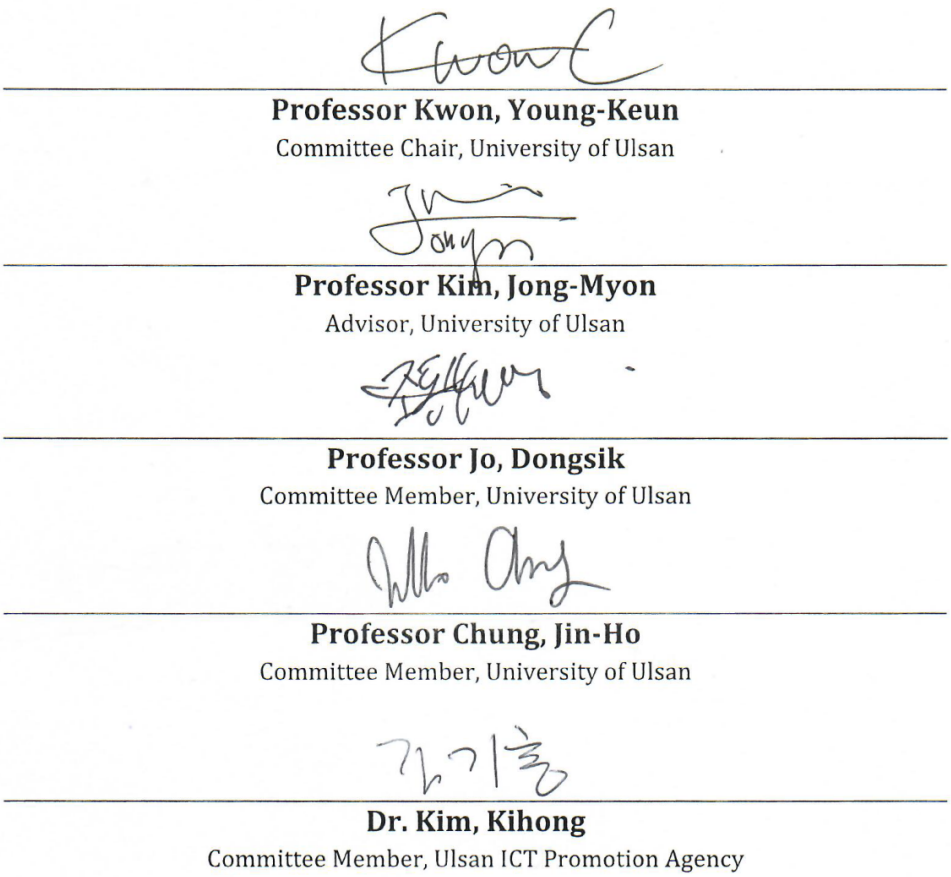
TERMS AND CONDITIONS RELATED TO COPYRIGHT

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g., Ph.D.) at the University of Ulsan, Republic of Korea. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Artificial Intelligence Techniques for Bearing Fault Diagnosis

This certifies that the dissertation of Md Junayed Hasan is approved.



Professor Kwon, Young-Keun
Committee Chair, University of Ulsan

Professor Kim, Jong-Myon
Advisor, University of Ulsan

Professor Jo, Dongsik
Committee Member, University of Ulsan

Professor Chung, Jin-Ho
Committee Member, University of Ulsan

Dr. Kim, Kihong
Committee Member, Ulsan ICT Promotion Agency

Department of Electrical, Electronic and Computer Engineering

UNIVERSITY OF ULSAN, REPUBLIC OF KOREA

May 2021

“ You may write me down in history

With your bitter, twisted lies,

You may trod me in the very dirt

But still, like dust, I'll rise.”

Maya Angelou,

"Still I Rise" from *And Still I Rise: A Book of Poem.*

VITA

MD JUNAYED HASAN was born in Jessore, Bangladesh, in 1993. He received the B.Sc. degree in Computer Science and Engineering from the Bangladesh Rural Advancement Committee (BRAC) University, Dhaka, Bangladesh, in 2016. Since 2017, he has been working towards pursuing the M.Sc. leading Ph.D. degree in Computer Engineering with the University of Ulsan (UOU), Ulsan, Republic of Korea under the supervision of Professor Jong-Myon, Kim.

His research interests include the data-driven fault diagnosis and anomaly detection of complex engineering systems using advanced signal, and image processing, fault feature extraction, feature engineering, deep learning, and explainable artificial intelligence techniques.

ACKNOWLEDGEMENTS

First, I am thankful to the Almighty for giving me the chance and blessing to complete this long journey. I would love to offer my gratitude to my dear parents who guided me and supported me morally and financially at every stage of my life to achieve this success, to my wife who always stands by me in all circumstances and supports me both physically and mentally.

Above all, I would like to express my profound gratitude to my supervisor, Professor Jong-Myon Kim, for allowing me to work under his kind guidance and supervision during my PhD. He has really given me the valuable knowledge both in technical and non-technical matters. I am very grateful to him for the generous support, encouragement, and the financial assistance during my stay in Korea.

I owe my deepest gratitude to Dr. Jia Uddin. Without his encouragement and support, it would be impossible for me to start this journey.

I am grateful to my friends, colleagues, and all the members of the Ulsan Industrial Artificial Intelligence Laboratory (UIAI LAB) at University of Ulsan, for the camaraderie, and their ever-enthusiastic supports and cheerfulness during my stay in Korea. My sincere gratitude goes to my thesis committee for their valuable suggestions, and constructive feedback.

Finally, I would also like to gratefully acknowledge the financial support of the BK21 plus scholarship program, Ministry of Science, ICT and future Planning of the Republic of Korea, the National Research Foundation of Korea (NRF), the Korea Institute of Energy Technology Evaluation and Planning (KETEP), the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea, the Technology Infrastructure Program funded by the Ministry of SMEs and Startups (MSS) of the Republic of Korea, and Electronics and Telecommunications Research Institute (ETRI), which allow me to pursue my study and research at University of Ulsan (UOU).

Md Junayed Hasan
University of Ulsan, Republic of Korea
May 2021

To my amazing parents, and beautiful wife.

Abstract

The condition monitoring of bearing by using vibration signal is an indispensable mission in today's manufacture processes. To develop the solution for monitoring the conditions of the bearing, signal processing plays a crucial role for finding out the meaningful information from the vibration signals. The existing signal processing approach present in the current literature mainly focuses on extracting the meaningful information from the bandwidths which contains the fault frequencies. However, when the system is changed, for the new dataset, these approaches failed to extract meaningful information. Moreover, due to noise, and non-stationarity, not always the fault frequencies are appeared into the expected region. The only solution is to explore and analyze the complete dataset from that new environment to make an adjustment with the signal processing approach. Currently, several Machine Learning (ML) and Deep Learning (DL)- based models have attained brilliant results in fault detection and diagnosis under consistent working conditions. Generally, the successful ML models consist of 4 stages, i.e., (1) data preprocessing, (2) feature extraction, (3) feature selection, and (4) classification. However, due to non-linearity, and non-stationarity, it is very difficult to extract and analyze the fault feature information from variable working conditions with the existing preprocessing steps. Therefore, the extracted feature information becomes easily vulnerable when the dataset along with the working conditions are changed. Moreover, the existing feature selection techniques lack of explain ability, which makes it difficult to interpret/justify the performance of the classifier. Furthermore, so far, the existing classifiers are used like a black-box model, which makes it even harder to debug the decision of the classifier. Thus, for every new dataset or application domain, there is a necessity to build a model from scratch. Therefore, to solve all these problems, a concept of explain ability is proposed in two steps for the first time in the field of bearing fault diagnosis: (a) incorporating explain ability of the feature selection process, and (b) interpretation of the machine learning classifier performance with respect to the selected features.

In this dissertation, an explainable ML based fault diagnosis model for bearing is proposed with 5-stages, i.e., (1) a data preprocessing method based on a Faster Discrete Orthogonal Stockwell Transformation (FDOST) Coefficient is proposed to analyze the vibration signals for capturing the invariant patterns from both time-frequency, and corresponding phase-angel information for variable speed, and load conditions, (2) a statistical feature extraction method is introduced to capture the significance from the invariant pattern of the preprocessed data obtained by FDOST, (3) an explainable feature selection process is proposed by introducing a wrapper based feature selector - Boruta, (4) a feature filtration method is proposed after the feature selection process to avoid the multicollinearity

problem by introducing Spearman's rank correlation coefficient, and finally, (5) an additive Shapley explanations followed by k-NN is proposed to diagnose, and to explain individual decision of the k-NN classifier for understanding, and debugging the performance of the model. Thus, the idea of explainability is introduced for the first time in the field of bearing fault diagnosis in two steps: (a) incorporating explainability to the feature selection process, and (b) interpretation of the classifier performance with respect to the selected features.

Further, to extend the explainable model to mitigate the training time, and to automate the feature extraction, selection, and filtration process, a Transfer Learning (TL) based DL algorithm is proposed by utilizing the visual patterns obtained from FDOST time-frequency coefficient-based Vibration Imaging (VI).

The effectiveness of each proposed model is demonstrated on two different datasets obtained from separate bearing testbeds containing different mechanical faults in rotating machinery along with variable load, and speed conditions. Lastly, an assessment of several state-of-art fault diagnosis algorithms in rotating machinery is included.

Table of Contents

List of Figures

List of Tables

Nomenclature

2D	Two – Dimensional
AI	Artificial Intelligence
ANN	Artificial Neural Network
AS	Accuracy Score
BN	Batch Normalization
BSS	Bootstrapped Set of Samples
BT	Ball Type
CL	Convolution Layer
CM	Confusion Matrix
CM-FD	Condition Monitoring and Fault Diagnosis
CNN	Convolution Neural Network
CV	Cross Validation
CWRU	Case Western Reserve University
CWT	Continuous Wavelet Transform
DCS	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DL	Deep Learning
DOST	Discrete Orthogonal Stockwell Transform
DST	Discrete Stockwell Transform
DT	Decision Tree
EE	Energy Entropy
EIS	Extended Information System
EMA	Exponential Moving Average
EMD	Empirical Mode Decomposition

EWT	Empirical Wavelet Transformation
FCL	Fully Connected Layer
FDOST	Fast Discrete Orthogonal Stockwell Transform
FFP	Final Feature Pool
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
FT	Fourier Transform
FTL	Fine Tuning-based Transfer Learning
GA	Genetic Algorithm
GP	Genetic Programming
ICS	Inter Class Separability
IFT	Inverse Fourier Transform
IRT	Inner Roller Type
k-NN	k-Nearest Neighbor
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LIME	Local Interpretable Model-Agnostic Explanations
LS	Laplacian Score
MDA	Mean Decrease in Accuracy
ML	Machine Learning
MW	Mother Wavelet
NB	Naïve Bayes
NT	Normal Type
OOB	Out of Bag
OR	Occam's Razor
ORT	Outer Roller Type

PCA	Principal Component Analysis
PL	Pooling Layer
RELU	Rectified Linear Unit
RF	Random Forest
RMS	Root Mean Square
RMSProp	Root Mean Square Propagation
RPM	Revolution Per Minute
RT	Roller Type
SA	Shadow Attribute
SFP	Statistical Feature Pool
SIP	Significant Information Pool
SRCC	Spearman's Rank Correlation Coefficient
ST	Stockwell Transform
STD	Standard Deviation
SVM	Support Vector Machine
TL	Transfer Learning
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
t-SNE	t-Stochastic Neighbor Embedding
UFP	Updated Feature Pool
VI	Vibration Imaging
VMD	Variational Mode Decomposition
WCC	Within Class Closeness
WPD	Wavelet Packet Decomposition
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

The motivations, contributions and outline of the dissertation are briefly summarized in this chapter. All research materials presented afterwards have been submitted to and published in peer-reviewed journals. Each of the following chapters addresses the core motivation and background knowledge of the problem addressed therein. While **Section 1.1** provides the overall motivations of this dissertation, **Section 1.2** and **Section 1.3** are dedicated to the main objectives, contribution, and the outline of the thesis, respectively.

1.1 Motivation

As greater demands are placed on existing assets in terms of higher output or increased efficiency, the need to understand when things are starting to go wrong is becoming more important. Add to this the increasing complexity and automation of plant and equipment, it becomes more important to have a properly structured and funded maintenance strategy. There is also a need to understand the operation of equipment so that improvements in plant output and efficiency can be realized. In today's increasingly competitive world all these issues are of key importance and can only be achieved through a properly structured and financed maintenance strategy that meets the business needs. In the manufacturing process, unexpected equipment failures can be costly and possibly tragic, resulting in unexpected manufacture downtime, costly substitution of parts and safety and ecological concerns. Predictive maintenance is a procedure for observing equipment during operation to pinpoint any deterioration, enabling maintenance to be planned and operational costs reduced [1]. Rolling element bearings are critical components used extensively in rotating equipment's

and, if they fail unexpectedly, can result in a devastating failure with associated high restoration and replacement costs [2,3]. Moreover, these are highly susceptible to the damage because of various factors such as variable speed and load, multiple fault severities, ample noise alternative load conditions, etc. [4]. As these components are highly likely to experience frequent wear and tear, therefore, become the primary reason for the sudden failure of the rotating machineries [2]. In return, industry may face, unexpected downtime, huge economic loss, and safety issues [5–8]. So, in past few decades, industries have recognized the significance of reliable and robust Condition Monitoring and Fault Diagnosis (CM-FD) techniques to mitigate these issues [9]. These CM-FD techniques can be developed by using data from different modalities including acoustic emissions [10], vibration acceleration signals [11], ultrasonic signals [12]. Among them, vibration signals are arguably the popular choice for the bearing CM-FD because these signals contain clear fault related signatures and can be explored easily through signals processing techniques [13,14].

For the past several years, researchers have been trying to develop a generalized approach for bearing CM-FD which can identify the faults from the given data with high efficacy. These data-driven approaches usually focus on two parts, i.e., **(a)** to identify the fault pattern from the extracted features by various signal processing algorithms, **(b)** by utilizing those distinguishable patterns, develop a classification/prediction algorithm using different Machine Learning (ML) or Deep Learning (DL) based approaches. However, these existing solutions have several limitations:

(1) The vibration signals of the bearing are prone to inconsistencies due to several factors which makes them difficult to analyze. Therefore, if there is a little variation in the given CM-FD scenario, the adopted signal processing techniques suffer to extract fault characteristics, properly. Moreover, when working condition is changed, the marginal probability distribution of the data is also changed. Therefore, the feature space of the acquired vibration signals from different working conditions are different hypothetically. With the existing signal processing approaches, if we want to preprocess the vibration data, it can successfully provide us meaningful information related to certain health type. However, when the working condition is changed, the extracted information for a certain health type is also changed. Thus, the marginal distribution of these preprocessed data, and/or the marginal distribution of these extracted feature space remain different. Therefore, the present signal processing techniques adapted into the current literature fails to create a CM-FD solution for different working conditions.

(2) For ML based analysis, feature extraction, and selection is necessary from the preprocessed data. For feature selection step, in the current literature, mostly we are habituated with the evolutionary based approaches (i.e., Genetic Algorithms (GA)), or wrapper-based approaches (i.e., forward-backward selection). These types of feature

selectors cannot explain the proper reasoning for selecting certain features. Therefore, the impact of the selected features on certain classifier can never be justified in accurate manner. Furthermore, the existing classifiers in CM-FD literature are working as a black-box model.

1.2 Thesis Objective and Contribution

In this dissertation, the main objective is to improve the existing fault diagnosis approaches for the condition monitoring of bearing for invariant working conditions by incorporating similarity among the feature spaces of different working conditions with explain ability, state-of-art accuracy, interpretability of the Artificial Intelligent (AI) models, and computational efficiency. The overall contributions are depicted into Figure 1.1.

(1) To capture the information of variable working conditions from the vibration signals of bearing both at low and high frequencies, a computationally advanced version of Stockwell Transform (ST) is considered as the signal preprocessing step. This proposed approach can bring similarity to the marginal distribution of the preprocessed data of certain health type for variable working conditions.

(2) To find out the best features from the statistical properties of the pre-processed data, a wrapper-based feature selector Boruta is proposed. This algorithm can justify the selection of each feature attribute with the help of embedded Random Forest (RF) classifier. Thus, the interpretation of feature selection process of bearing fault diagnosis is properly presented. Furthermore, a second stage of feature filtration technique is proposed by using Spearman's Rank Correlation Coefficient (SRCC) to create a biasfree feature set for the classifier. Thus, it helps the classifier to avoid the multicollinearity problem. Finally, to create a sweet spot in between the accuracy and the explain ability, a non-parametric classifier – k-Nearest Neighbor (k-NN) is proposed to generate the results of the proposed diagnostic model. The predictions of k-NN are explained via kernel SHAP.

(3) To automate the feature extraction, and selection process, from the ST based proposed preprocessing approach, a Two-Dimensional (2D) time-frequency image is formed to capture the invariant patterns from the data. Then, these images are converted into grayscale, coined as Vibration Imaging (VI), to utilize the Convolutional Neural Network (CNN) efficiently. Furthermore, to bring the computational efficiency to the DL based approaches, a CNN-based-Transfer Learning (TL) model is proposed.

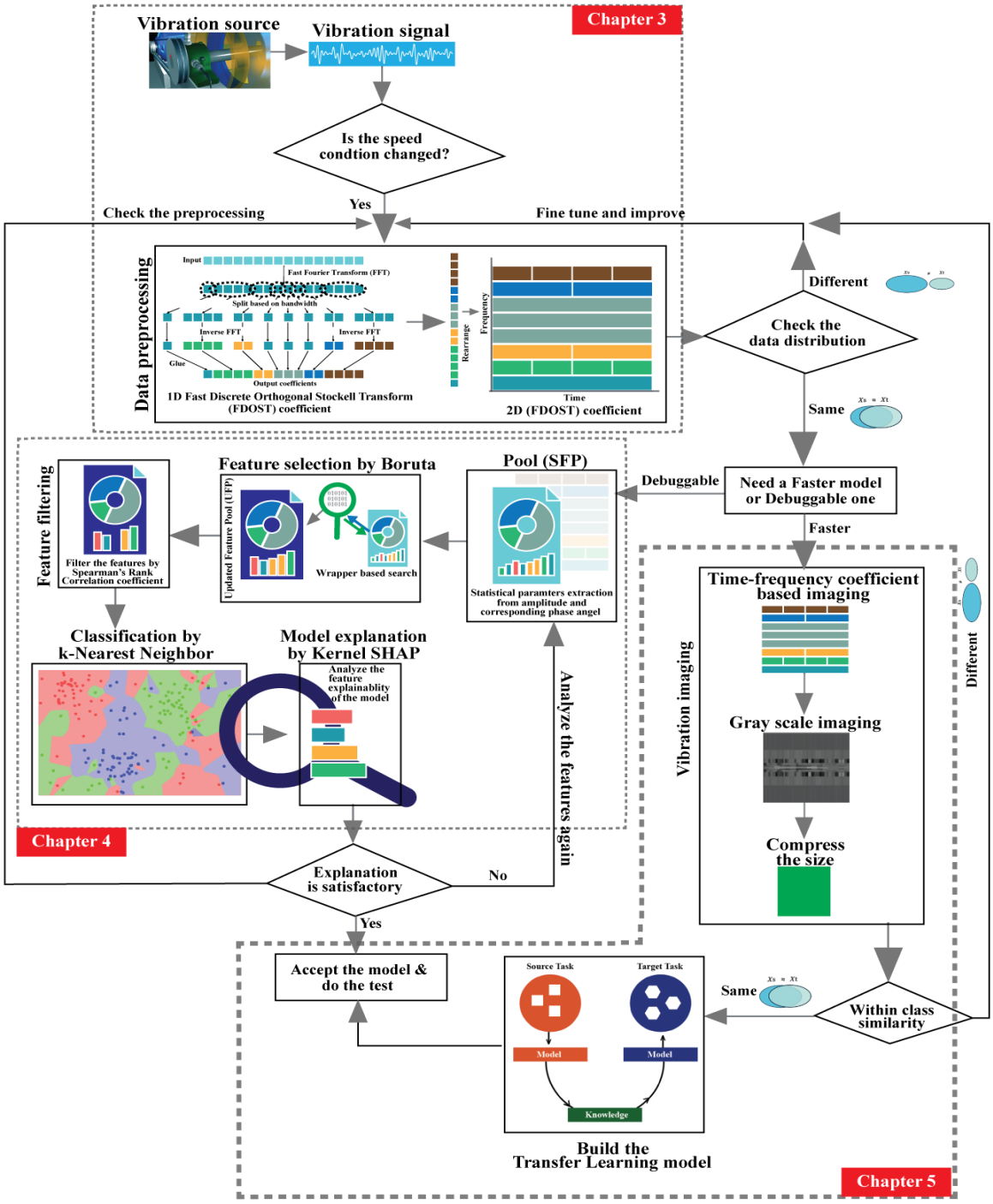


Figure 1.1: Data processing and overall system architecture.

1.3 Thesis Outline

The contents of this dissertation are briefly outlined below:

- **Chapter 1** briefly outlines the motivations, objectives, and outline of the thesis.
- **Chapter 2** presents the details of the considered datasets along with the experimental set-up.
- **Chapter 3** proposes a fault diagnosis model for bearing using Fast Discrete Orthogonal Stockwell Transformation Coefficient (FDOST) and Genetic Algorithm (GA).
- **Chapter 4** presents an Explainable AI (XAI) based approach for condition monitoring of bearing.
- **Chapter 5** introduces a Transfer Learning (TL) based approach for condition monitoring of bearing by using a Fast Discrete Orthonormal Discrete Stockwell Transformation (FDOST) coefficient-based Vibration Imaging (VI).
- The thesis is concluded in **Chapter 6** with a summary of the contributions made therein and a discussion of future research directions.

Chapter 2

Dataset Description

In this dissertation, to demonstrate every chapter, we are considering 2 datasets for every experimental analysis. Thus, the result analysis becomes consistent throughout the book. While considering the datasets, we have studied different load, and speed conditions. The crack size kept consistent for this study. Among the 2 datasets, one is considered from public repository to validate performances with benchmark studies, while another one is considered from the own experimental self-designed testbed to demonstrate the performance of the proposed algorithms in the real-world scenario.

2.1 Case Western Reserve University (CWRU) Bearing Dataset

For the case study 1 of every chapter, vibration signals of bearing are collected from a public repository provided by Case Western Reserve University (CWRU) [15]. In Figure 2.1, the experimental testbed is illustrated. As can be seen from this figure, the setup is composed of an induction motor of 2 horsepower, a dynamometer, and a transducer. The signals are collected from a drive end bearing with artificially seeded faults with the help of accelerometers mounted on the housing of the induction motor. With the help of dynamometer, several motor loads were applied while recoding the signals, as a result, variation in the motor shaft speeds was also observed, i.e., 1722-1797 Revolutions Per Minutes (RPMs). The signals are collected with a sampling frequency of 12 KiloHertz (kHz). The collected signals were associated with four types of health conditions of bearings, i.e., Normal Type (NT), Inner Raceway Type (IRT), Outer Raceway Type (ORT), Ball Type (BT). The details of this dataset are given into Table 2.1.

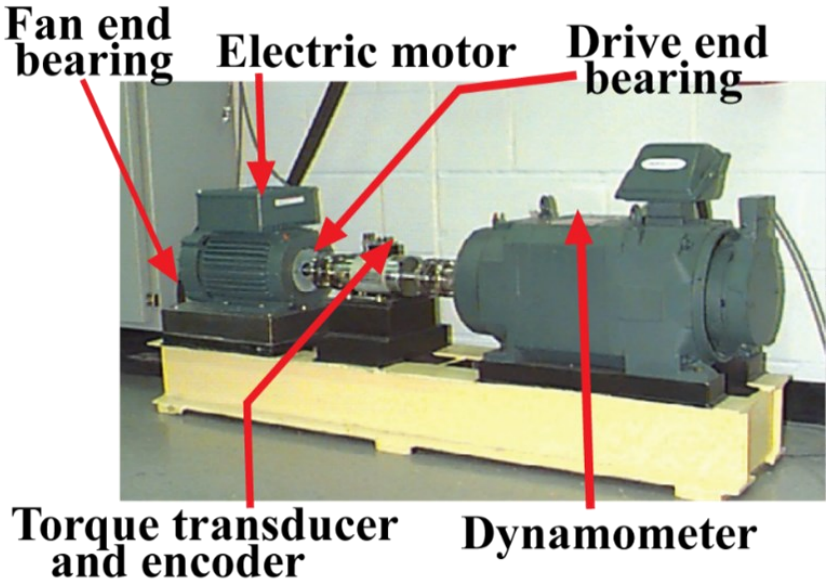


Figure 2.1: CWRU bearing testbed [15] for collecting vibration signals.

Table 2.1: Details of the dataset collected from CWRU bearing data bank used in case study 1.

	Health Type	Shaft Speed (RPM)	Load	Crack Size Length (inches)
Dataset 1	NT	1797	0	-
	IRT		0	0.007
	ORT		0	0.007
	BT		0	0.007
Dataset 2	NT	1772	1	-
	IRT		1	0.007
	ORT		1	0.007
	BT		1	0.007
Dataset 3	NT	1750	2	-
	IRT		2	0.007
	ORT		2	0.007
	BT		2	0.007

Before feeding to the proposed diagnostic model, we ensured that every health type has equal number of samples, there is no missing value in it. Thus, an ideal experimental test case is considered to analyze the performance of our algorithm.

2.2 Dataset from Self-designed Testbed

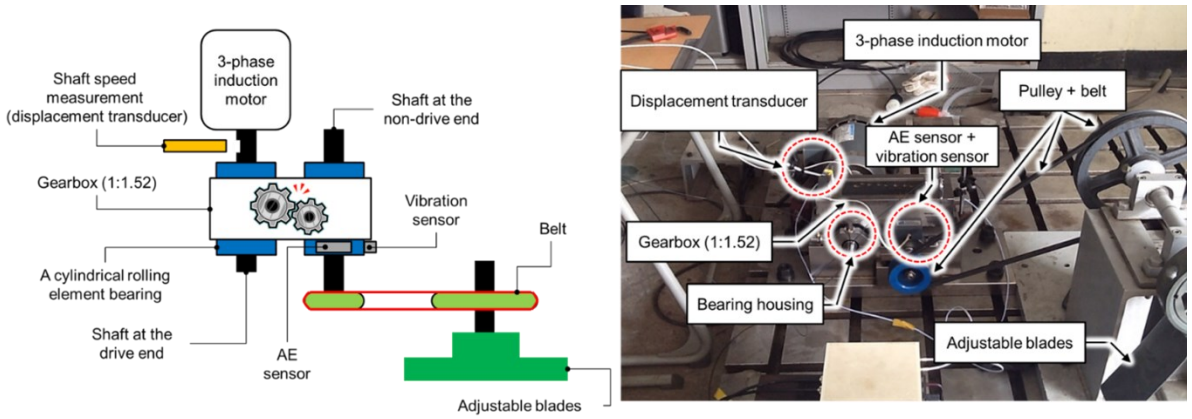


Figure 2.2: Self-designed testbed for collecting vibration signals.

Another test is conducted using vibration signals acquired from a self-designed test rig operated at three different motor speeds of 300, 400, and 500 RPMs. To conduct this test, a low-speed fault simulator was designed by using a cylindrical roller bearing (FAG NJ206-ETVP2) [16]. To capture the vibration signals, a wide-band vibration sensor is utilized at a sampling rate of 65536 Hz [17]. As illustrated in Figure 2.2, the whole set up is composed of two shafts, i.e., a drive end shaft, and a non-drive end shaft. A gearbox with a reduction ratio of 1.52:1 is used to connect these two shafts. At the non-drive end shaft of a three-phase induction motor, a displacement transducer is enclosed to measure the shaft speed. For 0 – 10 kHz frequency response, this transducer has the sensitivity range of +0 to -3 dB [16]. Hence, the vibration signals were recorded under three different motor speeds i.e., 300,400, and 500 RPM [17]. For the condition monitoring of bearing faults, three different types of seeded bearing defect were produced, as illustrated in Figure 2.4. For creating these artificial defects, a diamond cutter bit was used to generate cracks on the bearing surface. Therefore, the recorded signals were associated with four types of health conditions, i.e., Normal Type (NT), Inner Raceway Type (IRT), Outer Raceway Type (ORT), and Roller Type (RT). Before providing data to the model, it was ensured that every health type has an equal number of samples and that there are no missing values in it. The details of this dataset are listed in Table 2.2.

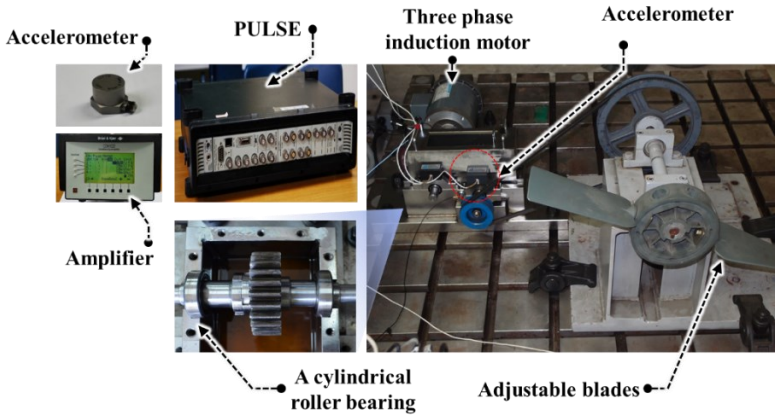


Figure 2.3: A snapshot of the real experimental testbed.

Figure 2.4: Fault types: (a) Outer Raceway Type (ORT), (b) Inner Raceway Type (IRT), and (c) Roller Type (RT).

Table 2.2: Details of the dataset from self-designed testbed used in case study 2.

	Health Type	Shaft Speed (rpm)	Crack Size Length (mm)
Dataset 1	NT	300	-
	IRT		6
	ORT		6
	RT		6
Dataset 2	NT	400	-
	IRT		6
	ORT		6
	RT		6
Dataset 3	NT	500	-
	IRT		6
	ORT		6
	RT		6

Chapter 3

Condition Monitoring of Bearing using Fast Discrete Orthogonal Stockwell Transformation Coefficient and Genetic Algorithm

3.1 Introduction

To get some meaningful insights from the signals to create a data-driven solution for the CM-FD of rolling element bearing, throughout these years, researchers have relied upon several signal processing techniques, such as, Fast Fourier Transformation (FFT) [18], Empirical Mode Decomposition (EMD) [19], Energy Entropy (EE) [20], Wavelet Packet Decomposition (WPD) [21], Empirical Wavelet Transformation (EWT) [22], Variational Mode Decomposition (VMD) [23], Continuous Wavelet Transform (CWT) [24], etc. These approaches provide satisfactory performance under static working conditions of rotary machines. However, due to several factors, i.e., friction among components of a machine, miss alignment, and noise, the acquired vibration signals acquired from bearing are non-linear, and non-stationary in nature, which create difficulties to extract and analyze the fault feature information [1,11,25,26]. Specifically, via the popular feature extraction methods which analyzes features from time domain, frequency domain, or time-frequency domain, it becomes very difficult to identify the fault characteristics for variable working conditions

[10,27,28]. To analyze such complex signals these signal processing techniques suffer some limitations given as follow:

(1) In practice bearing signals are prone to inconsistencies due to several factors which makes them difficult to analyze. Therefore, if there is a little variation in the given CM-FD scenario, the adopted signal processing techniques suffer to extract fault characteristics, properly.
(2) Wavelets transform based signal analysis usually experience the dilemma of appropriate Mother Wavelet (MW) selection and suffer from the drawbacks of poor noise immunity.

Therefore, it is inevitable to come up with a new and effective signal processing technique through which fault signature exploration can become reliable during the CM-FD of bearings for different speed conditions [29,30]. Recently, in this regard, Stockwell Transform (ST) is used for the fault signature identification and localization [31] for the following major advantages:

(1) It has better immunity to ample noise.
(2) It is free from the MW selection.
(3) It can obtain good resolutions from the signals both at low and high frequencies. Thus, it has built in adaptive ability to tackle inconsistencies in the observed data.

However, due to the redundant nature for selecting Gaussian windows ST suffers from a computational complexity of $\mathrm{O}\left(N^2 \log N\right)$. Therefore, in this study, a non-redundant faster variant of ST, namely known as Faster Discrete Orthogonal Stockwell Transform (FDOST) [32] with a computational complexity of $\mathrm{O}\left(N \log N\right)$, is proposed to explore unique pattern for a given fault. In this regard, first, FDOST co-efficient are computed from the vibration signals which are then utilized to extract statistical parameters from both time-frequency magnitudes and their corresponding phase angle information.

After the application of signal processing techniques, CM-FD pipeline normally utilize feature extraction and selection step. Generally, first, the statistical features are extracted from the processed signals, and afterwards, useful features are selected out of the whole feature pool through feature selection techniques which contain discriminant information for different fault types.

In this study, we have considered a Genetic Algorithm (GA) based feature selector to find out the most suitable features for analyzing the dataset. To decide the best classifier, we need to adapt a simple, yet powerful model, and which does not get affected by the heteroscedasticity

[33] of data. Therefore, k-Nearest Neighbor (k-NN) is considered as the classifier. Thus, in a nutshell, the contributions of this chapter can be summarized as follows:

(1) To capture the information of variable working conditions from the vibration signals both at low and high frequencies, a computationally advanced version of ST called FDOST is proposed as the signal preprocessing step. First, statistical parameters are extracted as features from the time-frequency magnitudes and their corresponding phase angles of the FDOST coefficients. The extracted statistical features are then arranged as a feature matrix which can be regarded as input in the proceedings step. Thus, a carefully curated statistical feature pool extracted from unique FDOST patterns for different types of bearing faults is proposed in this study, which is helpful for boosting up the classification performance of the subsequent classifier.

(2) To find out the best statistical properties (feature attributes) from the statistical feature pool, a GA based feature selector is designed to analyze the health conditions of the bearing with k-NN classifier.

Finally, to validate the proposed model, 2 different bearing datasets have been considered to conduct experimental case studies, among which, one is obtained from the public repository of CWRU [15], and other one is collected from a self-designed test bed which has already been discussed into **Chapter 2**. The performance of the signal processing step, feature selection process, and the classifier has been verified with several comparisons. The complete organization of this chapter can be summarized as follows: **Section 3.2** gives the theoretical and mathematical descriptions of the necessary backgrounds, **Section 3.3** discusses about the proposed methodology in a step-by-step procedure, **Section 3.4** highlights the experimental analysis with discussion, and **Section 3.5**; finally concludes this research work.

3.2 Technical Background

This section discusses the technical details of the Fast Discrete Orthogonal Stockwell Transformation (FDOST), Genetic Algorithm (GA), and k-Nearest Neighbor (k-NN) algorithm. The details related to the integration of these techniques into the proposed diagnosis framework have been discussed into the **proposed method** section.

3.2.1 Fast Discrete Orthogonal Stockwell Transformation

Stockwell Transformation (ST) is a multi-resolution evaluation-centered method which delivers frequency domain transformation like Fourier Transform (FT) [34] introduced by R.G Stockwell [35]. Initially it was defined to capture good resolutions from the signals both at low frequencies (wide window size), and high frequencies (small window size) like the CWT with a Gaussian window [36]. Moreover, it can also provide phase information associated with a signal, where CWT fails. The ST spectrum provides local phase information

which is helpful in defining distinct clusters for different classes [37]. However, the initially proposed algorithm with the Gaussian window had redundancy and created a high computational cost $\mathrm{O}(N^2 \log N)$ [37]. Later, a significant number of studies helped to resolve this issue [38]. In this study, one of the most efficient variation of ST proposed by U. Battisti et al. [39] is considered. This version of the algorithm offers a unified setting with a different admissible window to reduce the computational complexity to $\mathrm{O}(N \log N)$ while calculating the ST coefficients [39].

The ST of a function can be defined as

$$S(\tau, f) = \int_{-\infty}^{+\infty} h(t) \frac{|f|}{\sqrt{2\pi}} \exp\left(-\frac{(\tau - t)^2}{2f^2}\right) \exp(-i2\pi f t)\, dt \tag{3.1}$$

Where f is the frequency, t, and τ are the time variables, and $|f|/\sqrt{2\pi}$ is the normalizing coefficient factor. This continuous representation of ST can be represented into a discrete form by the following equation

$$S[j,n] = \sum_{m=0}^{N-1} H(m+n) \exp\left(-\frac{2\pi^2 m^2}{n^2}\right) \exp\left(i\frac{2\pi m j}{N}\right) \text{ for } n \neq 0 \tag{3.2}$$

Here to calculate the Discrete ST (DST), the following equivalence parameters are considered, i.e., $\int \to \sum, f \to n$, and $\tau \to j$. In Equation (2), $H[.]$ is the Discrete Fourier Transform (DFT) of $h[.]$, which is the FT in Equation (3.1). The Equation (3.2) can be simplified as

$$S[j,n] = \frac{1}{N} \sum_{m=0}^{N-1} h[k] \text{ where } k = 0, T, 2T, ..., (N-1)T \tag{3.3}$$

Where $h[k]$ is the discrete representation of $h(t)$. However, the Equation (3.2) can generate N^2 number of coefficients for a signal of length N. Thus, it creates a computational complexity of $\mathrm{O}(N^3)$. This high redundancy can be reduced by introducing N number of orthogonal basis vectors for the calculation of the number of coefficients for ST. This approach is established as Discrete Orthogonal Stockwell Transform (DOST). Mainly by the values of the 3 parameters, i.e., ν (center of a frequency band), β (width of the frequency band), and τ (location in time); k^{th} basis vectors of DOST can be represented as follows:

$$D[k]_{[\nu,\beta,\tau]} = \frac{1}{\sqrt{\beta}} \sum_{f=\nu-\beta/2}^{\nu+\beta/2-1} \exp\left(i\frac{2\pi k f}{N}\right) \exp\left(-i\frac{2\pi \tau f}{\beta}\right) \exp(-i\pi\tau) \tag{3.4}$$

Therefore, the DOST coefficient can be calculated as the inner product of the basis vector $D[k]_{[\nu,\beta,\tau]}$, and the input signal $h(k)$.

$$S_{[\nu,\beta,\tau]} = \left\langle D[k]_{[\nu,\beta,\tau]}, h(k) \right\rangle \text{ where } k = 0,1,2....,(N-1) \tag{3.5}$$

In Equation (3.5), the DOST coefficients are decreased to N with a computational complexity of $\mathrm{O}(N^2)$. However, to reduce it further, a faster technique is proposed by using the advantage of FFT in [38]. This approach reduced the computational complexity to $\mathrm{O}(N \log N)$ by using a fixed window size. Therefore, to introduce this Faster Discrete Orthogonal Stockwell Transform (FDOST), the coefficients are derived from Equation (3.5) as follows:

$$\begin{aligned} S_{[\nu,\beta,\tau]} &= \frac{1}{\sqrt{\beta}} \sum_{k=0}^{N-1} \sum_{f=\nu-\beta/2}^{\nu+\beta/2-1} \sum_{f=\nu-\beta/2}^{\nu+\beta/2-1} \exp\left(i\frac{2\pi k f}{N}\right) \exp\left(-i\frac{2\pi\tau f}{\beta}\right) \exp(-i\pi\tau) h(k) \\ &= \frac{1}{\sqrt{\beta}} \sum_{f=\nu-\beta/2}^{\nu+\beta/2-1} \exp(-i\pi\tau) \exp\left(-i\frac{2\pi\tau f}{\beta}\right) H[f] \end{aligned} \tag{3.6}$$

where $H[f]$ is the DFT of $h[k]$. To improve this representation, U. Battisti et al. [39] proposed a generalized window (φ) dependent basis function which can be defined as

$$E^{\varphi}_{[\mathrm{p,j}]}(k) = \frac{1}{\sqrt{\beta(p)}} \sum_{j=0}^{\beta(p)-1} \left[C^{\varphi}_{[\mathrm{p,j}]}(\nu(p)) \right]^{-1} \exp\left(2\pi i \left(\beta(p)+j\right) \left(\frac{k}{N} - \frac{\tau}{\beta(p)} \right) \right) \tag{3.7}$$

were, p denotes the frequency bands determined by ν , and β . The coefficient of FDOST can be calculated as

$$S^{\varphi}_{[p,\tau]} = \left\langle h(k), E^{\varphi}_{[p,\tau]} \right\rangle = \left\langle g, D[k]_{[p,\tau]} \right\rangle = \left\langle F^{-1} R^{\varphi} H, D[k]_{[p,\tau]} \right\rangle \tag{3.8}$$

were, F^{-1} is the Inverse Fourier Transform (IFT), H is the DFT of $h(k)$, R^{φ} is the sequence function, and $D[k]_{[p,\tau]}$ is the modified basis vector for calculating the coefficient of FDOST. Therefore, the updated Equation (3.4) for calculating the k^{th} basis vectors of FDOST with generalized window can be represented as follows:

$$D[k]_{[\mathrm{p},\tau]} = \frac{1}{\sqrt{\beta(p)}} \sum_{j=0}^{\beta(p)-1} \exp\left(i2\pi \left(\beta(p)+j\right) \frac{k}{N} \right) \exp\left(-i\frac{2\pi\tau j}{\beta(p)} \right) \tag{3.9}$$

The process of calculating the FDOST coefficient is illustrated into Figure 3.1 for a visual understanding.

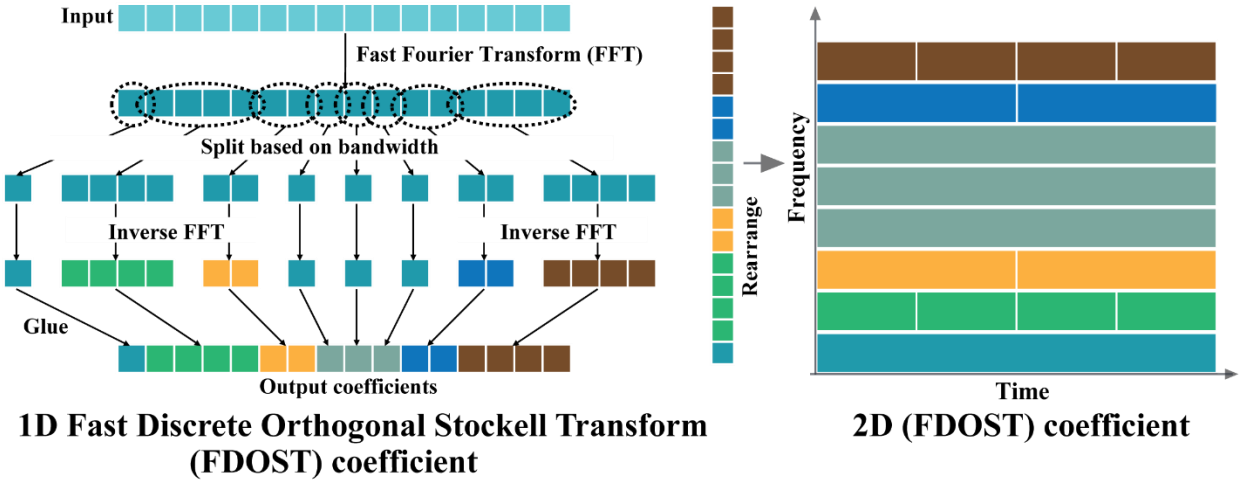


Figure 3.1: A visualization of FDOST coefficient calculation process.

3.2.2 Genetic Algorithm

Based on the evaluation theory, i.e., selection, crossover, mutation and replacement, GA determines the best combination of features with the most intrinsic class-wise information. To find the best feature combination that creates separability among classes, the Degree of Class Separation (DCS) is calculated by Equation (3.10).

$$\mathrm{DCS} = \frac{ICS}{WCC}, \tag{3.10}$$

where ICS is the Inter Class Separability parameter used to define the distance among different classes. Similarly, WCC is the Within Class Closeness that defines the closeness of the features within the same class. The Euclidian distance, given in Equation (3.11), is used to find the distance between two vectors.

$$D_{x,y} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.11}$$

A high rating for DCS is achieved when ICS is maximized and/or WCC is minimized. The ICS is determined based on the average distance of each feature vector of various classes with Equation (3.12).

$$\mathrm{ICS} = \frac{1}{{}_N c_2 \cdot n_A \cdot n_B} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k=1}^{n_A} \sum_{l=1}^{n_B} D_{i,j} \tag{3.12}$$

Similarly, WCC is obtained through the average distance of each feature vector of the same class by Equation (3.13).

$$\mathrm{WCC} = \frac{1}{N \cdot n_A \cdot n_B} \sum_{i=1}^{N} \sum_{j=1}^{n_A} \sum_{k=1}^{n_B} D_{i,j} \tag{3.13}$$

In these equations, N defines the number of classes, and $D_{i,j}$ is the Euclidian distance, derived from (3.11).

3.2.3 k-Nearest Neighbor Algorithm

The k-Nearest Algorithm (k-NN) is one of the simplest algorithms of ML [40,41]. This algorithm starts to work by assuming that the similar things are close to each other as described into Figure 3.2. Utilizing the idea of similarity based on proximity or distance, k-NN calculates the distance between 2 points by different approaches, i.e., Euclidian [42], Manhattan etc. [43]. However, in general, among all these approaches, Euclidian is the most widely approach, which can be expressed as follows:

$$d(m,b) = \sqrt{\sum_{i=1}^{n}(m_i - b_i)^2} \tag{3.14}$$

Where m,b are two points in an n dimensional Euclidian space.

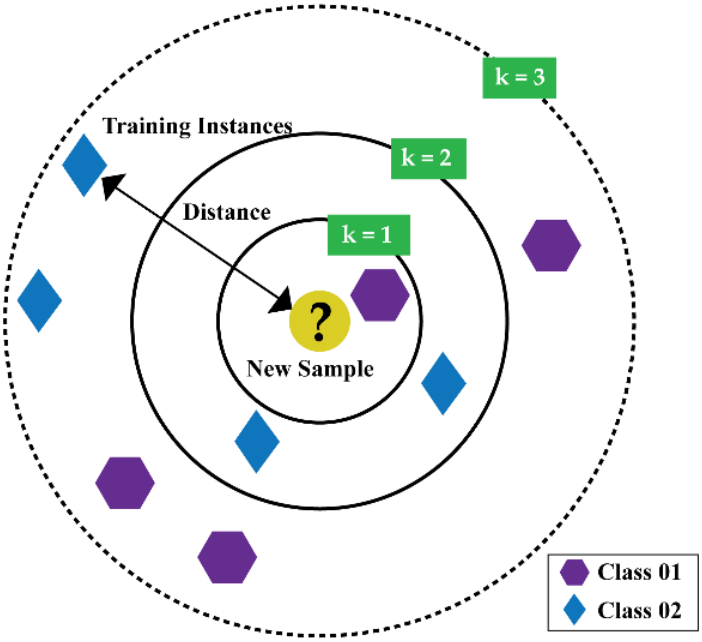


Figure 3.2: An illustration of k-NN algorithm.

To determine the chosen number of k (in another words, number of neighbors) for certain dataset, k-NN is needed to be run for several times with different values of k. From each run, select the value of k, which reduces the number of errors while making predictions. However, 4 things shall be considered while choosing the number of k.

(1) If the value of k is set to 1, the prediction turns out to be less stable.

(2) Inversely, if the value of is increased, the prediction becomes stable due to the majority voting/averaging. However, after a certain value for the target dataset, the number of errors will start to increase.

(3) The optimal k value can be determined by considering the square root of N, where N is the total number of samples. Then by using a grid search approach, from 1 to the optimal value of k, an error plot or accuracy plot is calculated to determine the most favorable k value.

(4) Usually, when considering the values for the range of k for the grid search, each value for k is set as an odd number, i.e., $[1,3,5,7,9.....,optimal-k-value]$.Thus, for the scenario while majority voting is necessary to determine the prediction, the tiebreaker becomes easy.

3.3 Proposed Method

The block diagram of the proposed method is illustrated into Figure 3.3. Based on this diagram, the core steps of the proposed algorithm are described into the following subsections.

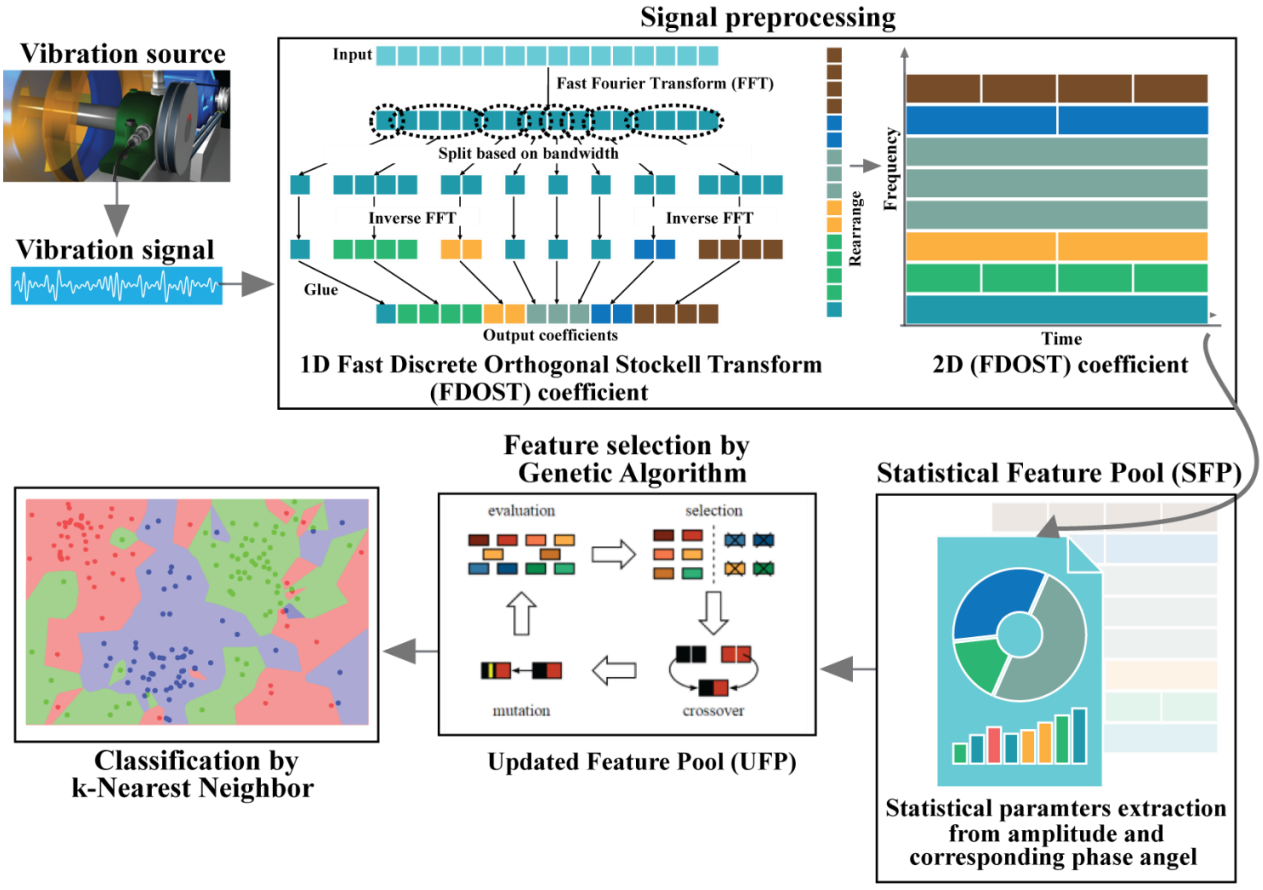


Figure 3.3: Block diagram of the proposed model for bearing fault diagnosis.

3.3.1 Data Preprocessing by Fast Discrete Orthogonal Stockwell Transformation (FDOST)

The vibration signals from bearing or any rotating machineries contains the information related to the different components with additive noise from the surrounding [4]. Thus, such signals are complex which possess non-stationary behavior. Therefore, it is difficult to extract required information form these signals with traditional statistical analysis in either time, or frequency domain [2]. To handle this issue, FDSOT has been adapted as the preprocessing step in this work. First, the raw signals are segmented into smaller sizes by adjustable overlapping sliding window [10,44]. Each of these segments contains the datapoints from at least one revolution. Then, from the time-domain signal, the FFT is calculated. Then, from the FFT, all the bandwidths are split. Except the lowest frequency bandwidth, on all the split portions, the IFFT is performed. Finally, all these portions of the bandwidths are merged to form the FDOST coefficient in 1D. Finally, the 1D representations are arranged into a 2D representation. The visual explanations of the full process are already given into Figure 3.1. These coefficient values are complex in nature; therefore, it holds the time-frequency information along with the corresponding phase-angle information from the signal.

3.3.2 Statistical Feature Pool Configuration

After computing FDOST of each signal, we obtain a set of complex coefficients $S_{r,c}$, which can be represented as follows:

$$S_{r,c} = A_{r,c} e^{j\phi_{r,c}} \text{ where } r \text{ represents row, and } c \text{ represents column} \tag{3.15}$$

Here, A and ϕ in (3.15) are the corresponding magnitude and phase angle of $S_{r,c}$ [45]. As previously discussed, $S_{r,c}$ contains good resolution from low, and high frequencies of the signal. Therefore, in this study, a set of statistical parameters are obtained from both $A_{r,c}$, and $\phi_{r,c}$ matrices. Later, the maximum values are calculated from $A_{r,c}$, $A_{c,r}$, $\phi_{r,c}$,and $\phi_{c,r}$ to create 4 Significant Information Pools (SIPs). Thus, the maximum coefficients with respect to frequency, and time for $A_{r,c}$, and $A_{c,r}$, and the maximum coefficients with respect to frequency ratio, and phase angle for $\phi_{r,c}$,and $\phi_{c,r}$ are determined from the 2D FDOST representation. Finally, from each of these SIPs, 5 statistical parameters (i.e., mean, Standard Deviation (STD), kurtosis, skewness, and Root-Mean-Square (RMS)) are extracted as features. With these 20 features, for each signal, a Statistical Feature Pool (SFP) is designed for further analysis. The details of these 20 features are enlisted in the Table 3.1.

Table 3.1: Feature attributes for the SFP configuration.

Input Definition	Feature Attributes				
	Mean	**Std**	**Kurtosis**	**Skewness**	**Rms**
$x = \max(A_{r,c})$	$\frac{1}{N}\sum_{n=1}^{N} x_n$	$\sqrt{\frac{\sum(x_i-\mu)}{N}}$	$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n-\bar{x}}{\sigma}\right)^4$	$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n-\bar{x}}{\sigma}\right)^3$	$\sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$
$x = \max(A_{c,r})$	$\frac{1}{N}\sum_{n=1}^{N} x_n$	$\sqrt{\frac{\sum(x_i-\mu)}{N}}$	$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n-\bar{x}}{\sigma}\right)^4$	$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n-\bar{x}}{\sigma}\right)^3$	$\sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$
$x = \max(\phi_{r,c})$	$\frac{1}{N}\sum_{n=1}^{N} x_n$	$\sqrt{\frac{\sum(x_i-\mu)}{N}}$	$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n-\bar{x}}{\sigma}\right)^4$	$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n-\bar{x}}{\sigma}\right)^3$	$\sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$
$x = \max(\phi_{c,r})$	$\frac{1}{N}\sum_{n=1}^{N} x_n$	$\sqrt{\frac{\sum(x_i-\mu)}{N}}$	$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n-\bar{x}}{\sigma}\right)^4$	$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n-\bar{x}}{\sigma}\right)^3$	$\sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$

3.3.3 Feature Selection by Genetic Algorithm (GA)

Through different search techniques, i.e., complete, sequential, and heuristic searches, the optimal subset of features can be determined [46,47]. However, the brute-force mechanism of the complete search adds additional computational complexity, and the sequential approach gives no guarantee that the best optimal feature subset is selected. On the other hand, using a heuristic search, the genetic approach with GA provides a balance between optimal selection and computational complexity. Therefore, in this study, GA is considered for the selection of the optimal feature subset from the SFP. Thus, after selecting the best features, we have prepared the Updated Feature Pool (UFP).

3.3.4 Classification by k-Nearest Neighbor (k-NN)

Finally, the UFP is fed to the k-NN to identify the types of health conditions. There are two main reasons behind selecting k-NN as a classifier in this study:

(1) It is a non-parametric clustering analysis-based model.

(2) Parameter's tuning is not a concern for k-NN as it is a non-parametric algorithm. Therefore, assumptions about the input data are not a concern while dealing with k-NN.

3.3.5 Performance Evaluation Criteria

To assess the performance of the proposed model two measurement matrices have been considered, i.e., (a) Accuracy Score (AS) [48], and (b) Confusion Matrix (CM) [49]. The AS can be calculated by the following equation,

$$AS = \left[\frac{(TP+TN)}{(TP+FP+TN+FN)}\times 100\right]\% \tag{3.16}$$

Here, TP, TN, FP, FN refers to True Positive, True Negative, False Positive, and False Negative, respectively. Moreover, to determine the best value of k for k-NN, a range of k-values are considered first. While considering the values within the range, only the odd values are contemplated. Therefore, if there are in total N samples present in the dataset, the range of k - values will be $\left[1,3,5,.....,\sqrt{N}\right]$. Then, a grid-search approach with K fold Cross Validation (K-CV) is used while evaluating the performance of the classifier, whereas the values of K ranges from $\left[1,2,3,.....,10\right]$. The upper range of K is arbitrary, however, the most common practice in ML is to use 10-CV by considering K=10. Thus, the best AS matrix is decided on the best choice of k (k of k-NN) along with the best performing K-fold.

3.4 Experimental Results Analysis

This section presents the description of experimental in a step-by-step manner for two separate case studies performed on two separate datasets. For every dataset, first, the performance of the classifier is analyzed for various load, and speed conditions. Later, to prove the robustness of the proposed model, the accuracy is compared with several popular state-of-art techniques as well.

3.4.1 Case Study 1 – CWRU Dataset

In this case, based on different load, and speed conditions, three different datasets were created, which has already been described in **Table 2.1 of Chapter 2**. Each of these datasets are divided into three subsets, i.e., training, testing, and validation. Training and validation datasets are used for training the model with 10-CV to figure out the best model configuration with highest performance. To generate unbiased results, model performance is calculated in terms of accuracy of the best model after 10-CV by using totally unseen test dataset. The train, test, and validation datasets ratios are highlighted in the table given below.

Table 3.2: The train, test, and validation datasets ratios.

Dataset	Train (70%)		Test (30%)	Total samples	Sample/Health Type
	Training (80%)	Validation (20%)			
1	261 samples	66 samples	141 samples	468	117
2	261 samples	66 samples	141 samples	468	117
3	261 samples	66 samples	141 samples	468	117

First, from the raw time domain signals, the 2D FDOST coefficient matrices have been calculated. As discussed earlier, this will help to preserve the time-frequency information along with the corresponding angle for both low, and high frequency components of a signal. The analyzed FDOST 2D representations are given in Figure 3.4 – Figure 3.6 for both the datasets for visual understanding of the described phenomenon.

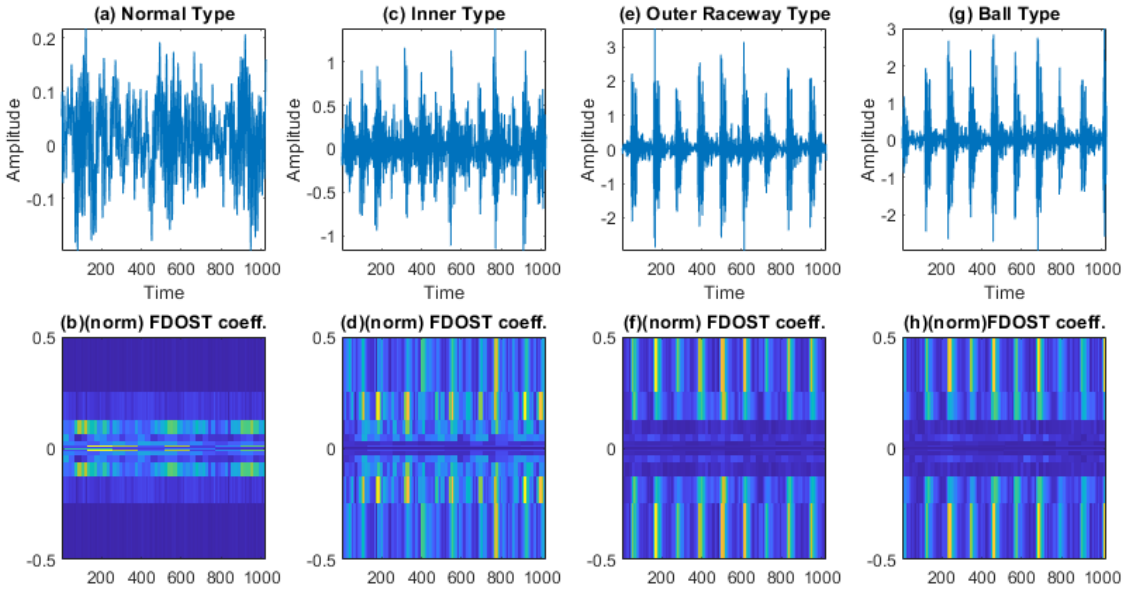


Figure 3.4: The visualization of the dataset 1 in terms of raw time domain signals and respective 2D FDOST coefficient matrix - (a, b): time domain, and 2D FDOST coefficient of NT, (c, d): time domain, and 2D FDOST coefficient of IRT, time domain, and 2D FDOST coefficient of ORT, and (g , h): time domain, and 2D FDOST coefficient of BT.

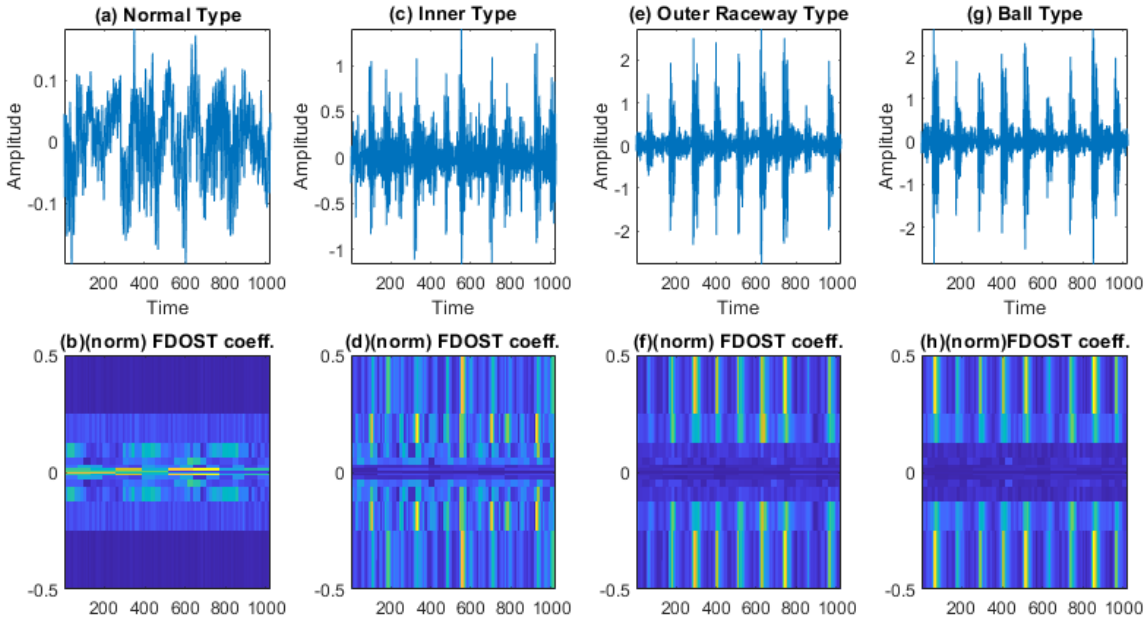


Figure 3.5: The visualization of the dataset 2 in terms of raw time domain signals and respective 2D FDOST coefficient matrix - (a, b): time domain, and 2D FDOST coefficient of NT, (c, d): time domain, and 2D FDOST coefficient of IRT, (e, f): time domain, and 2D FDOST coefficient of ORT, and (g , h): time domain, and 2D FDOST coefficient of BT.

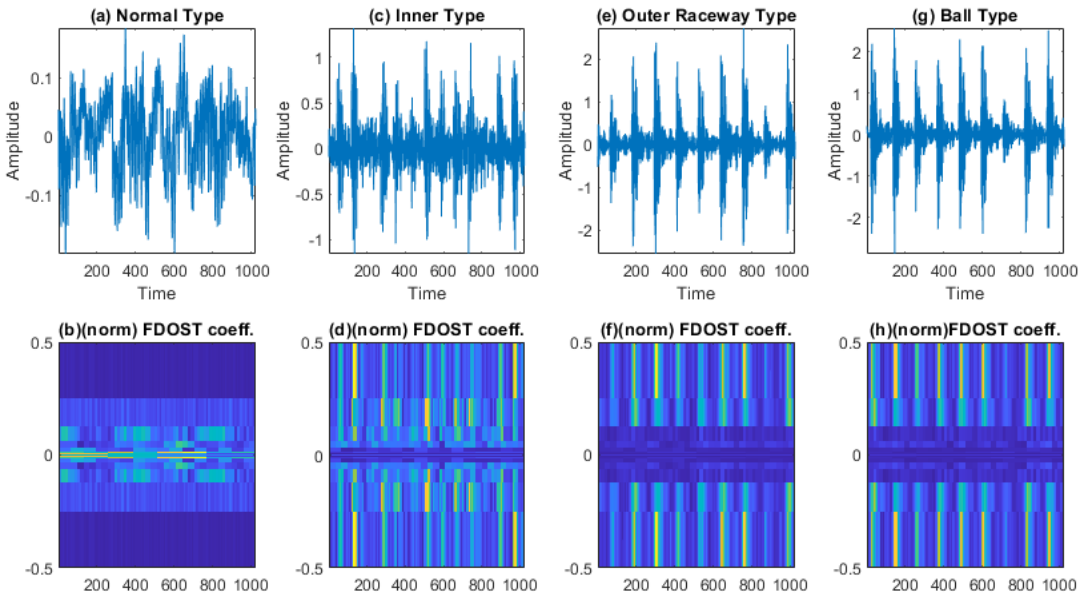


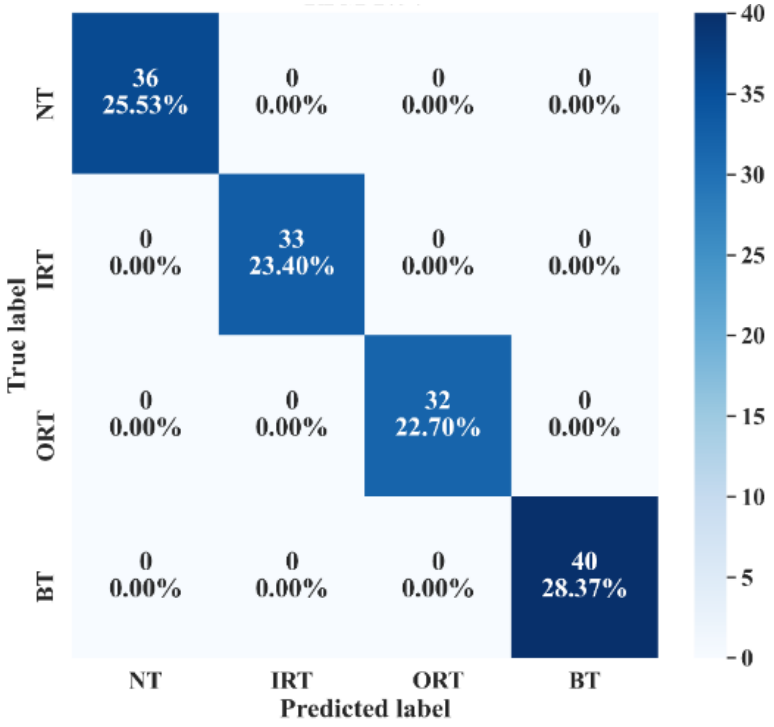
Figure 3.6: The analysis of dataset 3 - (a, b): time domain, and 2D FDOST coefficient of NT, (c, d): time domain, and 2D FDOST coefficient of IRT, (e, f): time domain, and 2D FDOST coefficient of ORT, and (g , h): time domain, and 2D FDOST coefficient of BT.

From Figure 3.4 – Figure 3.6, we can observe that time-frequency information's from FDOST are unique for each health condition. In addition, if we closely observe, we can understand that similar health condition from all the datasets exhibit similar type of pattern. The pattern of NT from datasets 1,2, and 3 are very identical. Similarly, the pattern of IRT is similar for 3 datasets. In the same way, it is observable and true for ORT, and BT. Moreover, among the patterns of NT, IRT, ORT, and BT, there is a clear distinction, by which, after observing these patterns, we can identify the individual health type. Therefore, it can be inferred that with the suggested signal preprocessing technique can extract unique patterns for a given health state if ever underlying dataset changes or some variations within the observed dataset occur due to the change in working condition of the machinery. In the next step, a SFP is formed by extracting 20 features according to the details of Table 3.1. Later, GA is applied to get the most relevant features for the given task. We are considering all the important feature which are identified as important by GA. Thus, after forming the UFP, the training data from all the datasets are normalized to train 3 separate model. Each model is trained by 10-CV. With this grid search approach, best model, with the most suitable number of neighbors (k of k-NN) is picked to calculate the performance of the model in terms of accuracy. The details of these tests along with the accuracy score are highlighted into Table 3.3.

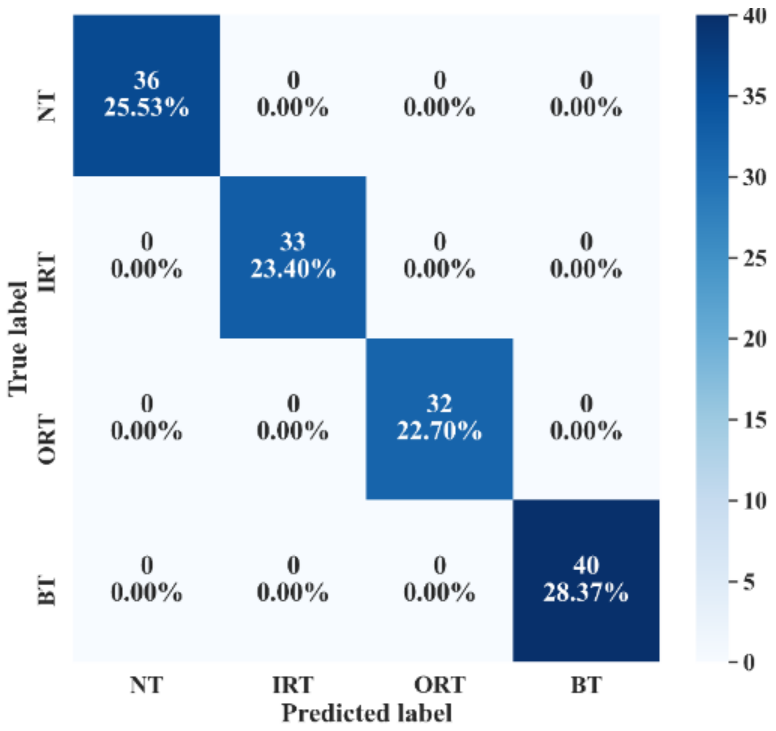
Table 3.3: Diagnostic performance of the proposed model.

Model	Dataset	UFP	Quantity of UFP	Common feature attributes	Best k	Accuracy (%)
1	1	'F2', 'F14', 'F3', 'F11', 'F10', 'F16', 'F9', 'F12', 'F1', 'F19'	10	'F2', F10', 'F12', 'F14', 'F16'	5	100
2	2	'F2', 'F14', 'F10', 'F16', 'F12'	6		3	100
3	3	'F1', 'F12', 'F16', 'F10', 'F14', 'F2'	6		3	100

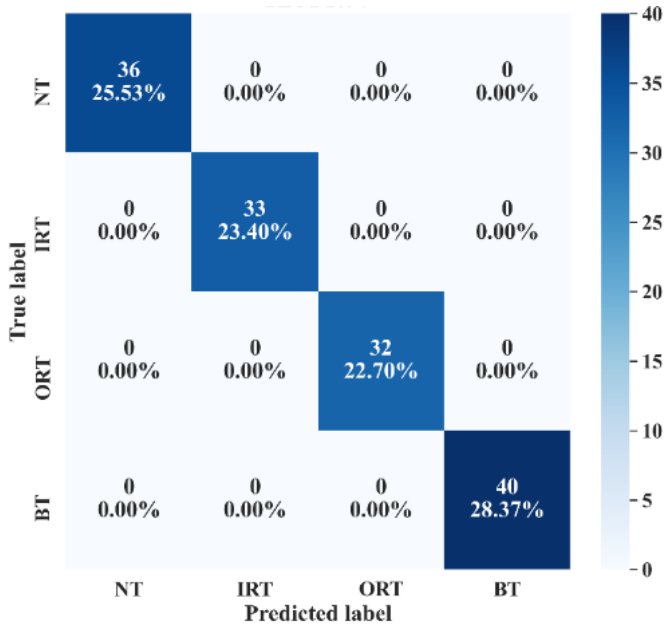
From Table 3.3, we can see that, for model 1, the selected features by GA are 12. However, the number of attributes in UFP for both models 2, and 3 are 6. For model 1, dataset 1 is used for training, as well as for testing. Similarly, for model 2, dataset 2 is used, and for model 3, dataset 3 is used. For each model, 70% data is used for training, and rest 30% is used for testing. The details about these train, test split is discussed into Table 3.2. All these models achieved 100% accuracy in total. The confusion matrices of these 3 models are given in Figure 3.7 that demonstrate the class-wise accuracy of the models when tested with the respective datasets.



(a)



(b)



(c)

Figure 3.7: The confusion matrices of the three models that demonstrates the class-wise test accuracies, i.e., (a) dataset 1, (b) dataset 2, and (c) dataset3.

It can be inferred from the figure that the accuracy of each model is 100% for the individual class. That means, each model achieves 100% True Positive Rate (TPR), and 100% True Negative Rate (TNR), which makes the performance of this model satisfactory with this dataset. In addition, in depth analysis of Table 3.3 revels that in FFP of all the three datasets there are 5 common features, i.e., 'F2', 'F10', 'F12', 'F14', and 'F16'. Therefore, it is safe to consider that these 5 features directly influence the final outcomes of all the three models while dealing with three different datasets. Thus, based on this analysis, to evaluate the performance of the proposed generalized CM-FD model for bearing, from each dataset, we have considered these 5 features to create the generalized version of the k-NN. To prove the robustness of this model, we have performed 3 sperate tests. Among them, for test 1, we trained, and validated (train: valid = 80:20) the model only with the samples of dataset 1, and then tested the trained model with dataset 2, and dataset 3. Similarly, test 2, we have used dataset 2 for training, and dataset 3 and 1 for testing. Likewise, for test 3, datasets 1, and 2 are used for testing while dataset 3 is used for training the model. The details of this analysis are enlisted into Table 3.4.

Table 3.4: Diagnostic performance of the invariant model.

Test	Training dataset	Test dataset	Performance measurement (%) – Avg.		
			TPR	TNR	Accuracy
1	1	2,3	100	100	100
2	2	3,1	100	100	100
3	3	1,2	100	100	100

All these models achieve 100% classification accuracy on the test datasets. Additionally, to prove the robustness of the proposed bearing CM-FD model few comparisons have been presented in Table 3.5 where the compared studies are conducted for the similar working environment of the machinery.

Table 3.5: Comparison Analysis.

Method no.	Ref.	Signal processing	Feature extraction	Feature selector	Classifier	Invariance capability	Accuracy (%)	Performance gap
1	[50]	No	Time domain features: waveform length, slope sign changes, simple sign integral and Wilson amplitude in addition to established mean absolute value and zero crossing	Laplacian Score (LS)	(Linear Discriminant Analysis) LDA, (Naïve Bayes) NB, SVM	No	99.6	0.4
2	[51]	Genetic Programming (GP)	Evolved features by GP stages	GP based filtering	k-NN	Yes	99.8	0.2
3	[52]	No	Local Binary Pattern (LBP)	No	Artificial Neural Network (ANN)	Yes	99.5	0.5

3.4.2 Case Study 2 – Dataset from the Self Designed Testbed

A second experiment was performed to confirm that the proposed model provides satisfactory results if evaluated using totally different dataset. In this case also, like case study 1, based on different load and speed conditions the data is divided into i.e., training, testing, and validation subsets as given in Table 3.6. Moreover, 10-CV is used to determine the best performing model. Test dataset remains totally unseen to the model to evaluate the model performance based on accuracy.

Table 3.6: Details of the data division.

Dataset	Train (70%)		Test (30%)	Total samples	Sample/Health Type
	Training (80%)	Validation (20%)			
1	717 samples	179 samples	384 samples	1280	320
2	717 samples	179 samples	384 samples	1280	320
3	717 samples	179 samples	384 samples	1280	320

Like the previous case study, first, from the raw time domain dataset, we have calculated the FDOST 2D coefficient matrices. The 2D FDOST representations for the three datasets are given in Figures 3.8 -3.10 for visual understanding of the described phenomenon.

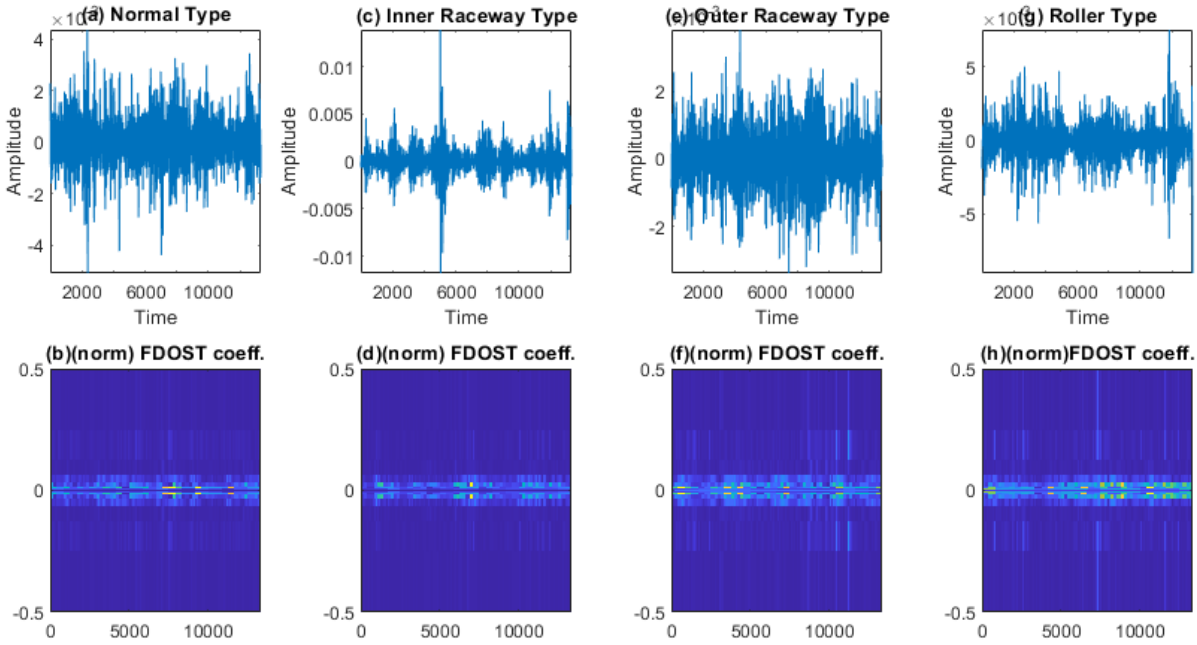


Figure 3.8: The visualization of the dataset 1 in terms of raw time domain signals and respective 2D FDOST coefficient matrix - (a, b): time domain, and 2D FDOST coefficient of NT, (c, d): time domain, and 2D FDOST coefficient of IRT, (e, f): time domain, and 2D FDOST coefficient of ORT, and (g , h): time domain, and 2D FDOST coefficient of RT.

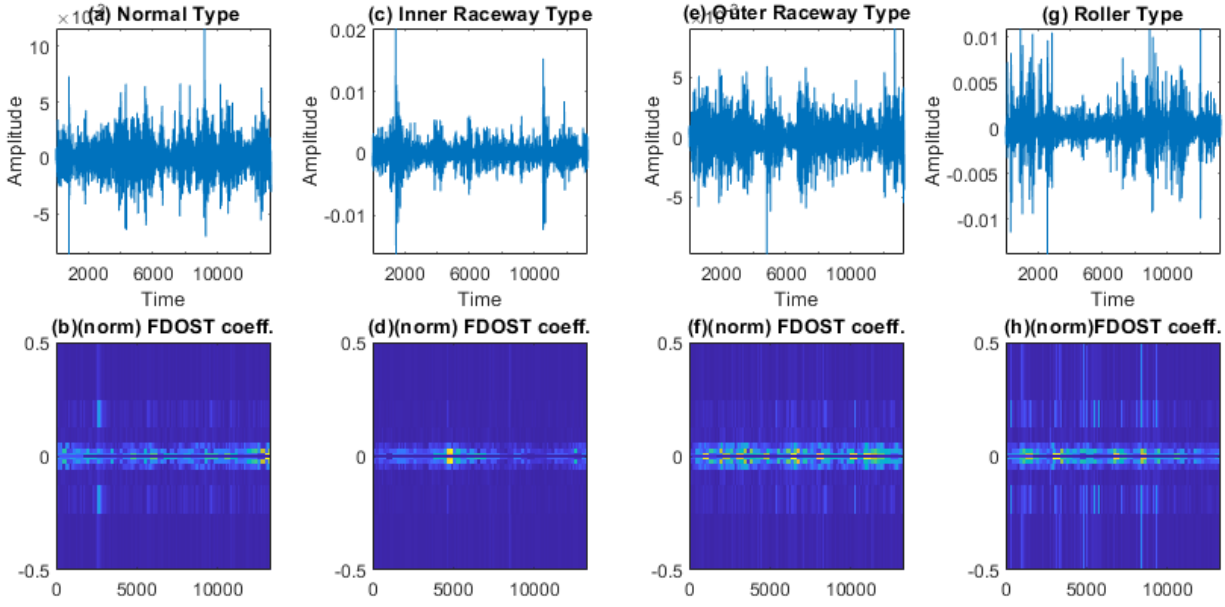


Figure 3.9: The visualization of the dataset 2 in terms of raw time domain signals and respective 2D FDOST coefficient matrix - (a, b): time domain, and 2D FDOST coefficient of NT, (c, d): time domain, and 2D FDOST

coefficient of IRT, (e, f): time domain, and 2D FDOST coefficient of ORT, and (g , h): time domain, and 2D FDOST coefficient of RT.

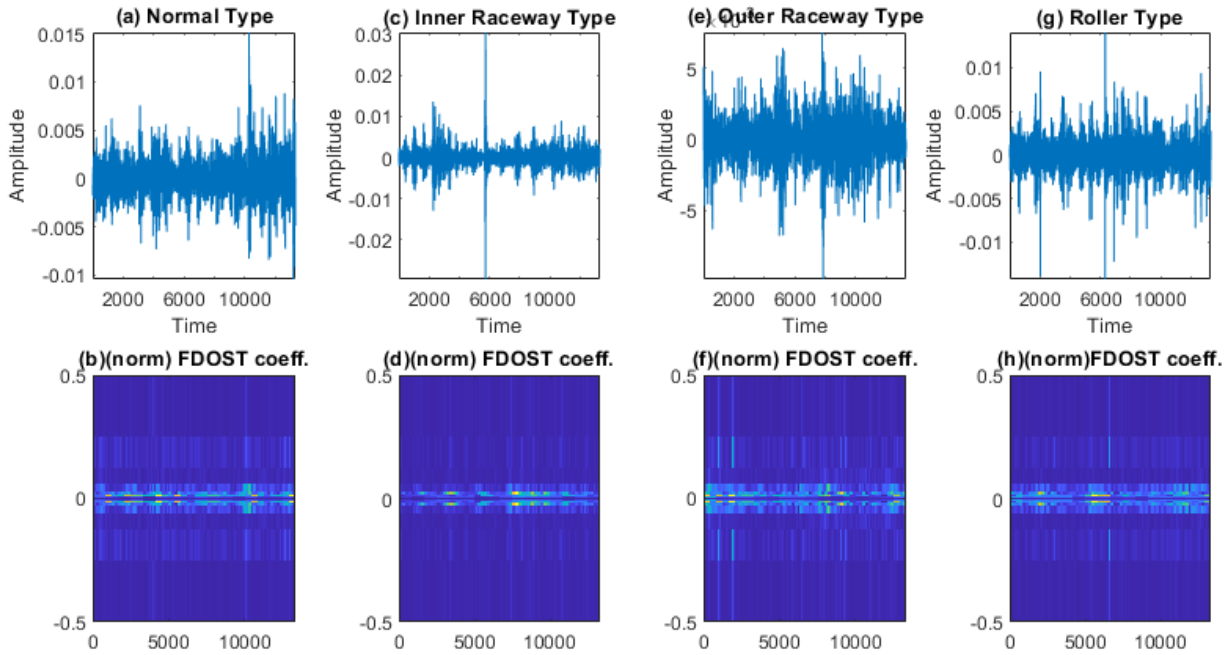


Figure 3.10: The analysis of dataset 3 - (a, b): time domain, and 2D FDOST coefficient of NT, (c, d): time domain, and 2D FDOST coefficient of IRT, (e, f): time domain, and 2D FDOST coefficient of ORT, and (g , h): time domain, and 2D FDOST coefficient of RT.

From Figure 3.8 – Figure 3.10, we can observe that for these datasets also time-frequency information's from FDOST are unique for each health condition. Moreover, if we closely observe, we can understand that similar health condition from all the datasets exhibit similar type of pattern. Although, the patterns in this case are not so vibrant as given in Figures 3.4 – Figure 3.6 because the sampling frequency of this dataset is at least 5 times higher than the previous dataset. Therefore, just based on the 2D FDOST analysis it cannot be confirmed that the classifier can produce satisfactory output under variable working condition of the machinery. Therefore, in this case, UFP will play a vital role in the development of a generalized model by properly selecting the important features for the task.

In the next step, a SFP is formed by extracting 20 features. After that, GA is applied on the feature pool to select the most relevant features. Once the UFP is obtained, the training data from all the datasets are normalized to train 3 separate model. Each model is trained by 10-CV. With this grid search approach, best model, with the most suitable number of neighbors (k) is picked to measure the classification accuracy. The details of these tests along with the accuracy score are highlighted into Table 3.7.

Table 3.7: Diagnostic performance of the proposed model.

Model	Dataset	UFP	Quantity of UFP	Common feature attributes	Best k	Accuracy (%)
1	1	'F3', 'F13', 'F17', 'F20', 'F9', 'F1', 'F11', 'F14', 'F7', 'F15', 'F4', 'F19', 'F12'	13	'F11','F13',' F14', 'F15','F19'	7	97.1
2	2	'F15', 'F14', 'F19', 'F11', 'F18', 'F13', 'F9', 'F3', 'F12', 'F16'	10		11	97.0
3	3	'F11', 'F19', 'F20', 'F1', 'F13', 'F15', 'F17', 'F14', 'F16'	9		7	96.1

Same as the previous case study, by considering the common feature attributes for creating the generalized model, three sperate tests were performed to evaluate the performance of the proposed model. The test 1, we train, and validate (train: valid = 80:20) the model only with all the samples of dataset 1, and then evaluate the performance with dataset 2, and dataset 3. Similarly, for test 2, we have used dataset 2 for training, and dataset 3, and 1 for testing. Likewise, for test 3, datasets 1, and 2 are used for testing while dataset 3 is used for training the model. All these tests achieve at least 97.0% classification accuracy on the test datasets. The details of this analysis are enlisted into Table 3.8.

Table 3.8: Diagnostic performance of the invariant model.

Test	Training dataset	Test dataset	TPR				TNR				Accuracy (%)
			NT	IRT	ORT	RT	NT	IRT	ORT	RT	
1	1	2	1.00	.93	.95	.98	.99	.99	.97	.99	97.0
		3	1.00	.95	.97	.98	.99	.99	.98	.99	97.8
2	2	3	1.00	.96	.97	.99	.99	.99	.98	.99	98.2
		1	1.00	.95	.95	.98	.99	.98	.98	.99	97.4
3	3	1	1.00	.95	.95	.98	.99	.98	.98	.99	97.1
		2	1.00	.95	.94	.97	.99	.99	.98	.99	97.0

3.5 Conclusions

This chapter proposes A 4-stage scheme to perform the fault diagnosis of bearing, i.e., **(1) data preprocessing:** a FDOST coefficient-based vibration signal preprocessing is proposed for capturing the invariant patterns from both time-frequency, and corresponding phase-angel information for variable speed, and load conditions, **(2) feature extraction:** statistical feature extraction is performed on FDOST coefficient to capture the significance from the invariant pattern of the preprocessed data, **(3) feature selection:** an evolutionary feature selection process is incorporated by introducing GA, and **(4) classifier:** a straight forward

k-NN classifier is used to classify the health conditions finally. The conducted case studies show proposed model can help to build a classifier for invariant working scenario.

However, this model has several limitations. The reason for selecting a feature attribute can be explained/understood by the proposed GA based feature selection. Thus, without the proper understanding of feature selection process, all the ML based classifiers only focus on the classification accuracy like a black box approach. Moreover, the feature selector cannot guarantee to remove colinear feature information. Thus, the classifier can be affected with the multicollinearity issue, which makes the model performance questionable. Therefore, in the next chapter, we have attempted to propose a solution by aiming an explainable fault diagnosis model in 2 steps, i.e., **(1)** introducing explain ability to the feature selection stage, and **(2)** incorporate explain ability to the classifier.

Chapter 4

An Explainable AI based approach for Condition Monitoring of Bearing

4.1 Introduction

In the previous Chapter, we have demonstrated that, after the application of signal processing techniques, CM-FD pipeline normally utilize feature extraction and selection step. Generally, first, the statistical features are extracted from the processed signals, and afterwards, useful features are selected out of the whole feature pool through feature selection techniques which contain discriminant information for different fault types. Likewise, a suitable 2D representation of the selected features is an appropriate choice when dealing with a DL algorithm to accomplish classification task. However, these approaches lack in the appropriate explanation of the feature selection process. Additionally, they are unable to explain the importance of each feature in the selected feature subset, as well as justification about their influence on the final output of the CM-FD pipeline is also missing. On the other hand, the ML based approaches adapted in the field of bearing fault diagnosis, completely suffer for the lack of feature and model explain ability . Thus, without a clear interpretation of model learning and working processes it difficult to debug the outcome of these diagnosis model. Moreover, these models are also not generalized, i.e., whenever, any of the underlying conditions changes regarding CM-FD process, changes are required in the suggested CM-FD approach to address the changes. This problem can be explained further by the sub-sequent points:

(1) The feature selectors available in the literature of bearing fault diagnosis are either the evolutionary algorithm-based approach [53,54] or the filter-based approach [55]. The evolutionary algorithms (i.e., GA) select features based on grid search and knap-snack mechanism [56] with gene cross-over, and mutation. Thus, the reason for selecting an individual feature attribute can never be explained or tracked down. In addition, both approaches rely on the technique of removing the redundant feature attributes from the original feature set. Moreover, several data compression techniques, i.e., principal component analysis [57], and different manifold learning techniques are also used to represent the data into lower dimension for diagnosis purpose [58]. However, these techniques are based on algebraic calculation and geometric projection to create the separability into the feature space. Therefore, explainability remains the problem for these linear, and non-linear dimensionality reduction-based techniques.

(2) Without the proper understanding of feature selection process, all the ML based classifiers only focus on the classification accuracy like a black box approach. This creates the adaptability issue for the model when the working environment changes a little.

In this study, to mitigate the limitations of the proposed model into **Chapter 3**, we have considered an Explainable Artificial Intelligent (XAI) model for bearing fault diagnosis which consists of a wrapper-based approach named Boruta, Spearman's Rank Correlation Coefficient (SRCC), and k-NN classifier with Shapley additive explanations. Boruta is built around a non-parametric classifier - Random Forest (RF). It first shuffles and permutes all the original feature attributes to create a randomized copy of original feature set. Then, this randomized set of features are added to the original feature set to perform the feature selection process by RF classifier [59]. Thus, by adding randomness to the feature attributes, this algorithm outperforms the conventional forward/backward selection-based wrapper approaches [60]. Subsequently, with the help of embedded RF classifier, the selection of every feature attribute can be explained. Before, feeding the selected features to the classifier, we have proposed an additional collinearity based feature filtration step by using SRCC [61,62] to obtain a bias-free accuracy measurement. This SRCC helps to remove the correlated feature attribute from the extracted statistical feature attributes. Thus, the explanations for selecting feature attributes give the opportunity to design a biasfree explainable ML model. To decide the best classifier, we need to adapt a simple, yet powerful model, and which does not get affected by the heteroscedasticity [33] of data. To satisfy all these conditions, in this study, we have considered k-Nearest Neighbor (k-NN) as a classifier. Finally, to interpret the decision of k-NN, we have proposed the kernel Shapley Additive Explanation (SHAP) [63] into our diagnosis model. There are 2 main advantages of this SHAP based model explanation.

(1) This interpretation of Shapley value is inspired from collaborative game theory scenario where contribution of each feature attribute the model performance is unequal but in cooperation with each other [64]. The Shapley value guarantees each feature attribute profits as much or more as they would have from performing independently.

(2) The SHAP kernel can provide a unique solution by satisfying local accuracy, missingness, and consistency on the basis of the original model, hence allows explainability of a model [63]. The proof of these properties will be discussed later into **Section 4.2.3** of this chapter.

Therefore, in a nutshell, the contributions of the proposed XAI model for bearing fault diagnosis can be summarized as follows:

(1) A wrapper-based feature selector named as Boruta is utilized to find the best features from the extracted statistical feature pool from the **previous Chapter**. This algorithm can justify the selection of each feature attribute with the help of an embedded RF classifier. Thus, the feature selection process is easily interpretable in the proposed bearing fault diagnosis model.

(2) In addition to interpretable feature selection process, a feature filtration technique is proposed by using SRCC to create a reduced feature set for classifier which can produce bias free results. Thus, it helps the classifier to avoid the multicollinearity trap.

(3) A correlation between the filtered feature pool and results of a non-parametric k-NN classifier is presented in this work, i.e., the predictions of k-NN are explained in context of SHAP values.

This chapter introduces a concept of explain ability for the first time in the field of bearing fault diagnosis in 2 steps: (a) incorporating explain ability of the feature selection process, and (b) interpretation of the ML classifier performance with respect to the selected features. Thus, it makes the proposed model a complete XAI based fault diagnosis state-of-art approach for bearing, which is applicable in real-world scenario. Like previous **Chapter**, to validate the proposed model, two bearing datasets have been considered, among which, one is obtained from the public repository of CWRU [15], and other one is collected from a self-designed test bed. The improvement in the performance of the proposed signal processing step, feature selection process, and classifier has been verified through several comparisons with sate of the art published studies. The complete organization of this chapter can be summarized as follows: **Section 4.2** gives the theoretical and mathematical descriptions of the necessary backgrounds, **Section 4.3** discusses about the proposed methodology in a step-by-step procedure, **Section 4.4** highlights the experimental analysis with discussion, and **Section 4.5**; finally concludes this research work.

4.2 Technical Background

This section discusses the technical details of the wrapper-based feature selector – Boruta, Spearman's Rank Correlation Coefficient (SRCC), and Shapley Additive Explanation (SHAP) for model interpretation. The details related to the integration of these techniques into the proposed diagnosis framework have been discussed into the **proposed method** section.

4.2.1 Wrapper based Feature Selector – Boruta

A wrapper approach usually uses a non-parametric classifier for the selection of important features from the complete feature pool. The reason behind considering a non-parametric classifier is to reduce the computational complexity [65] along with avoiding the multicollinearity problem [66]. Boruta is a wrapper-based approach that used Random Forest (RF) classifier for the feature selection process.

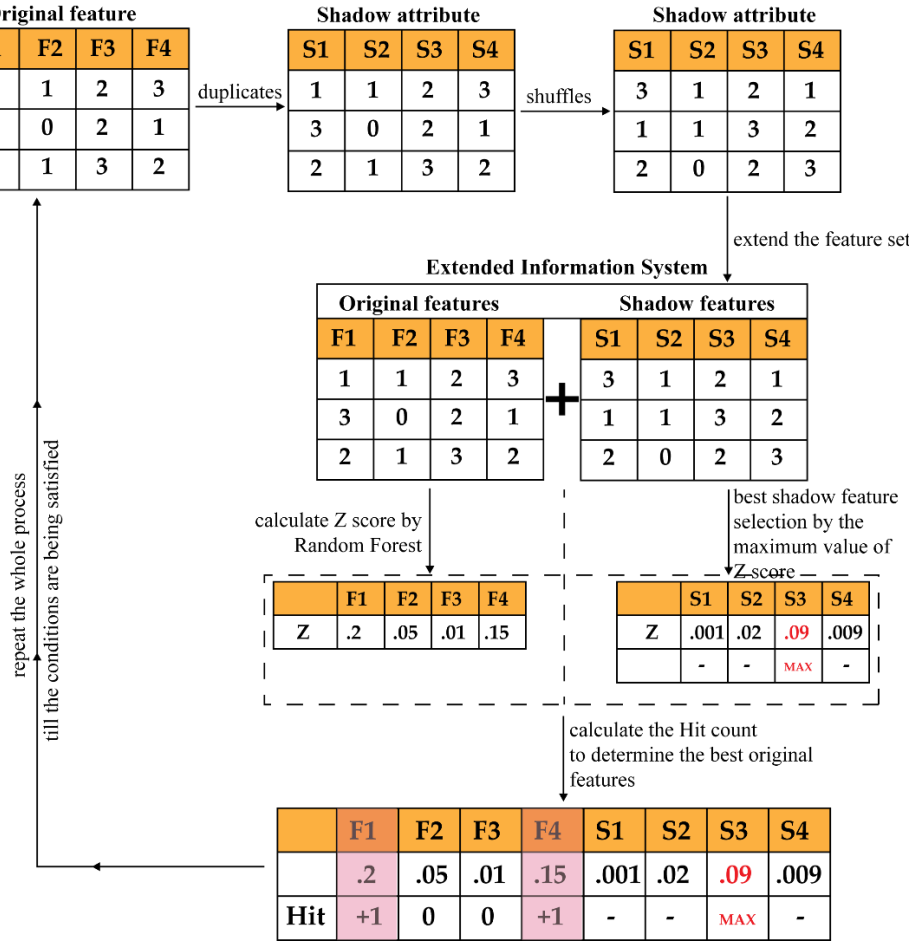


Figure 4.1: A visual representation of Boruta.

This algorithm first shuffles and permutes all the original feature attributes to create a randomized copy of original feature set. Then, this randomized set of features are added to

the original feature set to bring randomness to the feature attributes. Finally, it performs the feature selection process by RF classifier on those attributes [35]. Thus, by adding more randomness to the selection process, this algorithm is very powerful than the conventional forward/backward selection-based wrapper approaches [59,60]. The steps of the Boruta algorithm are described below:

(1) By shuffling and permuting all the original feature attributes, it first creates a randomized copy of original feature set. These replicated feature sets are known as Shadow Attributes (SAs). Then these SAs are merged with the original feature attributes to form the Extended Information System (EIS) [60]. At least 5 SAs are required to form the EIS.

(2) Then, the values of these SAs are randomly permuted, and shuffled. Thus, there is presence of pure randomness into all the replicated variables and the decision attributes.

(3) Afterwards, a RF classifier is fitted to the EIS several times, and SAs are randomized before each run. Therefore, for each run, the randomly updated part of the EIS is distinct.

(4) For each run, the importance of all feature attributes (Z score) is computed. To compute this importance the following steps are considered:

(a) The EIS is divided into several Bootstrapped Set of Samples (BSSs) (in another word, the training samples), equivalent to the considered number of Decision Trees (DTs) used for RF. Therefore, the samples for testing, which is commonly known as the Out of Bag Samples (OOBSs), are equivalent to the number of BSS.

(b) Then, each BSS is used for training individual DT, whereas the corresponding OOB is used for testing the performance of that DT. Thus, for each feature attributes from EIS, the number of votes for correct class are recorded.

(c) Later, the values of the feature attributes of OOBs are randomly permuted to record the votes for correct class once more like the previous step.

(d) Then the importance of the values of the attribute for a single DT can be calculated as follows:

$$MDA = CorrectVotesCast_{original} - CorrectVotesCast_{permuted} \tag{4.1}$$

This importance measure is known as the Mean Decrease in Accuracy (MDA).

(e) Finally, the importance of the values of the feature attribute (V_i) throughout the forest is computed as follows:

$$V_i = \frac{1}{N}\left(\sum_{n=1}^{N} MDA_n\right) \text{ where } N \text{ indicates the total number of DT.} \tag{4.2}$$

(f) Therefore, the final importance score is calculated,

$$Z = \frac{V_i}{\sigma_{V_i}} \tag{4.3}$$

(5) Find the Maximum Value of Z among the Shadow Attribute (MVSA). After that, assign a hit to every attribute that scored higher than MVSA. To determine the best attributes, the following steps are considered:
  (a) Consider an attribute as important if it performs significantly higher than MVSA.
  (b) Remove an attribute from the EIS as non-important if it performs significantly lower than MVSA.
  (c) For an attribute with undetermined importance, a two-sided test of equality is conducted with MVSA.
(6) Remove all the SA from EIS.
(7) Repeat the whole process till any of the following two cases are satisfied:
  (a) The importance is assigned to all the attributes.
  (b) The algorithm reach to the limit of defined number of RF runs.

The complete process is illustrated in Figure 4.1.

4.2.2 Spearman's Rank Correlation Coefficient

Correlation is a statistical relationship between two random variables in bivariate data [67]. In other words, it defines the degree of linear relationship between two random variables. However, correlation never imply causation of the data. The Correlation Coefficient (CC) is the statistical measure to calculate the relationship between the relative movements of two random variables. Among various CC measurements techniques [68,69], Spearman's Rank Correlation Coefficient (SRCC) is a non-parametric test to measure the correlation [61,62]. There are 2 main advantages of SRCC over rest of the CC measurement techniques:

(1) Due to the non-parametric test approach, it carries no assumptions of the distribution of the data [62].
(2) SRCC works on rank-ordered variables. Therefore, it works with the variables which have linear or, monotonic relationships [62,67].

The SRCC can be calculated by the following equation,

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{4.4}$$

Where, n resembles the number of observations, and d_i is the distance between the ranks of the corresponding variables.

4.2.3 Shapley Additive Explanation for Model Interpretation

The additive feature attribution method can interpret/explain the individual decision/prediction of classifier [63]. This method supports the model output as a sum of real values attributed to each input feature. Additionally, this method has a distinctive characteristic to provide a unique solution for the following 3 properties, i.e., (1) local accuracy, (2) missingness, and (3) consistency [63].

(1) Local accuracy: If x is the specific input to the original model f, then for approximating f, local accuracy requires the explanation model g to match its output with the output of f for any simplified given input x' to model g.

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i' = f\left(h_x(0)\right) + \sum_{i=1}^{M} \phi_i x_i' \text{ when } x = h_x(x'), \text{ and } \phi_i \in \mathbb{R} \tag{4.5}$$

Here, $x = h_x(x')$ is the mapping function, M denotes to the number of simplified input feature, and ϕ_i denotes the effect to each feature

(2) Missingness: If any feature is missing in the original input x, that feature x_i' has no attributed impact.

$$x_i' = 0 \Rightarrow \varphi_i = 0 \text{ when } x = h_x(x') \tag{4.6}$$

(3) Consistency: If a model a changed in such a way, so that a feature brings greater impact on the model, the attribution assigned to that feature will never decrease. Let $f_x(z') = f\left(h_x(z')\right)$ and $z' \setminus \mathrm{i}$ implies the setting for $z_i' = 0$ where $z' \in \{0,1\}^M$. For any 2 models f, and f', if

$$f_x'(z') - f_x'(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \Rightarrow \phi_i(f', x) \geq \phi_i(f, x) \tag{4.7}$$

From those above mentioned 3 properties, it can be easily understood that the considered additive feature attribution model is developed based on the classical methods [70,71] of estimating the solution concept in cooperative game theory called Shapley value [64].Therefore, SHAP values are calculated as a unified measure of feature importance as follows:

$$f_x(S) = f_x\left(h_x(z')\right) = E\left[f(x) | x_S\right] \text{ where } S \subseteq (index \neq 0) \in z' \tag{4.8}$$

Where local methods always try to ensure,

$$g(z') \approx f\left(h_x(z')\right) \text{ when } z' \approx x' \tag{4.9}$$

Previously, by Local Interpretable Model-Agnostic Explanations (LIME) [72], the explanation of this additive feature attribution method was introduced as a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{4.10}$$

To find the ϕ_i, LIME minimized the following objective function (loss function),

$$\xi = \arg\min_{g \in G} L(f, g, \pi_{x'}) + \Omega(g) \tag{4.11}$$

Where $\pi_{x'}$ is the local kernel to calculate the loss over set of samples in the simplified input space, and Ω denotes the penalizing constraint for reducing the time complexity [63]. On the other hand, with the Shapley regression values [63], feature importance was calculated in the presence of multicollinearity into the data by Equation (4.12),

$$\phi_i = \sum_{S \subseteq Q \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S) \right] \tag{4.12}$$

Where Q is the set of all input features. However, according to the proof of Lundberg et al. in [63], g can only satisfy the above mentioned 3 properties by this following equation,

$$\phi_i = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{4.13}$$

In this research, we have considered the kernel SHAP [ref] approach to bring the interpretability of the model. This approach uses Linear LIME (Equations (4.10), and (4.11)) with Shapley Values (Equation (4.13)) to calculate the feature importance.

By using Kernel SHAP, $L(f, g, \pi_{x'})$ from Equation(4.8) can be written as,

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f\left(h_x^{-1}(z')\right) - g(z') \right]^2 \left[\frac{(M-1)}{\left(M_{choose} |z'|\right) |z'| \left(M - |z'|\right)} \right] \tag{4.14}$$

This approach has two main advantages [63], i.e.,

(1) It can estimate the SHAP values without considering the model type.

(2) Feature importance is calculated in the presence of multicollinearity into the data.

4.3 Proposed Method

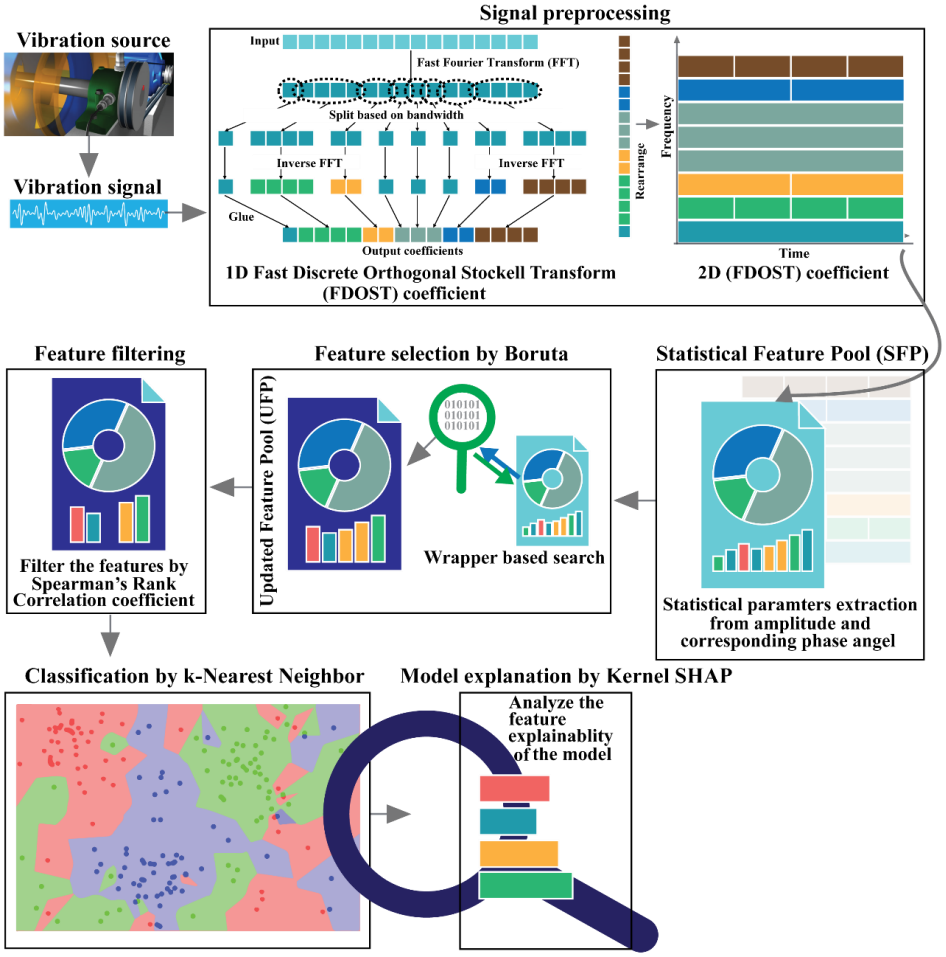


Figure 4.2: Block diagram of the proposed model for bearing fault diagnosis.

The main objective of this research is to introduce explain ability to the CM-FD process of bearing, where the feature selection process, and the importance of each selected feature for the ML classifier outcome can be easily interpretable. With this contribution the generalization capability of the proposed model can be significantly improved, and it can be effectively applied to different scenarios beside the considered ones in this work. The block diagram of the proposed method is illustrated into Figure 4.2. Based on this diagram, the core steps of the proposed algorithm are described into the flowing subsections.

4.3.1 Feature Selection by Boruta

After extracting the 20 statistical features from the FDOST coefficients like the previous chapter, the SFP is formed in the same way. From the obtained feature pool, now the aim is to identify the most important features for further analysis. In ML based analysis, it is necessary to remove the non-important features to avoid the unnecessary noise (based on

Occam's Razor (OR) principal [73,74]). Moreover, the filter-based feature selection mechanisms [55], and evolutionary algorithm-based feature selection approaches [53,54], mostly rely on the technique of removing the redundant features from the original feature set. However, the selection of all relevant features based on statistical justification, instead of just the non-redundant ones, is important for preventing the loss of any useful information from the overall feature set [60,75]. Therefore, to solve this issue, in this study, a wrapper-based feature selection approach - Boruta is considered. After providing the 20 statistical features enlisted SFP table to Boruta, it ranks those attributes based on the importance score termed as Z. Then, all the feature attributes ranked as important by Boruta, are considered to form the Updated Feature Pool (UFP).

4.3.2 Feature Filtering by Spearman's Rank Correlation Coefficient (SRCC)

In general, correlated features raise the issue of multicollinearity. Therefore, for some models, these features can adversely affect the performance due to multicollinearity issue [76]. In this study, the considered classification algorithm is k-NN. Although it is a non-parametric approach but the it can be affected because of the extra-weight carried by the correlated features on the distance calculation [77] resulting in the decline of the performance [78]. Usually, the performance discrepancy is low while dealing with a smaller feature set. Moreover, cross-validation approach also helps to overcome the issue. However, in this study, as we have focused on the interpretability of the classifier, therefore, a simpler model is necessary with less feature. This can be perceived as a special case of Occam's Razor principal [73,74] known as Minimum Description Length (MDL) [79,80]. Moreover, for better interpretation of k-NN outcome with Kernel SHAP, the SRCC is chosen as a medium for dimensionality reduction of the input data over the conventional technique, i.e., Principal Component Analysis (PCA). PCA or the techniques like PCA, do not deal with collinearity, instead they just compress the original data. However, with the proposed SRCC, we can reduce the dimensionality by removing the colinear feature information from the data. Therefore, without revoking the ranking of the features by Boruta, we filtered UFP again by the SRCC. Thus, the Final Feature Pool (FFP) is obtained for the classification.

4.3.3 Model Interpretation by Kernel SHAP

After obtaining the FFP, we have used k-NN as a classifier like the previous study. However, when a trained ML model yields an outcome for a regression or classification task, we might speculate while explaining the choice made by the model. With the accuracy-based evolution criteria of a model, we never get a complete description about the performance of a model. In the real-world cases, along with the accuracy measurement, it is necessary to understand the behavior of model with respect to inputs and output, so that when needed, it shall be easy to debug. Thus, with an interpretable model, we can learn more about the data, and

problems related to it. Therefore, the reason behind the failure of a model can be justified properly [81]. It is easy to explain a model if its decisions are easily understandable to humans. The best clarification for explaining the model is generally the model itself, since a simple model is easily represented and understood [82]. For example, the explanations of DT or RF can be easily interpreted by humans, whereas for the non-linear and complex models, it is often difficult for human to understand the decision-making process. However, the complex models can have higher accuracy scores than the simpler models for a larger dataset. Nevertheless, the interpretability of simpler models is higher than the complex ones. Therefore, in this study, to create a sweet spot between achieving high accuracy and interpretability of the model without sacrificing any of them, we have considered k-NN as the main classifier. Due to the simple architecture of the classifier with non-parametric behavior, it is comparatively easier to explain the model behavior by kernel SHAP [63] without compromising the accuracy as well.

4.3.4 Performance Evaluation Criteria

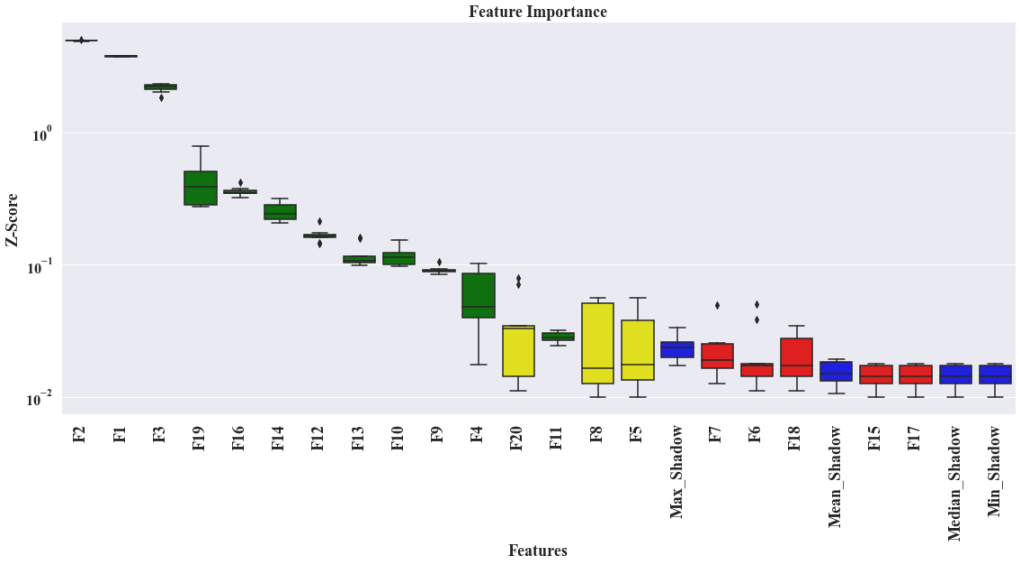
To evaluate the performance of the proposed interpretable model, we have considered 3 stages, i.e., (a) feature importance evaluation, (b) model performance evaluation, and (c) model interpretability evaluation. For feature importance evolution from FP, the Boruta Z score is considered. The details of this calculation procedure have already been described into **Section 4.2.1**. To assess the performance of the proposed model two measurement matrices have been considered, i.e., (a) Accuracy Score (AS) [48], and (b) Confusion Matrix (CM) [49] like the **previous Chapter**. Similarly, if there are in total N samples present in the dataset, the range of k - values will be $\left[1,3,5,.....,\sqrt{N}\right]$. Then, a grid-search approah with K fold cross validation (K-CV) is used while evaluating the performance of the classifier, whereas the values of K ranges from $[1,2,3,.....,10]$. The upper range of K is arbitrary, however, the most common practice in ML is to use 10-CV by considering K=10. Thus, the best AS matrix is decided on the best choice of k (k of k-NN) along with the best performing K-fold. Finally, the performance of the model is explained and justified by kernel SHAP.

4.4 Experimental Results Analysis

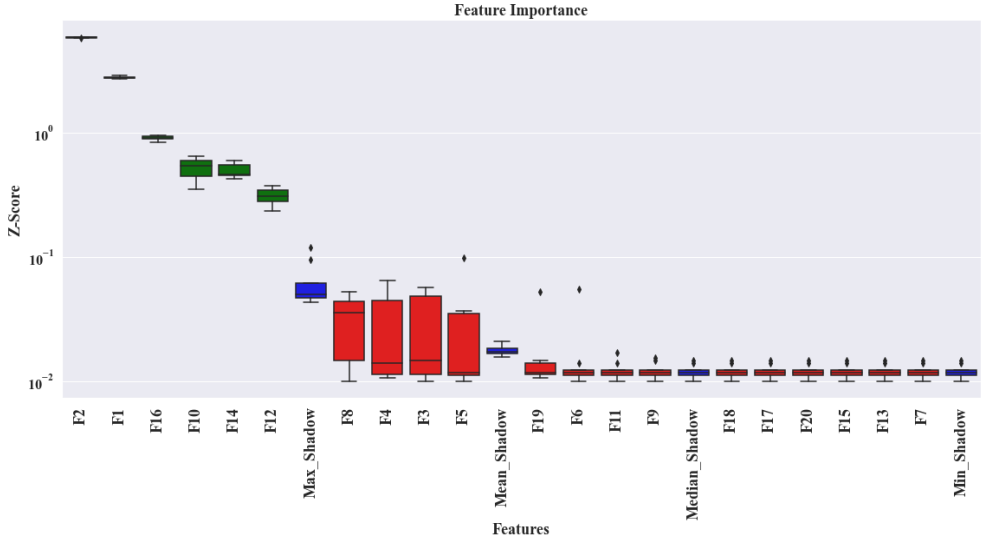
This section presents the description of experimental in a step-by-step manner for two separate case studies performed on two separate datasets like the previous chapter. For every dataset, first, the performance of the classifier is analyzed for various load, and speed conditions. After that, the performance score is justified based on Kernel SHAP explanatory graph. Finally, to prove the robustness of the proposed model, the accuracy is compared with several popular state-of-art techniques as well.

4.4.1 Case Study 1 – CWRU Dataset

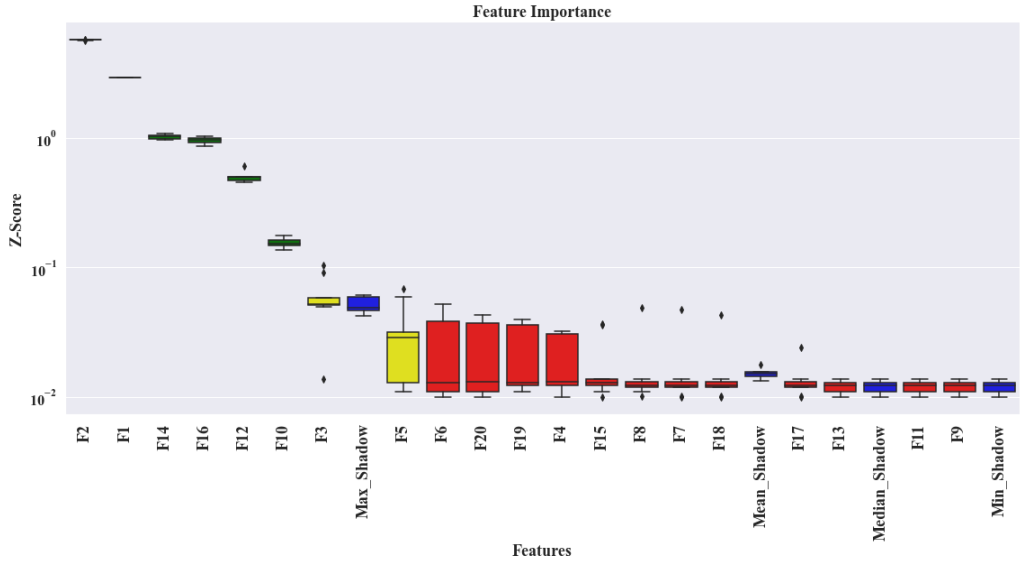
Till the SFP calculation, all the experimental setup for datasets, and the desired outcomes are identical like the **previous chapter (section 3.4.1)**. Later, Boruta is applied to get the most relevant features for the given task. We are considering all the important feature which are identified as important by Boruta. Intuitively, we can say that Boruta will pick only those set of feature attributes among all the extracted ones, which have class separability among different classes. Hence, to determine this phenomenon statistically, Z scores associated with each feature attributes are analyzed. Therefore, A graph with the Z scores associated with each feature are given below for all the datasets.



(a)



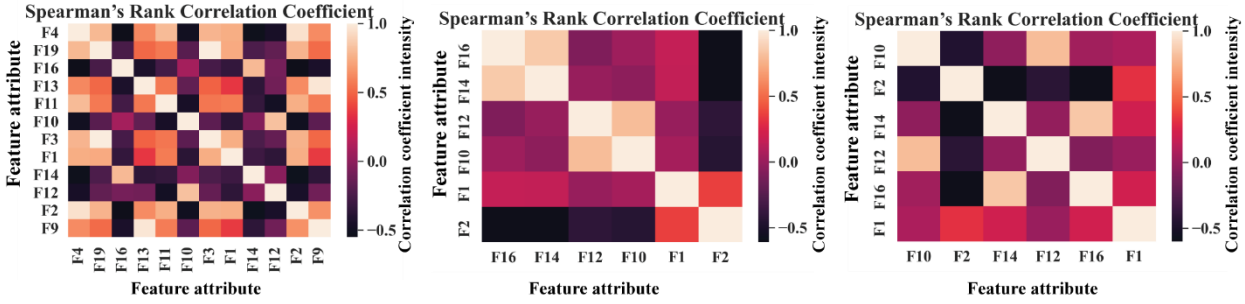
(b)



(c)

Figure 4.3: The Box-whisker plot presenting Z scores generated by Boruta for each feature associated with all the datasets, i.e., (a) dataset 1, (b) dataset 2, and (c) dataset3.

From these plots, the green means the most important features, yellow means tentative features, red means unimportant/rejected features, and blue corresponding to the shadow features. Furthermore, only green features (important ones) are considered to create the UFP. Then, by removing the colinear features having similarity score higher than 0.9 by SRCC (Figure 4.4), an FFP is formed which is later provided to the final classifier to identity the faults present in the bearing dataset. With the help of FFP, only the most relevant features which directly impact the outcome of the underlying classifier can be identified, hence, the influence of each feature on the outcome can be easily understood.



(a) **(b)** **(c)**

Figure 4.4: Representation of the feature correlation in the UFP of the three datasets: (a) UFP of dataset 1, (b) UFP of dataset 2, and (c) UFP of dataset3.

After forming the FFP, the training data from all the datasets are normalized to train 3 separate model. Each model is trained by 10-CV. With this grid search approach, best model,

with the most suitable number of neighbors (k of k-NN) is picked to calculate the performance of the model in terms of accuracy. The details of these tests along with the accuracy score are highlighted into Table 4.1.

Table 4.1: Diagnostic performance of the proposed model.

Model	Dataset	UFP	FFP	Common feature attributes	Best k	Accuracy (%)
1	1	'F13', 'F4', 'F2', 'F14', 'F3', 'F11', 'F10', 'F16', 'F9', 'F12', 'F1', 'F19'	'F13', 'F4', **'F14'**, 'F19', 'F11', **'F10'**, **'F16'**, **'F12'**, 'F1'	'F12', 'F16', 'F10', 'F14'	1	100
2	2	'F2', 'F14', 'F10', 'F16', 'F12', 'F1'	'F2', **'F14'**, **'F10'**, **'F16'**, **'F12'**, 'F1'		1	100
3	3	'F1', 'F12', 'F16', 'F10', 'F14', 'F2'	'F1', **'F12'**, **'F16'**, **'F10'**, **'F14'**, 'F2'		1	100

From Table 4.1, we can see that for model 1, Boruta picks 12 features as the most important ones (in UFP). Then, after removing the multicollinearity by SRCC, 9 features remain as the final candidate into the feature pool (in FFP). Now, let's observe the distribution of these 9 feature attributes to examine their class separability into the following Figure 4.5.

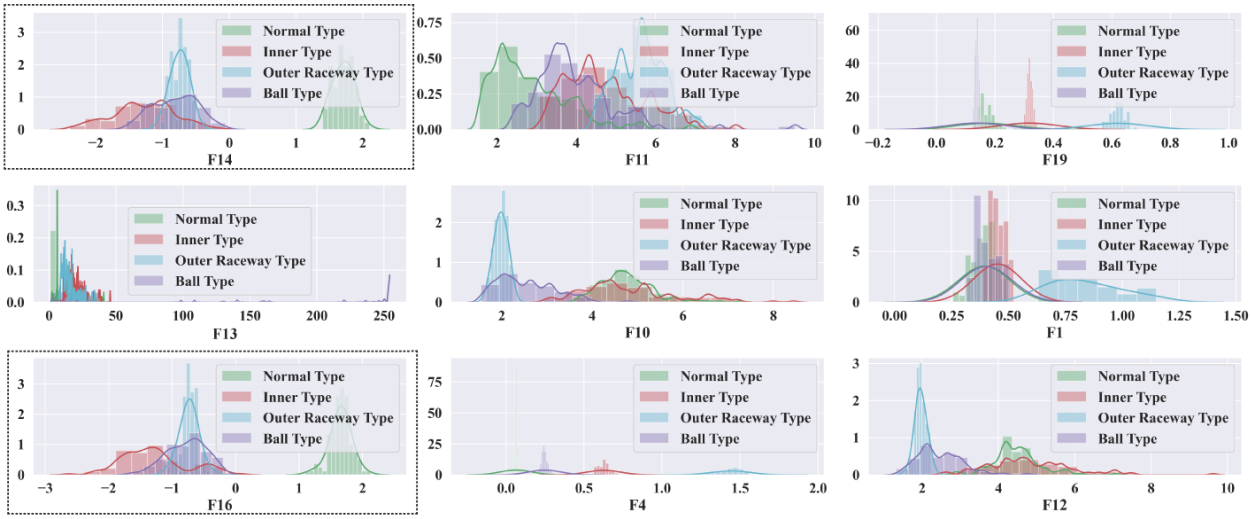
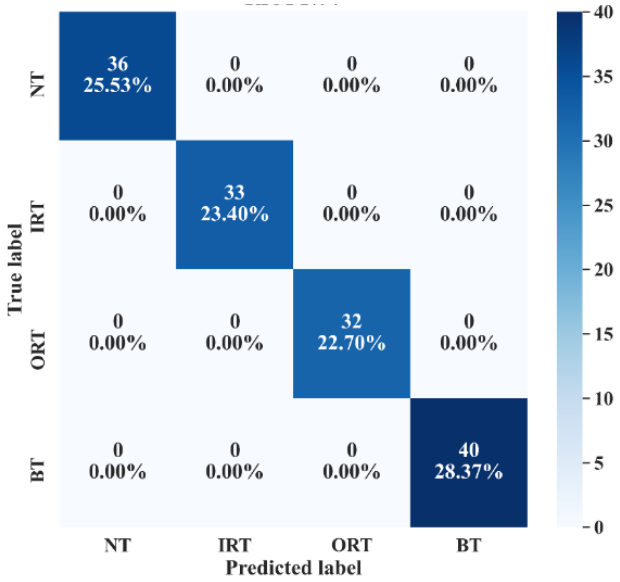


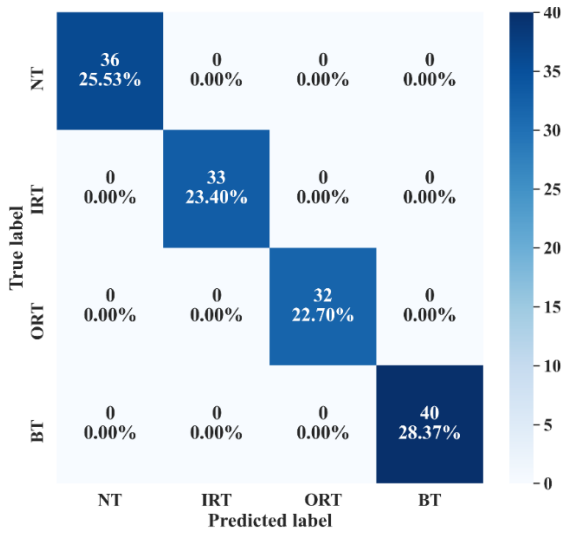
Figure 4.5: Feature distribution of FFP for model 1 in Table 4.1.

Now, by analyzing the distribution of these feature attributes, F14 , and F16 can provide a very interesting insight. For both, the distribution of Normal Type (NT) is to-tally separate than the faulty ones. Though, the faulty health conditions have some overlap into their distributions, still there are significant difference among the Inner Raceway Type (IRT), Outer Raceway Type (ORT), and Ball Type (BT). After that, F12, and F10 has a better

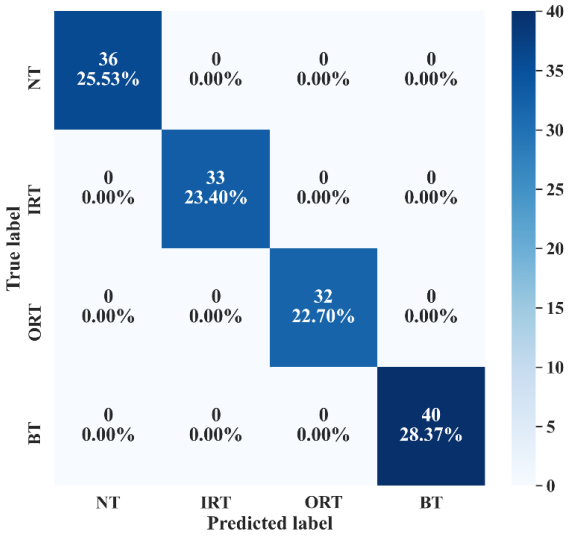
separability into the distribution than the others. Therefore, after passing all these selected features to the k-NN classifier, we have achieved 100% classification accuracy. However, according to our prior discussion, we need to check, for k-NN, which feature is the most prioritized one. We will analyze this part with the SHAPly additive explanations in the later portion. Before, zoom into that analysis, let's first create our own priority list of features from the distribution analysis. Based on our discussion, and observation from Figure 4.5, we think that the classifier will consider F14 or F16 as the first or second most important ones. After that, F12, and F10 will be prioritized. In a very similar way, the total number of selected feature attributes in UFP for both model 2 and model 3 are 9. Moreover, for these two models in UFP, there are no features which corelated more than 90% (Figure 4.4(b)-(c)). Therefore, no features are removed while forming the FFP through SRCC analysis. For model 1, dataset 1 is used for training as well as for testing. Similarly, for model 2 dataset 2 is used, and for model 3 dataset 3 is used. For each model, 70% of the data is used for training, and remaining 30% is used for testing. The details about the train and test split are discussed in **Table 3.2**. The confusion matrices of these three models are given in Figure 4.6. They demonstrate the class-wise accuracy of the models when tested with the respective datasets. From Figure 4.6, we can see that each model can classify all the test samples correctly from every health type, i.e., NT, IRT, ORT, and RT. That means each model achieves 100% True Positive Rate (TPR) and 100% True Negative Rate (TNR), which makes the performance of this model satisfactory with this dataset. Therefore, it can be inferred from this figure that the accuracy of each model is 100% for the individual class.



(a)



(b)

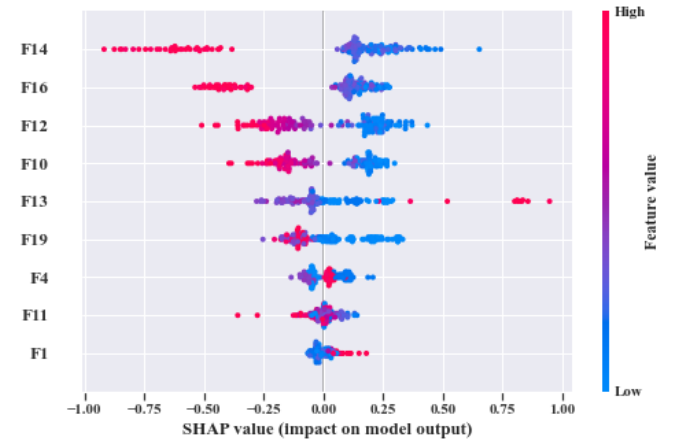


(c)

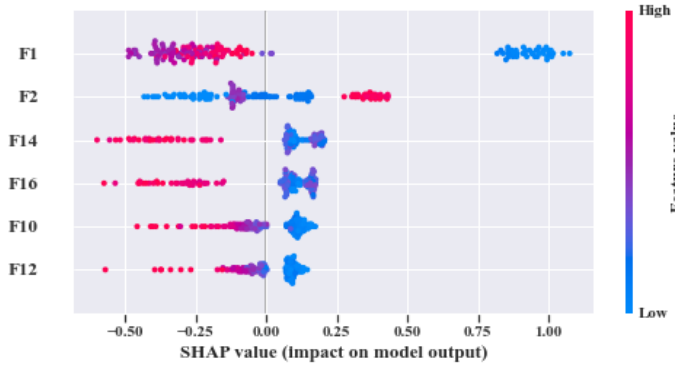
Figure 4.6: The confusion matrices of the three models that demonstrates the class-wise test accuracies, i.e., (a) dataset 1, (b) dataset 2, and (c) dataset3.

In addition, in depth analysis of Table 4.1 revels that in FFP of all the three datasets there are four common features, i.e., 'F12', 'F16', 'F10', 'F14'. Therefore, it is safe to consider that these four features directly influence the final outcomes of all the three models while dealing with three different datasets. Therefore, the analysis of these common features can divulge important details about the proposed CM-FD model. This information includes: (1) in which aspect and up to which degree these common features are affecting the outcome of the model individually as well as in collectively, (2) if the individual and collective effect of these

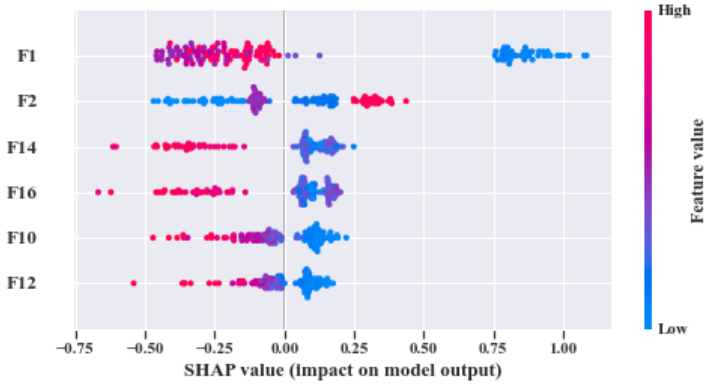
common features is positive and of higher degree then the model outcome can be explained with respect to these features alone. Once, it is ensured that the common features have significant impact on the performance of each model, only then, a generalized CM-FD model for bearing can be developed by using just these common features. A generalized CM-FD model will provide certain benefits, such as, application of the same model for the bearing fault identification no matter if the underlying factors that affects the machinery operation change. Furthermore, if the generalization power of a CM-FD model is high it can be utilized in the cross-domain applications. For this purpose, a SHAP kernel is used to explain the degree and nature of the impact these common features are having on the output of each model while dealing with the three datasets. In Figure 4.6, the impact of each feature on the performance of each model is explained with the help of SHAP value represented on x-axis. Likewise, the y-axis lists the importance of the features in ascending order, where, low represent the feature with least and high represents the feature with the highest impact on the output. For instance, from our previous discussion and analysis, for model 1 in Table 4.1, we concluded that the classifier might consider F14 or F16 as the first or second most important ones. Therefore, from Figure 4.7 (a), it can be observed that F14 is the feature with the highest impact, then next in line is F16, then comes F12, and so on. A feature is of highest importance or relevance when its SHAP value spans over greater range than the rest, as stated in [46]. Therefore, our theoretical intuitive hypothesis is validated. Thus, we know, why certain decision has been made, and which feature is responsible for certain decision made by the classifier. However, our goal is to understand the impact of the 4-common feature, i.e., 'F12', 'F16', 'F10', 'F14' on each model performance. For model 2, and 3, these four features have similar ranking. For these two models, 'F12', 'F16', 'F10', 'F14' are the third, fourth, fifth, and sixth ranked features, respectively. Moreover, the range of their SHAP values are also similar. Similarly, for model 1, these 4 features are the most dominant ones, where, in this case, 'F12' ranked ahead of 'F10'. Another interesting observation is that for all these models, the range of SHAP values for 'F14', and 'F16' are almost similar. In case of model 1, the range of SHAP values for these two features are even higher than the rest of the models. Therefore, from this analysis it can be concluded that these common features 'F14' and 'F16' has noticeable contribution on the performance of all the models. Moreover, in case of model 1 these features are the first two highly ranked features which means that the output of the model can be yield with the help of these features. Thus from this interpretation of the impact of common features on the output of the models with the help of SHAP kernel [63] it is safe to consider only two features i.e., 'F14', and 'F16' to build a generalized diagnostic model for bearing health state assessment while dealing with different sorts of datasets. Thus, with the help of these insights, number of task relevant features can be reduced with proper justification which makes our model faster, robust, and easily explainable in terms of feature importance for the outcome of the model.



(a)



(b)



(c)

Figure 4.7: Summary plots for all the test datasets with associated SHAP values: (a) SHAP values for model 1, (b) SHAP values for model 2, and (c) SHAP values for model 3.

Finally, based on the prior analysis, to evaluate the performance of the proposed generalized CM-FD model for bearing, from each dataset, we have considered only two features, i.e., 'F14',

and 'F16'. To prove the robustness of this model, we have performed 3 sperate tests. Among them, for test 1, we trained, and validated (train: valid = 80:20) the model only with the samples of dataset 1, and then tested the trained model with dataset 2, and dataset 3. Similarly, test 2, we have used dataset 2 for training, and dataset 3 and 1 for testing. Likewise, for test 3, datasets 1, and 2 are used for testing while dataset 3 is used for training the model. The details of this analysis are enlisted into Table 4.2.

Table 4.2: Diagnostic performance of the invariant model.

Test	Training dataset	Test dataset	Performance measurement (%) – Avg.		
			TPR	TNR	Accuracy
1	1	2,3	100	100	100
2	2	3,1	100	100	100
3	3	1,2	100	100	100

All these models achieve 100% classification accuracy on the test datasets. Therefore, it can be inferred form the given results that the reduced number of features that were selected with the help of Boruta and SHAP analysis were the most relevant and important features that helped the k-NN classifier to yield maximum performance.

Table 4.3: Comparison Analysis.

Method no.	Ref.	Signal processing	Feature extraction	Feature selector	Classifier	Invariance capability	Explain ability	Debuggable	Accuracy (%)	Performance gap
1	[50]	No	Time domain features: waveform length, slope sign changes, simple sign integral and Wilson amplitude in addition to establishe d mean	Laplacian Score (LS)	(Linear Discriminant Analysis) LDA, (Naïve Bayes) NB, SVM	No	No	No	99.6	0.4

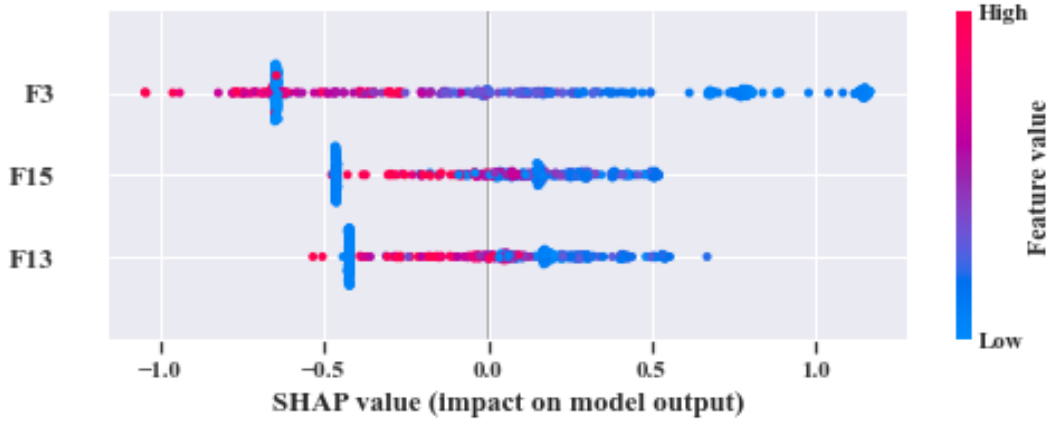
			absolute value and zero crossing							
2	[51]	Genetic Programming (GP)	Evolved features by GP stages	GP based filtering	k-NN	Yes	No	No	99.8	0.2
3	[52]	No	Local Binary Pattern (LBP)	No	Artificial Neural Network (ANN)	Yes	No	No	99.5	0.5
4	Chapter 3	ST	ST coefficient	GA	k-NN	Yes	No	No	100	0.0

It is noteworthy that, for the first-time bearing CM-FD is explained and its function is interpreted with different algorithms in this manner and this study shows the impact of such explanation on the development of a generalized CM-FD model. Additionally, to prove the robustness of the proposed bearing CM-FD model few comparisons have been presented in Table 4.3 where the compared studies are conducted for the similar working environment of the machinery. The details of these comparisons are enlisted into Table 4.3. From this table we can see that almost all the methods generate satisfactory classification accuracies for the considered scenarios. However, the compared studies have limitation in other aspects. For instance, method 1 although yields 99.6% accuracy but instead unable to generate same result if there working conditions changes, i.e., variation in the motor speed and load are encountered. Moreover, another disadvantage of method 1 is that this model cannot be interpreted/explained, thus, there is no way to debug the model for performance elevation or adaptability for other datasets. Similarly, Model 2,3, and 4 cannot be explained and debugged. However, among them, model 2, and 4 (proposed model in the previous chapter) gave 100% accuracy for datasets with variable speed and load conditions. Nevertheless, two things make our proposed model unique, and state-of-art, i.e., (a) justification for the selection of the feature attributes, and (b) the ability of the classifier to be explained and interpreted with respect to the selected features.

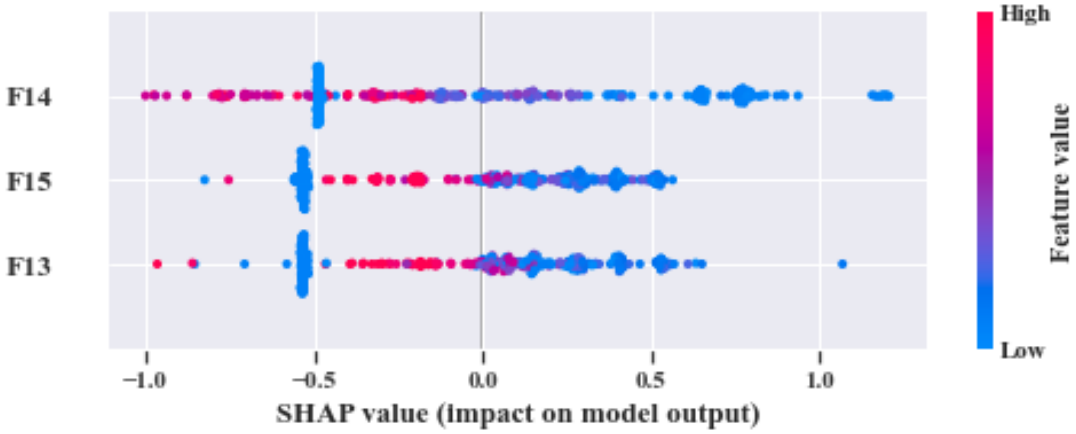
4.4.2 Case Study 2 – Dataset from the Self Designed Testbed

An additional experiment was performed to confirm that the proposed model provides satisfactory results if evaluated using totally different dataset. In this case as well, the experimental setup for the dataset along with the output of SFP are identical with the **previous Chapter (section 3.4.2)**. Like the case study 1, after forming the SFP, Boruta is applied on the feature pool to select the most relevant features as. The feature selection through Boruta is same as case study 1. Next, by removing the colinear features with score

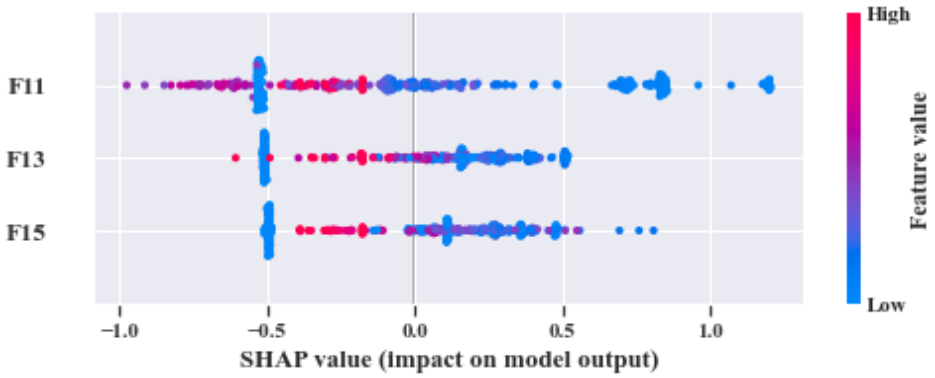
higher than 0.9 with the help of SRCC, FFP is formed. Once the FFP is obtained, the training data from all the datasets are normalized to train 3 separate model. Each model is trained by 10-CV. With this grid search approach, best model, with the most suitable number of neighbors (k) is picked to measure the classification accuracy. Later, by SHAP explanations is used to figure out common feature subset responsible for individual model performance. The details of these tests along with the accuracy score are highlighted into Table 4.4. From Table 4.4, we can observe that 'F13',and 'F15' are the common features for the 3 models. However, to understand the influence of each feature and to figure out the most influential features for the k-NN classifier output, we need to observe the SHAP values of the of each model. In Figure 4.7, given plots depict a summary of the SHAP values for individual models. By observing all the SHAP values from Figure 4.7 (a) – 4.7 (c), it can be inferred that F13', and 'F15' are highly impactful on the performance of all the three models. Therefore, to justify this analysis, k-NN classifier performance is analyzed by generating diagnostic results using these two features, i.e., 'F13', and 'F15'.



(a)



(b)



(c)

Figure 4.8: Summary plot for all the test dataset on the SHAP values of model (a) 1, (b) 2, and (c) 3.

Table 4.4: Diagnostic performance of the proposed model.

Model	Dataset	UFP	FFP	Common feature attributes	Best k	Accuracy (%)
1	1	'F3', 'F13', 'F16', 'F17', 'F20', 'F9', 'F1', 'F11', 'F14', 'F7', 'F15', 'F4', 'F19', 'F12', 'F18'	'F3', 'F13', 'F15'	'F13','F15'	5	99.2
2	2	'F15', 'F14', 'F19', 'F11', 'F18', 'F13', 'F9', 'F3', 'F12', 'F16', 'F20', 'F1'	'F15', 'F14', 'F13'		7	99.7
3	3	'F11', 'F19', 'F20', 'F1', 'F13', 'F15', 'F17', 'F14', 'F16', 'F3', 'F9', 'F4'	'F11', 'F13', 'F15'		5	98.1

Same as the previous case study, in this case as well, three sperate tests were performed to evaluate the performance of the proposed model. The test 1, we train, and validate (train: valid = 80:20) the model only with all the samples of dataset 1, and then evaluate the performance with dataset 2, and dataset 3. Similarly, for test 2, we have used dataset 2 for training, and dataset 3, and 1 for testing. Likewise, for test 3, datasets 1, and 2 are used for testing while dataset 3 is used for training the model. All these tests achieve at least 98.5% classification accuracy on the test datasets. The details of this analysis are enlisted into Table 4.5.

Table 4.5: Diagnostic performance of the invariant model.

Test	Training dataset	Test dataset	TPR				TNR				Accuracy (%)
			NT	IRT	ORT	RT	NT	IRT	ORT	RT	
1	1	2	1.00	.96	.99	.99	.99	.99	.99	.99	99.0
		3	1.00	.98	.99	.98	.99	.99	.98	.99	98.8
2	2	3	1.00	.99	.97	.99	.99	.99	.98	.99	98.5
		1	1.00	.99	.99	.97	.99	.98	.98	.99	98.5
3	3	1	1.00	.99	.99	.98	.99	.98	.98	.99	99.5
		2	1.00	.99	.98	.97	.99	.99	.98	.99	99.0

4.5 Conclusions

This chapter proposes an explainable AI based approach for the bearing fault diagnosis under variable speed and load conditions. A 5-stage scheme is suggested to identify faults in the observed bearing signals. The first stage is data preprocessing, in which FDOST coefficient-based vibration signal preprocessing is proposed for the exploration of invariant patterns from both the time-frequency and the corresponding phase-angel information for variable speed and load conditions. Thus, the heterogeneity is first created among different health types of specific working conditions so that the signals related to each fault type are easily identifiable. The exploration of heterogenous pattern among the signals of different health types is required to improve the efficacy of subsequent steps of fault diagnosis pipeline. The next stage is feature extraction, which consists of statistical feature extraction per-formed on FDOST coefficients to quantify the signals from the invariant pattern of the preprocessed data. After feature extraction, an explainable feature selection process is incorporated by introducing Boruta. With the introduction of these step, the behavior of feature selection process is being randomized to make it more robust, and bias-free. Moreover, the DT based wrapper-based classifier helps to identify the reason for selecting certain features from the input feature set. Next, a filtration method is introduced in addition to the feature selection to avoid the multicollinearity problem. Thus, the minimum description length is obtained from the selected feature set to satisfy the Occam's Razor principal. This reduced feature set contain non-overlapping information which is helpful in generating unbiased classification results. The last stage is about the classifier interpretation, in which an additive Shapley explanation followed by k-NN is proposed to diagnose the health conditions and to explain individual decision of the k-NN classifier for understanding and debugging the performance of the model. For the CWRU bearing dataset, our proposed approach gives 100% classification accuracy on average, and similarly, for our own experimental dataset, it gives around 97.0% classification accuracy. The conducted case studies show that the explainable model can help to build a robust classifier for invariant working scenarios with proper interpretations and explanations.

With the explainability of the feature selector and the interpretability of the classifier, we understand that for different datasets with different configurations, the proposed approach can be adopted. For adaptation of the current model for different datasets, some additional steps might be required during the feature analysis steps, i.e., changing the threshold values for discarding the colinear features, or smoothing the variables of the feature attributes to some extent. Nevertheless, having an explainable and interpretable architecture, the proposed model has better generalization capability, which can perform bearing fault diagnosis under different configurations for the variable working conditions of machines. However, this model has only one limitation. While developing a generalized version of the model for invariant working conditions, after all the analysis separately from all the working conditions, we need to first determine and fix the best number of features, and their attributes. Then, we need to build a classifier from scratch with the selected/determined number of features as input. To create the generalized version of the model, this additional step of training the classifier is the pitfall of ML based analysis. Therefore, in the next chapter, by utilizing the visual patterns of the FDOST coefficient, we have attempted to propose a deep learning (DL) based solution for bearing fault diagnosis for variable working conditions.

Chapter 5

A Transfer Learning based Approach for Condition Monitoring of Bearing by using a FDOST Coefficient-based Vibration Imaging

5.1 Introduction

In the **previous Chapter**, we have demonstrated that, after the application of signal processing techniques, CM-FD pipeline normally utilize feature extraction, explainable feature selection, and feature filtration for final classification by ML based classifier. However, while developing a generalized version of the model for invariant working conditions, after all the analysis separately from all the working conditions, we need to first determine and fix the best number of features, and their attributes. Then, we need to build a classifier from scratch with the selected/determined number of features as input. To create the generalized version of the model, this additional step of training the classifier is the pitfall of ML based analysis. Therefore, to mitigate this problem along with the feature extraction, and selection process automatic, in this chapter, a deep learning-based model is proposed after the proposed preprocessing step by FDOST.

Instead of selecting optimal features Two-dimensional (2D) FDOST coefficient directly, the co-efficient matrix itself used as the input to the DL based classifier. In the previous chapters (**Chapter 3** - **4**), we have already proved that our proposed preprocessing approach with FDOST coefficient can create invariant patterns for variable speeds, and/or load conditions. Therefore, by utilizing this similar visual pattern contained images, we can build a very powerful DL based approach for final identification of the health conditions of the bearing.

Thus, from these 2D images, the feature selection process is automated by employing a Transfer Learning (TL)–based Convolutional Neural Network (CNN). Conventional, straightforward neural networks make the feature selection process much easier through their convoluted encoding layers compared to traditional methods [83,84]. Yet, because these neural networks strain to deal with large amounts of data, to make the learning faster and effective, a transfer-learning-based neural network is proposed.

At the starting of 2017, several researches on TL-based Artificial Neural Network (ANN) for a bearing's raw vibration signals for variable working conditions of bearing was demonstrated [85]. However, for a raw signal, this approach cannot discover the critical features needed to transfer to the knowledge domain for further classification under different loads and speeds. Inadequate raw signal data extracted from mechanical sensors lead us to incomplete observation of critical patterns for the neural networks. To create invariant patterns, pre-processing steps are necessary since our data is limited. Due to white noise in signals, it is difficult to find out the exact information from additional properties mixed with domain data. These learning are passed through another working condition later thorough transfer learning. That is how, the network can learn from both working conditions and be fine-tuned by adjusting weights along with previous learning. By considering the limitations with the raw data, we have proposed a TL based approaches with the preprocessing step by FDOST coefficient in 2018 [44]. This work is still among one of the most pioneer works for TL based fault diagnosis of bearing for invariant working conditions. In this chapter, we have highlighted the proposed methodology of that work in details by incorporating the visual explain ability with DL based proposed model. The main contributions of this chapter can be highlighted as follows.

(1) After capturing the invariant patterns from FDOST coefficients from variable working conditions from the vibration signals both at low and high frequencies, these 2D FDOST coefficients are used as a 2D time-frequency image. Then, these are converted into grayscale, coined as Vibration Imaging (VI), to utilize CNN efficiently to automate the feature extraction and selection process.

(2) CNN-TL is proposed for fault diagnosis for vibration signals of bearings with variations of the shaft speed (e.g., RPMs). The proposed method (VI + CNN-TL), including VI for the RPM independent pattern and CNN-TL for fault diagnosis under variable RPMs, is suitably validated with extensive experiments and simulations, which justifies the potential of the proposed methodology over existing approaches in terms of achieving satisfactory theoretical results compared to experiment.

Like previous chapters (Chapter **3 – 4**), to validate the proposed model, two bearing datasets have been considered, among which, one is obtained from the public repository of CWRU

[12], and other one is collected from a self-designed test bed. The performance of the signal processing step, feature selection process, and the classifier has been verified with several comparisons. The complete organization of this chapter can be summarized as follows: **Section 5.2** gives the theoretical and mathematical descriptions of the necessary backgrounds, **Section 5.3** discusses about the proposed methodology in a step-by-step procedure, **Section 5.4** highlights the experimental analysis with discussion, and **Section 5.5**; finally concludes this research work.

5.2 Technical Background

This section discusses the technical details of the Convolutional Neural Network (CNN), and Transfer Learning (TL). The details related to the integration of these techniques into the proposed diagnosis framework have been discussed into the **proposed methodology** section.

5.2.1 Convolutional Neural Network

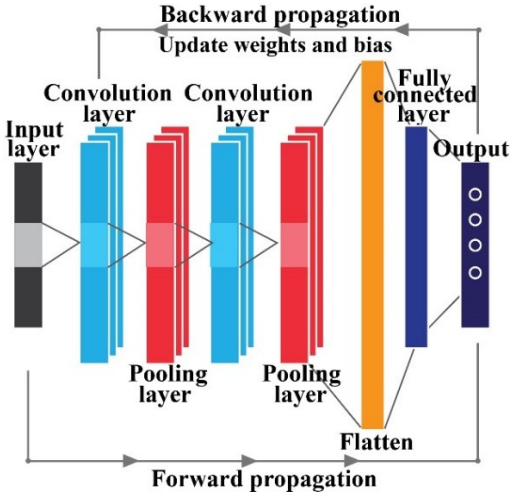


Figure 5.1: Common architecture of a convolution neural network (CNN).

The CNN architecture is usually formed with one input layer, a few convolution layers and pooling layers, several fully connected layers, and one final output layer to automate the feature extraction process [86]. The CNN successfully captures the spatial and temporal dependencies of an image through its different layers and preserves the features important for classification in a computationally powerful manner [87]. Additionally, incorporation of several optimization techniques in the recent few years that include Batch Normalization (BN), Dropout layer (DL), and Rectified Linear Units (ReLU) has improved the performance of CNNs [88–90]. The training process for a CNN is implemented in two stages, forward propagation stage and backward propagation stage (Figure 5.1). In the forward propagation stage, the CNN architecture extracts the spatial information from the input image throughout the designed layers [91]. In the backward propagation step, the network tries to update the

internal parameters in order to optimize the given objective function [92]. From the previous literature available on CNN, it is obvious that there is no rule of thumb for selecting the best number of layers in a CNN and the selection process for the total number of layers is a train-test based approach and is dependent on the nature of the input data. The forward and backward propagation are further explained as follows:

Forward Propagation

Convolution Layer: In this step, the Convolution Layers (CLs) learn the abstract features from the input image data to retain the relationship between pixels of the input while learning image attributes [91]. To achieve an enhancement of these convoluted features, an activation function with added weights and biases is applied [92]. This whole process can be expressed by the following relation:

$$x_n^m = f\left(\sum_{i \in K_n} x_i^{m-1} * w_{in}^m + b_n^m\right) \tag{5.1}$$

where x_n^m is the m th component of layer n, K_n is the n th convolution region of the $m-1$ layer feature map, w_{in}^m is the weight matrix, and b_n^m is the added bias. After calculating the summation of the total operation as described into Equation (5.1), a non-linear activation function f, called Leaky RELU, is applied on it. This function can be written as:

$$f(x) = \max(0.1x, x) \tag{5.2}$$

Pooling Layer: Immediately after the CLs, a Pooling Layer (PLs) is added to decrease the redundancy of the extracted features from the previous layer. In this study, max pooling is used as the pooling layer [93], which can achieve the maximum value of the convolutional output x_n^m as follows:

$$x_n^m = f\left(w_n^m * \max(x_n^{m-1}) + b_n^m\right) \tag{5.3}$$

where, the output x_n^m of the convolution layer is down sampled, w_n^m , and b_n^m represents the weight and bias matrix, respectively. The $\max(x_n^{m-1})$ denotes the max pooling function used to lessen the dimensions of the attained convoluted feature maps.

Fully Connected Layer: To increase the depth of the network architecture, several CLs and PLs are stacked together. Usually, several Fully Connected Layers (FCLs) are arranged layer by layer till the final one is reached, which alters the resultant filter matrix to a column or row [94]. Thus, at the end, the output feature can be obtained by the final fully connected layer which is given as:

$$y^z = f\left(w^z x^{z-1} + b^z\right) \tag{5.4}$$

where z represents the continuous order of the network architecture and y^z is the output of the final fully connected layer, f is the activation function to give the probabilistic output from the input. In this research paper, SoftMax [94] is considered as the final activation function.

Backward Propagation

After completion of the forward propagation, the objective function (commonly known as, loss function) is calculated to acquire the target data in accordance with the input data. Once the loss function is calculated, the parameters (i.e., weights, and biases) of the network architectures are updated in a reverse manner. This is achieved by propagating the loss function error in the backward direction. In this study, the cross-entropy loss function [92] can be expressed as follows:

$$E(w) = \frac{1}{n}\sum_{z=1}^{n}\left[y_z \ln \overline{y_z} + (1-y_z)\ln\left(1-\overline{y_z}\right)\right] \tag{5.5}$$

where y_z and $\overline{y_z}$ are the actual target and the predicted value of the zth sample, respectively. During the training process, the stochastic gradient descent method is utilized to minimize the loss function and update the network weights and biases. While training the neural network, the entire input dataset is divided into several smaller groups called batches, and multiple batches are supplied to train the CNN [95]. Thus, to minimize the loss function and avoid overfitting or underfitting problems, the entire training process is realized over several epochs [10,95].

5.2.2 Transfer Learning

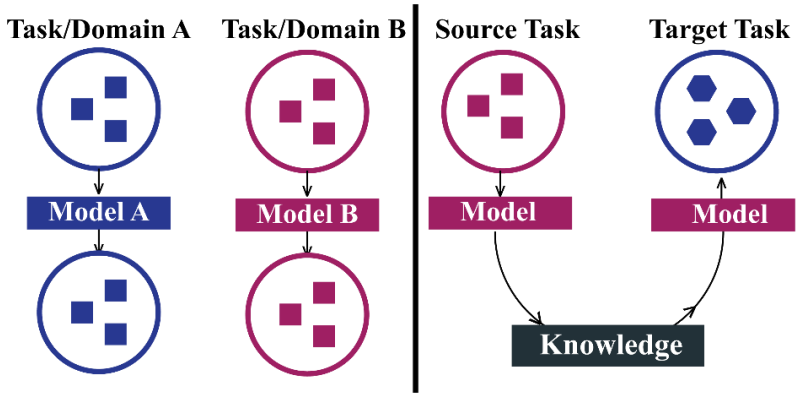


Figure 5.2: The left side shows the conventional learning process while the right side shows the concept of TL.

According to the definition, the main goal of TL is to transfer knowledge obtained from one task to another task that resembles it closely to improve the diagnostic performance of the new task in a short amount of time [44]. Fine-tuning-based TL (FTL) is one of the most popular approaches for designing a TL-based architecture [96]. In FTL, knowledge of the source task is transferred to the target task by transferring the learned parameters only. The source task is the task on which the proposed CNN is trained, and the parameters are adjusted accordingly. Alternatively, the target task is a task that is very similar to the source task; the learned parameters can be transferred from the source task to the target task to make the training process faster [10]. In this study, for the source task, a CNN architecture is trained to identify the bearing health conditions under a certain speed condition. Then, in the target task, this learned knowledge and the network parameters are passed to identify the health condition of bearing for a different speed condition. Thus, the necessity of training the target task from scratch is mitigated. For example, the output of the source task can be expressed as:

$$S_{y_n^m} = S_{f_m}\left(S_{x_n^m}\right) \tag{5.6}$$

where S_{f_m} is the final objective function or mapping function of the source task. Similarly, like the source task, the relatively similar target task can be expressed as:

$$T_{y_n^m} = T_{f_m}\left(T_{x_n^m}\right) \tag{5.7}$$

Where T_{f_m} is the final objective function of the target task. In the TL framework, by using the CNN architecture, the network first learns the mapping function S_{f_m}. After that, S_{f_m} is transferred to $T_{y_n^m}$ for obtaining the optimized objective function T_{f_m}, which improves the learning process. The visual concept of TL is illustrated into Figure 5.2.

5.3 Proposed Methodology

The block diagram of the proposed method is illustrated into Figure 5.3. Based on this diagram, the core steps of the proposed algorithm are described into the flowing subsections.

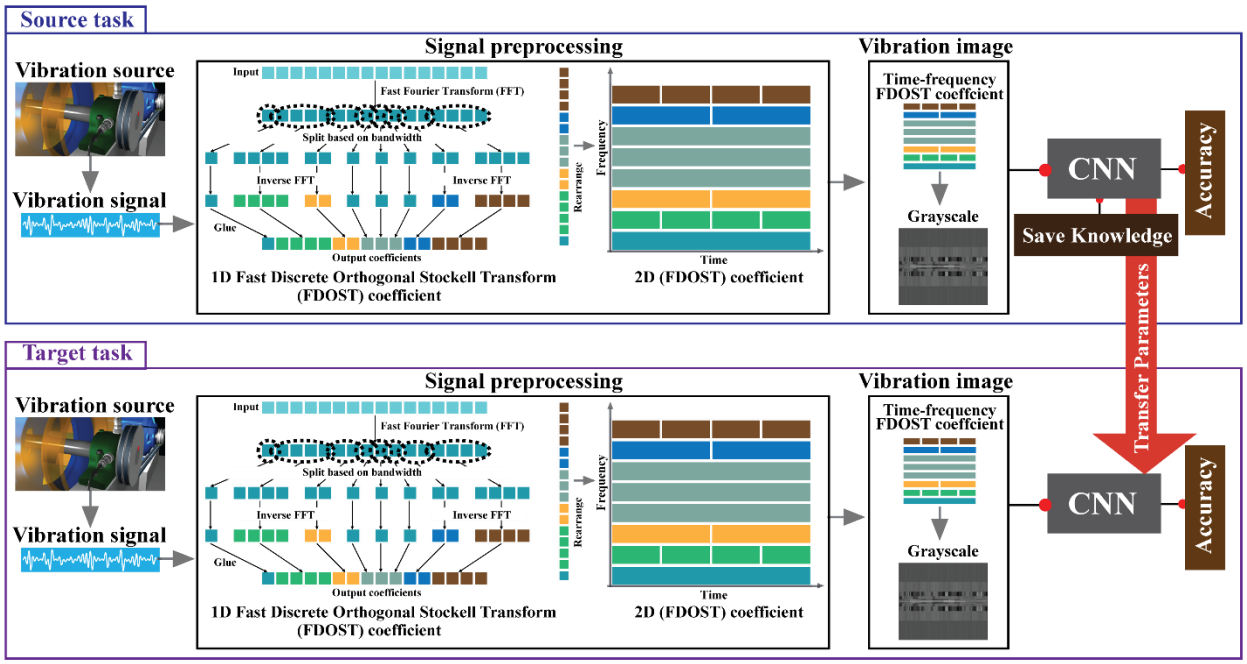


Figure 5.3: The diagram of the proposed model.

In this study, our target is to classify machine faults under variable speed conditions (i.e., RPM) using a TL model, such as proposed CNN-TL. There are three main steps in the proposed methodology (VI + CNN-TL): the source task, transfer, and target task. In the source task, we first apply FDOST to obtain the time-frequency coefficient to capture the invariant patterns of the data. Then, this matrix is converted into a 2D time frequency image, and finally converted into grayscale image, which is referred to Vibration Imaging (VI). This allows us to visualize the bearing health condition, and the 2D images are fed to the convolutional neural network (CNN) for model parameter optimization. The transfer block mainly passes the knowledge gathered from the source task network to the target network to complete the TL. In the target task, we test the model for classifying faults for variable RPMs.

5.3.1 Data Pre-processing by Vibration Imaging (VI)

Preprocessing of vibration signal plays a vital role, particularly in neural network-based fault diagnosis techniques [97–99]. The VI framework is executed into two steps, (i) The acquired time-domain vibration signals are decomposed via the FDOST coefficient, and 2D time-frequency images are obtained. These 2D images retains the information about the energy distribution across the time-frequency plane for different health conditions [100–102], and (ii) the resulting time-frequency images are converted into gray-scale images using a weighted sum of the red, green, and blue intensity pixels [103]. This adds computational benefits for the neural network-based analysis. For simplicity, these images have been referred to as VI in the paper. In order to meet the size restraints of the developed TL-CNN

architecture, the VI's are constricted [91]. Therefore, each VI is compressed into $256\times256\times1$ dimensions.

5.3.2 Proposed CNN Architecture

The proposed CNN architecture is inspired by original Le-Net5 [104]. The proposed CNN architecture is illustrated into Figure 5.4.

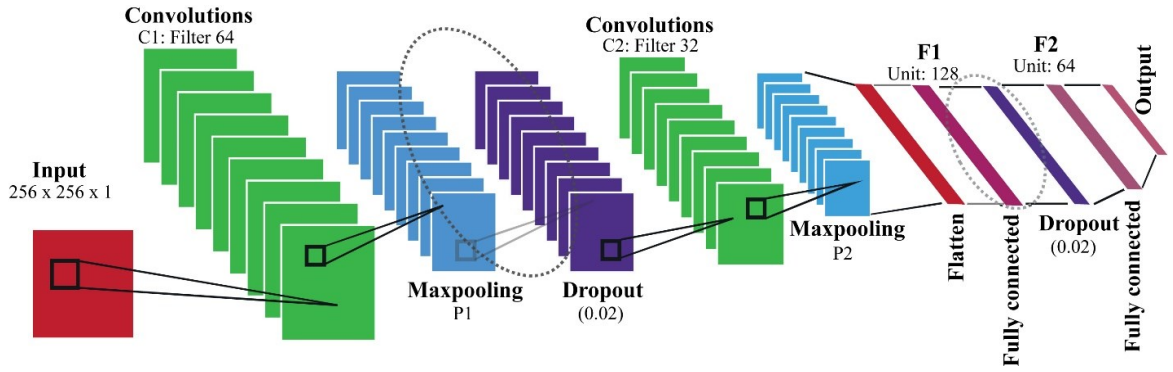


Figure 5.4: Architecture of the proposed CNN.

The proposed CNN architecture can be explained as follows: the CNN has a total of ten layers which starts with the input layer followed by a total of two CLs. After CLs, the architecture has two PLs, two DLs, two FCLs and finally one output layer. The size of the input layer is determined according to VI's (256X256X1). To reduce the number of parameters and to enhance the training efficiency the kernel size is determined to be 3X3. The CL1 and CL2 consist of 64 and 32 filters. The CL1 size is down sampled by the PL2. The FCL1 merges all the feature maps of the CL2 into a 1-D form. The FCL2 facilitates the final layer to classify the input data into its respective classes. The valid convolution method utilized in this neural architecture accepts the size of the feature maps to remain unchanged. Furthermore, the two dropout layers allow the network to generalize data to reduce over-fitting problems [88,90]. As mentioned earlier, the training process of the neural network is conducted through the Backward Propagation Stage (BPS). The main goal of training the network is to minimize the objective function error by updating the weights and biases through the BPS. In the training stage, a deep learning rate is considered to provide the CNN structure. This deep learning can optimize the performance of the neural network and avoids the convergence of the objective function to a local minimum. Furthermore, for updating the weights into the CNN, an adaptive moment estimation method (Adam) is considered [105]. Adam combines the advantages of deep gradient algorithms (AdaGrad) to deal with the sparse gradients and a Root-Mean-Square Propagation (RMSProp) algorithm capable of performing at non-stationary settings. Adam preserves the Exponential Moving Averages (EMA) of the gradient and its square in every update, which are related as follows:

$$w = w - \alpha \frac{B_{mt_1}}{\sqrt{B_{mt_2} + \varepsilon}} \tag{5.8}$$

where, $B_{mt_1} = \beta_1 B_{mt_1 - 1} + (1 - \beta_1) \frac{\partial}{\partial w} \cos t(w)$ here $\beta_1 \approx 1$ (5.9)

$$B_{mt_2} = \beta_1 B_{mt_2 - 1} + (1 - \beta_2) \frac{\partial^2}{\partial w^2} \cos t(w) \text{ here } \beta_2 \approx 1 \tag{5.10}$$

where w is the weight parameter and α is the positive scalar step size. Here, B_{mt_1} and B_{mt_2} are the first and second moment bias correction respectively and β1, β2 are the decay rates. From Equation (5.9) - (5.10), it is observable that the step size α and decay rates β1, β2 are small. Thus, the weight update process described into Equation (5.8) offers a nearly optimal learning rate selection [105]. The hyperparameters including dropout rate, learning rate, momentum, and number of batch sizes are tuned in grid-search based approach with 10-CV. In the TL, the learning of these hidden layers is then passed to the target task to boost the learning of the target task. For the neural network, these learnings are stored in the form of weights. The layers that are transferred to the target network are listed in Table 5.1.

Table 5.1: The proposed CNN structure with TL specifications.

Layers	Params.	Observations	Height	Width	Depth	Trainable	Stage Specs.	Transfer
Input			256	256	1			
Conv. 1	Kernel Size	Filter	3	3		Yes	1	Yes
	Padding	Zero						Yes
	Depth	Filter number			64			Yes
	Output		256	256	64			Yes
Pool 1	Kernel Size	Filter	3	3		No		Yes
	Padding	Zero						Yes
	Output		85	85	64			Yes
Dropout	Output		85	85	64	No		Yes
Conv. 2	Kernel Size	Filter	3	3		Yes	2	No
	Padding	Zero						Yes
	Depth	Filter number			32			Yes
	Output		85	85	32			Yes
Pool 2	Kernel Size	Filter	3	3		No		Yes
	Padding	Zero						Yes
	Output		28	28	32			Yes

Dropout	Output		28	28	32	No		Yes
F1	Nodes	Flatten into 1D	128			Yes	3	Yes
F2	Nodes	Flatten into 1D	64			Yes		Yes
Output	Nodes	Flatten into 1D	4			Classify	4	No

5.3.3 Performance Evaluation Criteria

To evaluate the performance of the proposed framework, different performance evaluation matrices have been considered in this work, including the (a) F1 score (F1), (b) Accuracy Score (AS), and (c) loss function graph. The F1 score can be calculated by Equation 5.11.

$$F1 = \left[\frac{2TP}{2TP + FN + FP} \times 100 \right]\% \tag{5.11}$$

Moreover, to adjust the overfitting and underfitting problems, the total loss of the model is observed until 3000 epochs. Further, to visualize the class separability, the feature space of the output layer is visualized by t-Stochastic Neighbor Embeddings (t-SNEs) [106]. Subsequently, to remove the bias from the data along with the evaluation parameters, 10-CV [107] is used for each experiment like previous chapters.

5.4 Experimental Results Analysis

This section presents the description of experimental in a step-by-step manner for two separate case studies performed on two separate datasets. For every dataset, first, the performance of the classifier is analyzed for various load, and speed conditions. Later, to prove the robustness of the proposed model, the accuracy is compared with several popular state-of-art techniques as well.

5.4.1 Case Study 1 – CWRU Dataset

In this case, based on different load, and speed conditions, three different datasets were created, which has already been described in **Chapter 2**. Before preprocessing the data, an overlapping segmentation [10] is considered to generate more number of samples in order to utilize the full potential of DL algorithm. To generate results, a total of three experiments are conducted. In experiment 1, dataset 1 is used for the source task, while datasets 2 and 3 are considered in the target task. At first, the VI are attained from dataset 1, and then the TL-CNN is trained and tested with 90% and 10% of the data, respectively. The trained TL-CNN architecture with learned weights is then saved to use for the target task. In the target task, the VI based inputs are calculated from datasets 2 and 3. Then, the TL-CNN architecture with the learned knowledge is used to adjust the target task's TL-CNN for measuring the final diagnostic performance. In this case, 15% of data from datasets 2 and 3 are used for training the network, and the remaining 85% of the data from both datasets are used for testing

purposes. Similarly, for experiment 2, dataset 2 is considered for the source task and datasets 1 and 3 are considered for the target task. For experiment 3, dataset 3 is considered for the source task and datasets 1 and 2 are considered for the target task. Therefore, in this diagnostic framework, the dataset is divided separately for two different tasks, i.e., the source task and the target task, as shown in Table 5.2.

Table 5.2: Data division.

Source task details	Dataset	Train (90%)		Test (10%)
		Training (80%)	Validation (20%)	
	1	1944 samples	216 samples	240 samples
	2	1944 samples	216 samples	240 samples
	3	1944 samples	216 samples	240 samples
Target task details		**Train (15%)**		**Test (85%)**
		Training (90%)	**Validation (10%)**	
	1	324 samples	36 samples	2040 samples
	2	324 samples	36 samples	2040 samples
	3	324 samples	36 samples	2040 samples

Additionally, to remove the bias from the datasets, an equivalent number of samples are considered for each health type. Moreover, the model is trained for 3000 epochs for the source tasks. For each experiment, once the network is trained in the source task, the performance of the target task is observed to measure the final classification accuracy. The diagnostic performances of the three experiments are listed in Table 5.3.

Table 5.3: Diagnostic performance of case study 1.

Exp.	Source Task	Target Task	Health type	F1 (%)	AS(%)
1	Dataset 1	Dataset 2, 3	NT	100	100
			IRT	100	
			ORT	100	
			BT	100	
2	Dataset 2	Dataset 3, 1	NT	100	100
			IRT	100	
			ORT	100	
			BT	100	
3	Dataset 3	Dataset 1, 2	NT	100	100
			IRT	100	
			ORT	100	
			BT	100	

As can be seen from this Table, the target tasks for all three experiments, achieved 100% accuracy. To show the detailed analysis of this specific result, from the source task of experiment 1, the graph of the loss function and the last layer feature separability of the source network obtained by t-SNE, are highlighted in Figure 5.5.

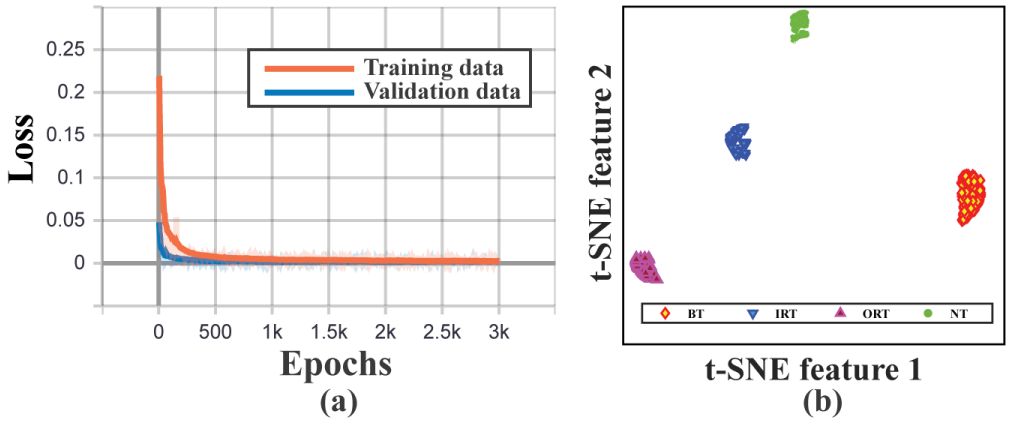


Figure 5.5: For experiment 1 – source task (dataset1) (a) loss function, (b) bottom neck layer features.

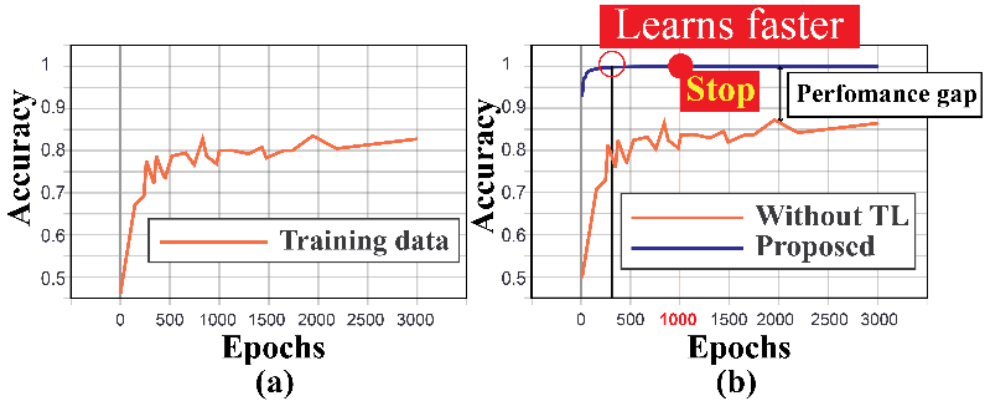


Figure 5.6: (a) The training accuracy typically achieved with dataset 1 and 2 for experiment 3: target task and (b) comparison of the training accuracies for the two approaches (without TL vs. the proposed approach).

As mentioned earlier, the TL-based approach can learn faster with a smaller amount of data. To establish this fact, experiment 3 is further analyzed. A test is conducted where the proposed TL-CNN is trained from scratch with the target dataset independently, without passing on any knowledge from the source task. Like the previous test, 15% of the data are used for training and 85% of the data are used for testing the network. The data division is identical so that the performance can be compared on the same scale. In Figure 5.6(a), it can be observed that the TL-CNN without TL does not yield the same performance; it provides 83.3% accuracy for overall training. Furthermore, from Figure 5.6(b), it is observed that the proposed approach learns faster than the TL-CNN without the TL approach, and it also achieved 100% training accuracies, which lead to a better diagnostic performance (as discussed in Table 5.3). Hence, it can be concluded that the proposed model provides a faster convergence rate with enhanced classification performance.

To establish the validity and robustness of the proposed diagnostic framework, several approaches are considered from the literature [85,91,108] and adopted using an

experimental setup similar to the one in this case-study. The AS accuracy is considered to compare the performances of these methods. These methods include:

(1) WC + CNN-TL: the input is transformed into 2D Wavelet Coefficient (WC) matrices and then supplied to the TL-based deep architectures [108] to perform the TL-based analysis on the target task.

(2) TFI + CNN-TL: the input is transformed into several Time-Frequency Images (TFI) to create the multi-fusion input [91] and then passed to the TL-CNN architecture based on the CNN model mentioned in [91]. Finally, the TL-based approach from the proposed framework is adopted to perform the final analysis.

(3) RAW + CNN-TL: the input is directly fed to the adopted CNN architecture derived from [85], and then the knowledge attained from the source task is transferred to the target task to determine for final classification accuracy.

These methods are compared, and the improvement details of the proposed framework are discussed in Table 5.4. It is clear that the proposed framework outperforms these other state-of-the-art approaches [85,91,108], showing an improvement of 2.8% to 8.0% in terms of the AS score.

Table 5.4: Comparison of the diagnostic performance of case study 1.

Methods	Exp.	AS (%)	Improvement (%)
WC + CNN-TL	1	96.2	**3.8**
	2	96.3	**3.7**
	3	96.3	**3.7**
TFI + CNN-TL	1	97.2	**2.8**
	2	96.5	**3.5**
	3	96.5	**3.5**
RAW + CNN-TL	1	92.1	**7.9**
	2	92.3	**7.7**
	3	92.0	**8.0**

5.4.2 Case Study 2 – Dataset from the Self Designed Testbed

A second experiment was performed to confirm that the proposed model provides satisfactory results if evaluated using totally different dataset. In this case also, like case study 1, based on different load, and speed conditions, three different datasets were created, which has already been described in **Chapter 2**. Before preprocessing the data, an overlapping segmentation [10] is considered to generate more number of samples in order to utilize the full potential of DL algorithm. To generate results, a total of three experiments are conducted. In experiment 1, dataset 1 is used for the source task, while datasets 2 and 3 are considered in the target task. At first, the VI are attained from dataset 1, and then the TL-CNN is trained and tested with 90% and 10% of the data, respectively. The trained TL-CNN architecture with learned weights is then saved to use for the target task. In the target task, the VI based inputs are calculated from datasets 2 and 3. Then, the TL-CNN architecture with

the learned knowledge is used to adjust the target task's TL-CNN for measuring the final diagnostic performance. In this case, 15% of data from datasets 2 and 3 are used for training the network, and the remaining 85% of the data from both datasets are used for testing purposes. Similarly, for experiment 2, dataset 2 is considered for the source task and datasets 1 and 3 are considered for the target task. For experiment 3, dataset 3 is considered for the source task and datasets 1 and 2 are considered for the target task. Therefore, in this diagnostic framework, the dataset is divided separately for two different tasks, i.e., the source task and the target task, as shown in Table 5.5. Additionally, to remove the bias from the datasets, an equivalent number of samples are considered for each health type. Moreover, the model is trained for 3000 epochs for the source tasks. For each experiment, once the network is trained in the source task, the performance of the target task is observed to measure the final classification accuracy. The diagnostic performances of the three experiments are listed in Table 5.6.

Table 5.5: Data division.

Source task details	**Dataset**	**Train (90%)**		**Test (10%)**
		Training (80%)	**Validation (20%)**	
	1	1152 samples	288 samples	160 samples
	2	1152 samples	288 samples	160 samples
	3	1152 samples	288 samples	160 samples
Target task details		**Train (15%)**		**Test (85%)**
		Training (90%)	**Validation (10%)**	
	1	216 samples	24 samples	1360 samples
	2	216 samples	24 samples	1360 samples
	3	216 samples	24 samples	1360 samples

Table 5.6: Diagnostic performance of case study 2.

Exp.	**Source Task**	**Target Task**	**Health type**	**F1 (%)**	**AS(%)**
1	Dataset 1	Dataset 2, 3	NT	100	100
			IRT	100	
			ORT	100	
			BT	100	
2	Dataset 2	Dataset 3, 1	NT	100	100
			IRT	100	
			ORT	100	
			BT	100	
3	Dataset 3	Dataset 1, 2	NT	100	100
			IRT	100	
			ORT	100	
			BT	100	

5.5 Conclusions

This chapter proposes a TL-CNN based approach for condition monitoring of bearing with variable speed conditions. With the help of FDOST, we have created the invariant scenario from variable working conditions. Thus, the feature similarity comes into source and target task datasets. Therefore, by utilizing this similarity in patterns, TL based approach perfectly utilized the power of proposed CNN architecture for diagnosing the health types of bearing. By outperforming the conventional approaches, it stands as state-of-art among all the proposed approaches proposed into this dissertation.

However, this method has the limitations of explain ability, and interpretability from the statistical point of view related to the feature spaces. Therefore, the future direction of this work is to explore the possibilities of explanations and interpretations for a complete DL based explainable model.

Chapter 6

Contributions and Future Directions

6.1 Summary of Contributions

In this dissertation, the main goal is to improve the existing fault CM-FD approaches bearing for invariant working conditions by incorporating similarity among the feature spaces of different working conditions with explain ability, state-of-art accuracy, interpretability of the AI models, and computational efficiency.

In **Chapter 3**, to capture the information of variable working conditions from the vibration signals of bearing both at low and high frequencies, a computationally advanced version of ST called FDOST is proposed as the signal preprocessing step. First, statistical parameters are extracted as features from the time-frequency magnitudes and their corresponding phase angles of the FDOST coefficients. The extracted statistical features are then arranged as a feature matrix which can be regarded as input in the proceedings step. Thus, a carefully curated statistical feature pool extracted from unique FDOST patterns for different types of bearing faults is proposed in this study, which is helpful for boosting up the classification performance of the subsequent classifier.

In **Chapter 4**, a wrapper-based feature selector named as Boruta is utilized to find the best features from the extracted statistical feature pool. This algorithm can justify the selection of each feature attribute with the help of an embedded RF classifier. Thus, the feature selection process is easily interpretable in the proposed bearing fault diagnosis model. In

addition to interpretable feature selection process, a feature filtration technique is proposed by using SRCC to create a reduced feature set for classifier which can produce bias free results. Thus, it helps the classifier to avoid the multicollinearity trap. Finally, a correlation between the filtered feature pool and results of a non-parametric k-NN classifier is presented in this work, i.e., the predictions of k-NN are explained in context of SHAP values.

In **Chapter 5**, to automate the feature extraction, and selection process, after capturing the invariant patterns from FDOST coefficients from variable working conditions from the vibration signals both at low and high frequencies, these 2D FDOST coefficients are used as a 2D time-frequency image. Then, these are converted into grayscale, coined as Vibration Imaging (VI), to utilize CNN efficiently to automate the feature extraction and selection process. Furthermore, to bring the computational efficiency to the DL based approaches, a CNN-TL model is proposed for fault diagnosis for vibration signals of bearings with variations of the shaft speed (e.g., RPMs).

All these approaches are suitably validated with extensive experiments and simulations with two different bearing datasets discussed into chapter 2, which justifies the potential of the proposed methodology over existing approaches in terms of achieving satisfactory theoretical results compared to experiment.

With all these proposed approaches, we can reach to the following decisions:

(1) If we want a model which can create similarity among the feature spaces of different working conditions with explain ability, state-of-art accuracy, and interpretability of the AI models, then we can rely on the proposed solution given into **Chapter 4**.

(2) If the demand is computation efficiency over the explain ability, and interpretability of the model, then we can consider the proposed solution presented into **Chapter 5**. This model includes similarity among the feature spaces of different working conditions by generating similar visual patterns, state-of-art accuracy, and computational efficiency.

6.2 Future Directions

This dissertation gives us clear idea about several future directions of expanding the proposed solutions:

(1) For the ML based interpretable model, we can further investigate the feature space, and interpretability of the AI based models by incorporating the causal analysis of the extracted features. This will provide us the direction of bringing the explain ability on the data preprocessing step as well.

(2) For the TL based model, the future direction is to explore the possibilities of explanations and interpretations for the proposed CNN architecture to understand feature contribution from the input of both source and target task.

(3) To create a more robust solution, the data preprocessing step can be further improved. Moreover, the invariant images created by the FDOST coefficients can be converted into sound to analyze the melody patterns of different health types. Therefore, individual sound can be utilized for feature extraction or TL based classification. Besides, a hybrid approach by considering both the FDOST coefficient-based vibration images, and melody analysis would be an interesting solution to create a robust condition monitoring model for bearing.

List of Research Publications

Peer-Reviewed Journals (Index: SCI-E)

Published

(1) **Hasan M.J.;** Sohaib M.; Kim JM. **(2021).** An Explainable Artificial Intelligence based Diagnostic Framework for Bearing, Sensors 21.12: 4070. (SCIE, IF 3.576)

(2) **Hasan, M. J**.; Ahmed, Z.; Rai, A.; Kim, J. M. **(2021)**. A Fault Diagnosis Framework for Centrifugal Pumps by Scalogram Based Imaging and Deep Learning, IEEE Access (SCIE, IF 3.745).

(3) **Hasan, M. J**.; Sohaib M., Kim, J. M. **(2020)**. A Multitask-Aided Transfer Learning-Based Diagnostic Framework for Bearings under Inconsistent Working Conditions. Sensors 2020, 20 (24), 7205. (SCIE,IF 3.576)

(4) **Hasan, M. J**.; Shon, D.; Im, K.; Choi, H.-K.; Yoo, D.-S.; Kim, J.-M. **(2020).** Sleep State Classification Using Power Spectral Density and Residual Neural Network with Multichannel EEG Signals. Applied Sciences. 2020, 10, 7639. (SCIE,IF 2.679)

(5) **Hasan, M. J**.; Islam, M. M.; & Kim, J. M. **(2020)**. Multi-Sensor Fusion-based Time-Frequency Imaging and Transfer Learning for Spherical Tank Crack Diagnosis Under Variable Pressure Conditions. *Measurement* (2020): 108478.(SCIE,IF 3.927)

(6) **Hasan, M. J**.; Kim, J.; Kim, C. H.; & Kim, J. M. **(2020)**. Health State Classification of a Spherical Tank Using a Hybrid Bag of Features and K-Nearest Neighbor. *Applied Sciences*, *10*(7), 2525. (SCIE,IF 2.679)

(7) **Hasan, M.J**.; Kim, J. M. **(2019)**. A Hybrid Feature Pool-Based Emotional Stress State Detection Algorithm Using EEG Signals. Brain Sci. 2019, 9, 376. (SCIE, IF 3.394)

(8) **Hasan, M.J**.; Kim, J. M. **(2019)**. Fault Detection of a Spherical Tank Using a Genetic Algorithm-Based Hybrid Feature Pool, and k-Nearest Neighbor Algorithm. *Energies* 2019, *12*, 991. (SCIE, IF 3.004) **[Featured Journal]**

(9) **Hasan, M. J**.; Islam, M. M.; & Kim, J. M. **(2019)**. Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. Measurement, 138, 620-631. (SCIE,IF 3.927)

(10) **Hasan, M.J**.; Kim, J.-M. **(2018)**. Bearing Fault Diagnosis under Variable Rotational Speeds Using Stockwell Transform-Based Vibration Imaging and Transfer Learning. *Appl. Sci.* 2018, *8*, 2357. (SCIE,IF 2.679)

Conferences

(1) **Hasan, M. J.**; Kim J. **(2019).** Deep Convolutional Neural Network with 2D Spectral Energy Maps for Fault Diagnosis of Gearboxes Under Variable Speed, Third

Mediterranean Conference, MedPRAI 2019, December 22–23, 2019, Istanbul, Turkey. (extended as Scopus proceeding)

(2) **Hasan, M. J**.; Kim, Jaeyoung & Kim, J. M. **(2019)**. Health State Classification of a Spherical Tank Using a Hybrid Set of Feature Analysis, The 14th KIPS International Conference on Ubiquitous Information Technologies and Applications (CUTE 2019), Dec 18-20, 2019, Macau, China.

(3) **Hasan, M. J.**; Sohaib M.; Kim JM. **(2018)**. 1D CNN-Based Transfer Learning Model for Bearing Fault Diagnosis Under Variable Working Conditions, International Conference on Computational Intelligence in Information System (CIIS - 2018), November 16 – 18, 2018, Gadong, Brunei. (extended as Scopus proceeding)

(4) **Hasan, M. J**.; Islam, M. M.; Kim, C. H.; & Kim, J. M. **(2017)**. An Improved Feature Extraction Method using Global Neighborhood Structure Mapping for Textures Under Varying Illumination, THE ENGINEERING AND ARTS SOCIETY IN KOREA-2017,(2017), 100-102, Ulsan, Republic of Korea.

(5) **Hasan, M. J.**; Appana, D. K.; Muhammad, S.; & Kim, J. M. **(2017)**. Automated Feature Extraction for detecting Brain Tumors in Magnetic Resonance Imaging using a Genetic Algorithm, 12th International Forum on Strategic Technology (IFOST - 2017), Ulsan, Republic of Korea.

LNCS Book Chapters (Index: Scopus)

(1) **Hasan M.J**.; Kim J. (2020). Deep Convolutional Neural Network with 2D Spectral Energy Maps for Fault Diagnosis of Gearboxes Under Variable Speed. In: Djeddi C., Jamil A., Siddiqi I. (eds) Pattern Recognition and Artificial Intelligence. MedPRAI 2019. Communications in Computer and Information Science, vol 1144. Springer, Cham. (Scopus)

(2) **Hasan M.J.**; Sohaib M.; Kim JM. (2019). 1D CNN-Based Transfer Learning Model for Bearing Fault Diagnosis Under Variable Working Conditions. In: Omar S., Haji Suhaili W., Phon-Amnuaisuk S. (eds) Computational Intelligence in Information Systems. CIIS 2018. Advances in Intelligent Systems and Computing, vol 888. Springer, Cham (Scopus)

References

[1] X. Yan, Y. Liu, M. Jia, Y. Zhu, A multi-stage hybrid fault diagnosis approach for rolling element bearing under various working conditions, IEEE Access. 7 (2019) 138426–138441.

[2] X. Yan, M. Jia, A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing, Neurocomputing. 313 (2018) 47–64.

[3] Y. Lei, Z. He, Y. Zi, A new approach to intelligent fault diagnosis of rotating machinery, Expert Syst. Appl. 35 (2008) 1593–1600.

[4] M.J. Hasan, M. Sohaib, J.-M. Kim, A Multitask-Aided Transfer Learning-Based Diagnostic Framework for Bearings under Inconsistent Working Conditions, Sensors. 20 (2020) 7205.

[5] L. Cui, J. Huang, F. Zhang, Quantitative and localization diagnosis of a defective ball bearing based on vertical–horizontal synchronization signal analysis, IEEE Trans. Ind. Electron. 64 (2017) 8695–8706.

[6] J. Tian, Y. Ai, C. Fei, M. Zhao, F. Zhang, Z. Wang, Fault diagnosis of intershaft bearings using fusion information exergy distance method, Shock Vib. 2018 (2018).

[7] D.T. Hoang, H.J. Kang, A Motor Current Signal-Based Bearing Fault Diagnosis Using Deep Learning and Information Fusion, IEEE Trans. Instrum. Meas. 69 (2019) 3325–3333.

[8] W. Mao, J. Chen, X. Liang, X. Zhang, A new online detection approach for rolling bearing incipient fault via self-adaptive deep feature matching, IEEE Trans. Instrum. Meas. 69 (2019) 443–456.

[9] A. Rai, J.-M. Kim, A novel health indicator based on the Lyapunov exponent, a probabilistic self-organizing map, and the Gini-Simpson index for calculating the RUL of bearings, Measurement. (2020) 108002.

[10] M.J. Hasan, M.M.M. Islam, J.M. Kim, Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions, Meas. J. Int. Meas. Confed. 138 (2019) 620–631. https://doi.org/10.1016/j.measurement.2019.02.075.

[11] M. Sohaib, C.-H. Kim, J.-M. Kim, A Hybrid Feature Model and Deep-Learning-Based Bearing Fault Diagnosis, Sensors. 17 (2017) 2876. https://doi.org/10.3390/s17122876.

[12] M. Meng, Y.J. Chua, E. Wouterson, C.P.K. Ong, Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks, Neurocomputing. 257 (2017) 128–135.

[13] H. Zheng, R. Wang, Y. Yang, Y. Li, M. Xu, Intelligent fault identification based on multisource domain generalization towards actual diagnosis scenario, IEEE Trans.

Ind. Electron. 67 (2019) 1293–1304.

[14] H. Oh, J.H. Jung, B.C. Jeon, B.D. Youn, Scalable and unsupervised feature engineering using vibration-imaging and deep learning for rotor system diagnosis, IEEE Trans. Ind. Electron. 65 (2017) 3539–3549.

[15] C.W.R. University, Case Western Bearing Data Center, (2017). http://csegroups.case.edu/bearingdatacenter/pages/download-data-file.

[16] M. Kang, J. Kim, J. Kim, High-Performance and Energy-Efficient Fault Diagnosis Using Effective Envelope Analysis Processing Unit, IEEE Trans. Power Electron. 30 (2015) 2763–2776.

[17] M. Sohaib, J.-M. Kim, Fault diagnosis of rotary machine bearings under inconsistent working conditions, IEEE Trans. Instrum. Meas. 69 (2019) 3334–3347.

[18] S.-Y. Shao, W.-J. Sun, R.-Q. Yan, P. Wang, R.X. Gao, A deep learning approach for fault diagnosis of induction motors in manufacturing, Chinese J. Mech. Eng. 30 (2017) 1347–1356.

[19] Y. Sun, S. Li, X. Wang, Bearing fault diagnosis based on EMD and improved Chebyshev distance in SDP image, Measurement. (2021) 109100.

[20] J. Ben Ali, N. Fnaiech, L. Saidi, B. Chebel-Morello, F. Fnaiech, Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals, Appl. Acoust. 89 (2015) 16–27.

[21] L.-Y. Zhao, L. Wang, R.-Q. Yan, Rolling bearing fault diagnosis based on wavelet packet decomposition and multi-scale permutation entropy, Entropy. 17 (2015) 6447–6461.

[22] Z. Qiao, Y. Liu, Y. Liao, An improved method of EWT and its application in rolling bearings fault diagnosis, Shock Vib. 2020 (2020).

[23] R. Gu, J. Chen, R. Hong, H. Wang, W. Wu, Incipient fault diagnosis of rolling bearings based on adaptive variational mode decomposition and Teager energy operator, Measurement. 149 (2020) 106941.

[24] Y. Cheng, M. Lin, J. Wu, H. Zhu, X. Shao, Intelligent fault diagnosis of rotating machinery based on continuous wavelet transform-local binary convolutional neural network, Knowledge-Based Syst. 216 (2021) 106796. https://doi.org/https://doi.org/10.1016/j.knosys.2021.106796.

[25] M. Kang, J. Kim, J.M. Kim, A.C.C. Tan, E.Y. Kim, B.K. Choi, Reliable fault diagnosis for low-speed bearings using individually trained support vector machines with kernel discriminative feature analysis, IEEE Trans. Power Electron. 30 (2015) 2786–2797. https://doi.org/10.1109/TPEL.2014.2358494.

[26] A. Rai, S.H. Upadhyay, A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings, Tribol. Int. 96 (2016) 289–306.

[27] S.A. Khan, J.-M. Kim, Rotational speed invariant fault diagnosis in bearings using vibration signal imaging and local binary patterns, J. Acoust. Soc. Am. 139 (2016)

EL100–EL104. https://doi.org/10.1121/1.4945818.

[28] M.M.M. Islam, J. Myon, Time–frequency envelope analysis-based sub-band selection and probabilistic support vector machines for multi-fault diagnosis of low-speed bearings, J. Ambient Intell. Humaniz. Comput. 0 (2017) 0. https://doi.org/10.1007/s12652-017-0585-2.

[29] J. Qu, Z. Zhang, T. Gong, A novel intelligent method for mechanical fault diagnosis based on dual-tree complex wavelet packet transform and multiple classifier fusion, Neurocomputing. 171 (2016) 837–853.

[30] G. Chen, F. Liu, W. Huang, Sparse discriminant manifold projections for bearing fault diagnosis, J. Sound Vib. 399 (2017) 330–344.

[31] X. Zheng, Y. Wei, J. Liu, H. Jiang, Multi-synchrosqueezing S-transform for fault diagnosis in rolling bearings, Meas. Sci. Technol. 32 (2020) 25013. https://doi.org/10.1088/1361-6501/abb620.

[32] U. Battisti, L. Riba, Window-dependent bases for efficient representations of the Stockwell transform, Appl. Comput. Harmon. Anal. 40 (2016) 292–320. https://doi.org/10.1016/j.acha.2015.02.002.

[33] T.S. Breusch, A.R. Pagan, A simple test for heteroscedasticity and random coefficient variation, Econom. J. Econom. Soc. (1979) 1287–1294.

[34] S. Mishra, C.N. Bhende, B.K. Panigrahi, Detection and classification of power quality disturbances using S-transform and probabilistic neural network, IEEE Trans. Power Deliv. 23 (2007) 280–287.

[35] R.G. Stockwell, L. Mansinha, R.P. Lowe, Localization of the complex spectrum: The S transform, IEEE Trans. Signal Process. 44 (1996) 998–1001. https://doi.org/10.1109/78.492555.

[36] R.G. Stockwell, A basis for efficient representation of the S-transform, Digit. Signal Process. A Rev. J. 17 (2007) 371–393. https://doi.org/10.1016/j.dsp.2006.04.006.

[37] B. Patel, A new FDOST entropy based intelligent digital relaying for detection, classification and localization of faults on the hybrid transmission line, Electr. Power Syst. Res. 157 (2018) 39–47.

[38] M.D. Kaba, M.K. Camlibel, Y. Wang, J. Orchard, Fast discrete orthonormal Stockwell transform, SIAM J. Sci. Comput. 31 (2009) 4000–4012.

[39] U. Battisti, L. Riba, Window-dependent bases for efficient representations of the Stockwell transform, Appl. Comput. Harmon. Anal. 40 (2016) 292–320. https://doi.org/10.1016/j.acha.2015.02.002.

[40] M.J. Hasan, J.M. Kim, Fault detection of a spherical tank using a genetic algorithm-based hybrid feature pool and k-nearest neighbor algorithm, Energies. 12 (2019). https://doi.org/10.3390/en12060991.

[41] M.J. Hasan, J. Kim, C.H. Kim, J.-M. Kim, Health State Classification of a Spherical Tank

Using a Hybrid Bag of Features and K-Nearest Neighbor, Appl. Sci. 10 (2020) 2525. https://doi.org/10.3390/app10072525.

[42] P.-E. Danielsson, Euclidean distance mapping, Comput. Graph. Image Process. 14 (1980) 227–248.

[43] L. Greche, M. Jazouli, N. Es-Sbai, A. Majda, A. Zarghili, Comparison between Euclidean and Manhattan distance measure for facial expressions classification, in: 2017 Int. Conf. Wirel. Technol. Embed. Intell. Syst., IEEE, 2017: pp. 1–4.

[44] M.J. Hasan, J.-M. Kim, Bearing Fault Diagnosis under Variable Rotational Speeds Using Stockwell Transform-Based Vibration Imaging and Transfer Learning, Appl. Sci. 8 (2018) 2357. https://doi.org/10.3390/app8122357.

[45] S. Bag, A.K. Pradhan, S. Das, S. Dalai, B. Chatterjee, S-transform aided random forest based PD location detection employing signature of optical sensor, IEEE Trans. Power Deliv. 34 (2018) 1261–1268.

[46] M. Kang, J. Kim, L.M. Wills, J.M. Kim, Time-varying and multiresolution envelope analysis and discriminative feature analysis for bearing fault diagnosis, IEEE Trans. Ind. Electron. 62 (2015) 7749–7761. https://doi.org/10.1109/TIE.2015.2460242.

[47] R. Islam, S.A. Khan, J.M. Kim, Discriminant Feature Distribution Analysis-Based Hybrid Feature Selection for Online Bearing Fault Diagnosis in Induction Motors, J. Sensors. 2016 (2016). https://doi.org/10.1155/2016/7145715.

[48] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and F-score, with implication for evaluation, in: Eur. Conf. Inf. Retr., Springer, 2005: pp. 345–359.

[49] A. Luque, A. Carrasco, A. Martín, A. de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, Pattern Recognit. 91 (2019) 216–231.

[50] B.R. Nayana, P. Geethanjali, Analysis of statistical time-domain features effectiveness in identification of bearing faults from vibration signal, IEEE Sens. J. 17 (2017) 5618–5625.

[51] B. Peng, S. Wan, Y. Bi, B. Xue, M. Zhang, Automatic Feature Extraction and Construction Using Genetic Programming for Rotating Machinery Fault Diagnosis, IEEE Trans. Cybern. (2020) 1–15. https://doi.org/10.1109/TCYB.2020.3032945.

[52] K. Kaplan, Y. Kaya, M. Kuncan, M.R. Minaz, H.M. Ertunç, An improved feature extraction method using texture analysis with LBP for bearing fault diagnosis, Appl. Soft Comput. 87 (2020) 106019.

[53] R. Leardi, Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection, J. Chemom. 8 (1994) 65–79.

[54] S. Oreski, G. Oreski, Genetic algorithm-based heuristic for feature selection in credit risk assessment, Expert Syst. Appl. 41 (2014) 2052–2064.

[55] M.A. Ambusaidi, X. He, P. Nanda, Z. Tan, Building an intrusion detection system using

a filter-based feature selection algorithm, IEEE Trans. Comput. 65 (2016) 2986–2998.

[56] P.C. Chu, J.E. Beasley, A genetic algorithm for the multidimensional knapsack problem, J. Heuristics. 4 (1998) 63–86.

[57] H. Xie, J. Li, H. Xue, A survey of dimensionality reduction techniques based on random projection, (2017) 1–35. https://doi.org/arXiv:1403.2877.

[58] A. Refahi Oskouei, H. Heidary, M. Ahmadi, M. Farajpur, Unsupervised acoustic emission data clustering for the analysis of damage mechanisms in glass/polyester composites, Mater. Des. 37 (2012) 416–422. https://doi.org/10.1016/j.matdes.2012.01.018.

[59] M.B. Kursa, A. Jankowski, W.R. Rudnicki, Boruta–a system for feature selection, Fundam. Informaticae. 101 (2010) 271–285.

[60] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J Stat Softw. 36 (2010) 1–13.

[61] J.H. Zar, Spearman rank correlation, Encycl. Biostat. 7 (2005).

[62] J.H. Zar, Significance testing of the Spearman rank correlation coefficient, J. Am. Stat. Assoc. 67 (1972) 578–580.

[63] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, ArXiv Prepr. ArXiv1705.07874. (2017).

[64] L.S. Shapley, A Value for n-person Games", Contribution to the Theory of Games vol. II (HW Kuhn and AW Tucker eds), Ann. Math. Stud. 28 (n.d.).

[65] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Trans. Syst. Man. Cybern. 21 (1991) 660–674.

[66] S.B. Kotsiantis, Decision trees: a recent overview, Artif. Intell. Rev. 39 (2013) 261–283.

[67] P.Y. Chen, M. Smithson, P.M. Popovich, Correlation: Parametric and nonparametric measures, Sage, 2002.

[68] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise Reduct. Speech Process., Springer, 2009: pp. 1–4.

[69] D.P. Francis, A.J.S. Coats, D.G. Gibson, How high can a correlation coefficient be? Effects of limited reproducibility of common cardiological measures, Int. J. Cardiol. 69 (1999) 185–189.

[70] S. Lipovetsky, M. Conklin, Analysis of regression in game theory approach, Appl. Stoch. Model. Bus. Ind. 17 (2001) 319–330.

[71] A. Datta, S. Sen, Y. Zick, Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems, in: 2016 IEEE Symp. Secur. Priv., IEEE, 2016: pp. 598–617.

[72] M.T. Ribeiro, S. Singh, C. Guestrin, " Why should i trust you?" Explaining the predictions of any classifier, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016: pp. 1135–1144.

[73] P. Domingos, The role of Occam's razor in knowledge discovery, Data Min. Knowl. Discov. 3 (1999) 409–425.

[74] J. Feldman, The simplicity principle in perception and cognition, Wiley Interdiscip. Rev. Cogn. Sci. 7 (2016) 330–340.

[75] R. Nilsson, J.M. Peña, J. Björkegren, J. Tegnér, Consistent feature selection for pattern recognition in polynomial time, J. Mach. Learn. Res. 8 (2007) 589–612.

[76] A. Alin, Multicollinearity, Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 370–374.

[77] H. Yigit, A weighting approach for KNN classifier, 2013 Int. Conf. Electron. Comput. Comput. ICECCO 2013. 1 (2013) 228–231. https://doi.org/10.1109/ICECCO.2013.6718270.

[78] O.S. Ahmed, S.E. Franklin, M.A. Wulder, J.C. White, Extending airborne lidar-derived estimates of forest canopy cover and height over large areas using knn with landsat time series data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (2015) 3489–3496.

[79] H. Zenil, N.A. Kiani, A.A. Zea, J. Tegnér, Causal deconvolution by algorithmic generative models, Nat. Mach. Intell. 1 (2019) 58–66.

[80] Remodelling machine learning: An AI that thinks like a scientist, Nat. Mach. Intell. (2019). https://doi.org/10.1038/s42256-019-0026-3.

[81] M.V. García, J.L. Aznarte, Shapley additive explanations for NO2 forecasting, Ecol. Inform. 56 (2020) 101039.

[82] M.T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, ArXiv Prepr. ArXiv1606.05386. (2016).

[83] D.K. Appana, W. Ahmad, J.-M. Kim, Speed Invariant Bearing Fault Characterization Using Convolutional Neural Networks, in: S. Phon-Amnuaisuk, S.-P. Ang, S.-Y. Lee (Eds.), Multi-Disciplinary Trends Artif. Intell., Springer International Publishing, Cham, 2017: pp. 189–198.

[84] A. Prosvirin, J. Kim, J.M. Kim, Bearing fault diagnosis based on convolutional neural networks with kurtogram representation of acoustic emission signals, Lect. Notes Electr. Eng. 474 (2018) 21–26. https://doi.org/10.1007/978-981-10-7605-3_4.

[85] R. Zhang, H. Tao, L. Wu, Y. Guan, Transfer Learning with Neural Networks for Bearing Fault Diagnosis in Changing Working Conditions, IEEE Access. 5 (2017) 14347–14357. https://doi.org/10.1109/ACCESS.2017.2720965.

[86] Y. LeCun, L.D. Jackel, L. Bottou, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, U.A. Muller, E. Sackinger, P. Simard, Learning algorithms for classification: A comparison on handwritten digit recognition, Neural Networks Stat. Mech. Perspect. 261 (1995) 276.

[87] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444. https://doi.org/10.1038/nature14539.

[88] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014)

1929–1958.

[89] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, ArXiv Prepr. ArXiv1502.03167. (2015).

[90] G.E. Dahl, T.N. Sainath, G.E. Hinton, Improving deep neural networks for LVCSR using rectified linear units and dropout, in: 2013 IEEE Int. Conf. Acoust. Speech Signal Process., IEEE, 2013: pp. 8609–8613.

[91] J. Wang, Z. Mo, H. Zhang, Q. Miao, A deep learning method for bearing fault diagnosis based on time-frequency image, IEEE Access. 7 (2019) 42373–42383.

[92] H. Wang, J. Xu, R. Yan, R.X. Gao, A New Intelligent Bearing Fault Diagnosis Method Using SDP Representation and SE-CNN, IEEE Trans. Instrum. Meas. 69 (2019) 2377–2389.

[93] L. Jing, M. Zhao, P. Li, X. Xu, A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox, Measurement. 111 (2017) 1–10.

[94] J. Ma, F. Wu, J. Zhu, D. Xu, D. Kong, A pre-trained convolutional neural network based method for thyroid nodule diagnosis, Ultrasonics. 73 (2017) 221–230.

[95] J. Brownlee, What is the Difference Between a Batch and an Epoch in a Neural Network?, Mach. Learn. Mastery. (2018).

[96] M. Dąbrowski, T. Michalik, How effective is Transfer Learning method for image classification, in: Position Pap. 2017 Fed. Conf. Comput. Sci. Inf. Syst., 2017: pp. 3–9. https://doi.org/10.15439/2017f526.

[97] D.-T. Hoang, H.-J. Kang, Rolling element bearing fault diagnosis using convolutional neural network and vibration image, Cogn. Syst. Res. 53 (2019) 42–50.

[98] M. Sohaib, J.-M. Kim, A robust deep learning based fault diagnosis of rotary machine bearings, Adv. Sci. Lett. 23 (2017) 12797–12801.

[99] M. Zhao, M. Kang, B. Tang, M. Pecht, Deep Residual Networks with Dynamically Weighted Wavelet Coefficients for Fault Diagnosis of Planetary Gearboxes, IEEE Trans. Ind. Electron. 65 (2018) 4290–4300. https://doi.org/10.1109/TIE.2017.2762639.

[100] M. Misiti, Y. Misiti, G. Oppenheim, J.-M. Poggi, Wavelets and their Applications, John Wiley & Sons, 2013.

[101] Y. Falamarzi, N. Palizdan, Y.F. Huang, T.S. Lee, Estimating evapotranspiration from temperature and wind speed data using artificial and wavelet neural networks (WNNs), Agric. Water Manag. 140 (2014) 26–36.

[102] Ö. Türk, M.S. Özerdem, Epilepsy detection by using scalogram based convolutional neural network from EEG signals, Brain Sci. 9 (2019) 115.

[103] R. Bala, K.M. Braun, Color-to-grayscale conversion to maintain discriminability, in: Color Imaging IX Process. Hardcopy, Appl., International Society for Optics and Photonics, 2003: pp. 196–202.

[104] Y. LeCun, LeNet-5, convolutional neural networks, URL Http//Yann. Lecun.

Com/Exdb/Lenet. 20 (2015) 5.

[105] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, ArXiv Prepr. ArXiv1412.6980. (2014).

[106] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[107] M.W. Browne, Cross-validation methods, J. Math. Psychol. 44 (2000) 108–132.

[108] S. Guo, B. Zhang, T. Yang, D. Lyu, W. Gao, Multitask Convolutional Neural Network With Information Fusion for Bearing Fault Diagnosis and Localization, IEEE Trans. Ind. Electron. 67 (2019) 8005–8015.

“ If I have seen further than others, it is by standing upon the shoulders of giants.”

Isaac Newton.

THE END.