Surekha Medapati

Master's Thesis

Experimental Analysis of Multi-Sensor Data for Motor Condition Monitoring.

School of Technology and Innovation.
Master's Thesis of
Sustainable and Autonomous Systems.

Vaasa 2026

UNIVERSITY OF VAASA
School of Technology and Innovation
Author: Surekha Medapati
Title of the thesis: Master's Thesis : Experimental Analysis of Multi-Sensor Data for Motor Condition Monitoring.
Degree: Master of Computer Science
Discipline: Sustainable and Autonomous Systems
Supervisor: Professor Petri Välisuo
Venna Pradeep Bonda, Doctoral Student
R&D Team manager Daniel Barg ABB Oy
Principal R &D Engineer Hauke Carstensen ABB Oy
Year: 2026 **Pages:** 108

ABSTRACT :

Electric motor bearing failures constitute one of the most frequent causes of unplanned industrial downtime, yet how can industrial operators maintain absolute confidence in AI-driven diagnostic systems when electric motors constantly shift speeds and evolve over years of service? While Deep learning has demonstrated significant potential in laboratory settings, most models remain "black boxes" validated on narrow, single-session datasets that fail to reflect the unpredictable nature of real-world industrial environments.

This thesis bridges the gap between academic theory and industrial reliability by presenting a robust machine learning framework for bearing fault diagnosis, validated against a four-year longitudinal dataset (2022-2025), spanning from 0-1500 RPM. Moving beyond traditional scalar thresholds, this study evaluates four supervised classification models and three anomaly detectors across 57 unique motor speeds using a rigorous Leave-One-Speed-Out (LOSO) protocol to ensure that every evaluation speed was unseen during training phase of AI models.

The experimental results show that traditional fault indicators like kurtosis and RMS fail to generalize across speed folds as a multi-scale Residual Raw CNN achieves a 99.81% binary and 98.47% three-class accuracy. Furthermore, this work utilizes DeepSHAP attribution to "look inside" these models providing the physical evidence that X-radial axis remains the primary diagnostic indicator except at high-speed exceeding 1200 RPM. Among the unsupervised detectors, CNN autoencoder delivered the most operationally balanced performance with a 92.89% detection rate and only a 5.25% false alarm rate. Together these findings demonstrate that trustworthy, interpretable, longitudinally stable and speed robust bearing health monitoring is achievable within the ABB Detect-Predict-Recommend operational framework.

KEYWORDS: (Bearing fault diagnosis, convolutional neural network, condition monitoring, explainable AI, anomaly detection, predictive maintenance, variable-speed drives).

Contents

Figures

Tables

Abbreviations

AC	Alternating Current
ABB	ASEA Brown Boveri
ACS355	ABB Variable Speed Drive model
AI	Artificial Intelligence
BPFI	Ball Pass Frequency Inner Race
BPFO	Ball Pass Frequency Outer Race
BSF	Ball Spin Frequency
CAGR	Compound Annual Growth Rate
CNN	Convolutional Neural Network
CNN-FFT	Convolutional Neural Network with FFT features
CWRU	Case Western Reserve University
DC	Direct Current
DeepSHAP	Deep Learning-based SHAP attribution method
DL	Deep Learning
DR	Detection Rate
F1	F1 Score
FAR	False Acceptance Rate
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
FTF	Fundamental Train Frequency
Hz	Hertz
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
LOSO	Leave-One-Speed-Out cross-validation
M3AA	ABB asynchronous motor model designation
MFPT	Machinery Failure Prevention Technology
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
RMS	Root Mean Square
RPM	Revolutions Per Minute
RQ1	Research Question 1
RQ2	Research Question 2
RQ3	Research Question 3
RQ4	Research Question 4

SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
t-SNE	t-Distributed Stochastic Neighbour Embedding
USD	United States Dollar
VSD	Variable Speed Drive
XAI	Explainable Artificial Intelligence

1 Introduction

Electric motors are everywhere. They power manufacturing lines, pumps, fans, compressors and any rotating machine. Nowadays it is hard to think of any production facility without motors. When these machines break down, especially high priority assets unexpectedly everything stops. Production halts, maintenance engineers are called in. Every hour of waiting translates directly into the company's financial loss. To avoid this type of unexpected failures industries started scheduling maintenance at regular intervals. However, scheduling maintenance has its own drawback.

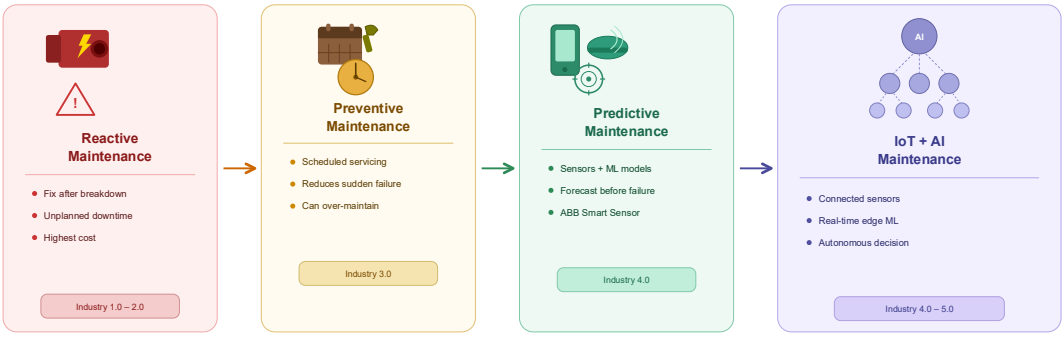


Figure 1. Industry Maintenance evolution from industry 1.0 to 5.0 (Sanakkayala et al., 2022).

Figure 1 represents the four-stage progression of Industrial maintenance across industrial revolution from 1.0 to 5.0. During the earliest manufacturing era, that is Industry 1.0 to 2.0 maintenance was entirely reactive. This process leads to unplanned downtime. Motors simply run until they fail and then engineers are called to fix these machines. Industry 3.0 approach is a preventive maintenance. This approach reduces catastrophic failures, but it has own inefficiencies. Machines were often serviced when they did not need it, which was wasting labor resources and time.

Industry 4.0 marked as a fundamental shift as embedded sensors, digital communication networks and machine learning algorithms. The key advantage is monitoring the motor condition continuously. The main advantage is it also predicts machine failure before they occur. Predictive maintenance heavily depends on IoT sensors, data storage and digitalization. Industry estimates suggest that AI driven predictive maintenance can decrease the maintenance cost by 10-40%. It also reduces time by 50% which in turn extends equipment life by 20-40% (Dilda et al., 2017). The industry 4.0 system continuously collects data from each motor. It triggers alert when there is an abnormality that is detected. It allows engineers to intervene before a catastrophic failure occurs. This data allows engineers and experts to pinpoint the exact reason for failure and precisely decode what, where and the cause of the failure. This becomes very powerful for large industrial fleets.

When hundreds of motors are operating simultaneously each carries unique Identity numbers and a faulty motor is easily located instantly without the manual inspection of every unit. The data is stored digitally so the customers can monitor their entire fleet remotely from a mobile phone or laptop. Scaling up this type of system is very easy and covers more motors as industry expands. It just requires software updates.

Declining sensor prices, growing adaptation of cloud-edge architectures and industry digitization collectively helped in the acceleration of Industry 4.0 development. According to Mordor Intelligence (2026), the predictive maintenance market was valued at 14.09 billion in 2025, and it is estimated to grow from 18.9 billion USD in 2026 to 82.17 billion by 2031 at CAGR of 34.41 percentage. Advancements in AI, especially Agentic AI, suggest that enterprise industries shift from Industry 4.0 to 5.0 Integrating real time AI inference with autonomous decision making and the human expert at supervisory level.

1.1 ABB Company Oy and the Industrial Context

The client of this research thesis is ABB Oy IEC Low Voltage motors. ABB is a global technology leader in automation and electrification and stands for sustainable and sufficient resource future (Sustainability | ABB, 2025) according to ABB's own global survey 2023 of over 3200 plant maintenance leaders found that unplanned downtime costs (ABB,2023). The financial results of industrial equipment failure are 125,000 USD per hour and raised to 170,000 per hour (ABB Ability Digital Powertrain Enabling Hardware, 2026). The ABB is already in industry 4.0 predictive maintenance infrastructure. With ABB Ability Digital power train Condition monitoring of rotating equipment fitted with ABB Ability Smart Sensors.

Figure 2. ABB Ability Smart Sensor.

Figure 2 to is the ABB Ability Smart Sensor is a condition monitoring device designed to track the health and performance of rotating machines. It monitors the electric machines such as motors, pumps, and fans. The main advantage of this type of solution is that it removes the need for manual inspections. Without interrupting normal operations, users can monitor motor heath condition.

For motors specifically the sensor keeps track of a wide range of health indicators. These can be grouped into two categories:

Health indicators:

- Overall condition score
- Vibration velocity
- Bearing condition
- Skin temperature

Operating parameters:

- 3-axis vibration
- Rotational speed in RPM
- Total running time
- Supply frequency and output power

The data collected by the sensors sent to the cloud for further processing and analysis. One of the examples of this ABB condition monitoring services saved the Mokrá cement plant up to 210K USD and increased its operational efficiency (ABB, 2021). ABB organizes its predictive maintenance around a 3-stage operational framework: Detect, Predict and Recommend (ABB Ability Digital Powertrain Enabling Hardware, 2026). In the Detect stage the data is collected securely and analyzed to capture the early deviations from the motor normal behavior. The predict stage applies advanced analytics to identify the failure risks and classify the severity. The Recommend stage provides guidance to maintenance team to act before the occurrence of downtime. ABB condition monitoring

has a digital platform which is a self-service platform that provides a real time dashboard which shows the motor health condition and analytics powered alerts. The data is collected securely and stored in cloud platforms or can be on edge device to customer's in-house storage system which is depending on the customers' preference. This ABB's operational framework is the main motivation of the machine learning system developed in this thesis.

1.2 Research Problem:

The main faults of electrical machines are Electrical and Mechanical faults. In electrical faults they have unbalanced voltage or currents, earth faults, die electric and inter turn short circuits. In mechanical faults mainly has Gear box faults, Rotor bar damage, Air-gap eccentricity, stator winding damage, misaligned shaft, unbalanced rotor and Bearing faults. In this critical motor infrastructure rolling element bearing as the most failure component. Bearing faults are the most common causes of unplanned stoppages of all rotating machines, accounting for approximately 40-50% (Randall & Antoni, 2011; Tiboni et al., 2022). The IEEE reports states that approximately 42% of all induction motor failures are due to bearing faults (Albrecht et al., 1986). The downstream consequences are very severe when the bearing fails without warning.

Bearing degradation follows a progressive trajectory as healthy bearings exhibit broadband vibration caused by normal motor rotor dynamics, lightly damaged bearings introduce a periodic impulse at defect characteristic frequencies which were initially buried in background noise. Heavily damaged bearings produce a strong clearly visible impulse signature.

The industry's practical answer to bearing failure risk has been deployment of sensors on motor to do condition monitoring. Devices such as the ABB smart sensor attached to motor capture main multi parameter data streams. In practice, however most deployed analytics pipelines reduce these vibration signals to handful of scalar statistics flagging alerts only when an RMS or when kurtosis exceeds a fixed ceiling. Such threshold-

based rules can confirm that vibration is elevated but cannot adapt to variable operating speeds and offer no explanation for their decisions (Randall & Antoni, 2011).

However, the existing research has three important key gaps that prevent monitoring systems from reaching their full potential. First, virtually all published fault diagnosis is validated on single session benchmark datasets with a narrow speed range: The most popular and benchmark dataset from Case Western Reserve University (CWRU) dataset spans only 4 speeds across the 67 RPM speed range. The Paderborn dataset applies a fixed load protocol and the MFPT bearing dataset covers 4 operating conditions (Loparo, 2012). In real factory environment motors do not run at one speed. The motors speed up and slow down and useful AI models must be able to diagnose a fault at 1500 RPM even if it was only trained on data from 1000 RPM which is called generalization. Uniquely, this thesis has a 4-year longitudinal dataset collected from the same physical ABB test bed spanning from 2022 to 2025. Second, based on the literature review of this thesis as no prior study has evaluated bearing fault diagnosis models on multi-year longitudinal data from a real industrial testbed leaving unanswered the question of how model performance evolves as cumulative bearing wear, temperature drift and environmental variation over time. Third, the relative diagnostic contribution of the three accelerometer axes radial X, axial Y and radial Z as a function of rotational speed has not been systematically characterized using attribution methods making it impossible to guide feature selection decisions from empirical evidence.

1.2.1 Data driven approaches to Bearing faults diagnosis:

Machine learning and deep learning approaches have shown considerable promise in bearing faults diagnosis precisely because they can extract complex nonlinear patterns directly from vibration signals which does not require manual feature engineering. The authors Randall and Antoni state that Conventional neural networks operating on raw waveforms or frequency spectra, random forest classifiers trained on engineering FFT features and ensemble methods combining representations have strong classification accuracy in controlled laboratory environment. These data driven models

offer a clear advance over scalar threshold rules capturing damage related temporal patterns and spectral structures that amplitude statistics cannot encode.

Despite these advance algorithms several limitations constrain the real-world reliability of purely data driven models. As noted in section 1.2, most benchmark datasets cover only a narrow range of operating speeds meaning published accuracy figures do not reflect how a model will perform at speeds it was never trained on. Furthermore, scalar feature representations such as kurtosis which measures impulses or the heaviness of the tails of a probability distribution which is widely reported in the literature have been shown empirically to fail at consistent faulty severity discrimination across a broad speed range and their inclusion in classification pipelines without validation under such condition's risks overstating diagnostic capability.

A further limitation is the black-box nature of deep learning. Even when a CNN achieves high average accuracy it provides no inherent explanation of which signal features or frequency bands resulted in a particular health state prediction. In safety and reliable maintenance settings this lack of interpretability erodes operator and customer trust and makes it difficult to audit model behavior when operating conditions shift. Addressing interpretability is therefore not optional enhancement but a prerequisite for responsible deployment of data driven bearing diagnostics in industrial environments.

1.2.2 Explainability and longitudinal validation:

This thesis addresses this problem by implementing explainable artificial intelligence which offers a principled route to overcoming the interpretability barrier in bearing fault diagnosis. Attribution methods such as DeepSHAP compute feature level importance feature level importance scores that reveal which parts of the input specific frequency bands, time domain segments or individual accelerometer axes are most responsible for a model's classification decision (Lundberg & Lee, 2017).Applied to CNN

architecture trained on vibration data these tools can translate an deep learning model into an auditable diagnostic instrument.

This thesis has the availability of longitude vibration dataset spanning multiple recording years on the same testbed opens a further methodological opportunity that single session benchmark studies cannot address. Training and evaluating models on data collected from 2022 to 2025 across 57 unique rotational speed points allow assessment of how classifier performance evolves as environmental variation, temperature drift and cumulative bearing wear over time. Leave One Speed Out cross validation which withholds an entire speed fold from training provides a substantially more realistic measure of generalization than conventional random train test splits directly simulating the deployment scenario in which a model encounters speeds it has never seen during training.

The connection between this research and ABB Digital framework is direct. The ABB Digital Platform organizes its predictive maintenance offering around the 3 stage Detect-Predict –Recommend pipeline (ABB Ability Digital Powertrain Enabling Hardware,2026). The framework's structure maps directly onto the machine learning pipeline developed in this research work is unsupervised anomaly detection fulfils the Detect stage by flagging the deviations from bearing healthy state behavior. Supervised severity classification fulfils the Predict stage by characterizing the degree of damage. A confidence threshold reaction option returning an Unknow or Inspection decision when the inference model falls beyond 0.80 which fulfils Recommend state by communicating uncertainty rather than forcing a potentially erroneous diagnosis onto maintenance engineers.

Together explainability methods, longitudinal data validation and integration within an industrially grounded framework represent a coherent opportunity to move bearing health monitoring beyond threshold rules. This shift points towards genuinely trustworthy and deployment ready intelligence models.

1.3 Research Questions

The research is structured around four questions that together address the identified gaps in the bearing fault diagnosis literature:

- RQ1: Which feature representation gives the most reliable and accurate bearing damage classification across a full 0-1500 RPM operational speed range?

- RQ2: Can a deep learning model trained on raw vibrational signals generalize to motor speeds it has never seen during training?

- RQ3: Do the three sensor axes contribute equally to bearing fault detection and does their contribution change at different motor speed?

- RQ4: Can a faulty detection system train only on healthy bearing data reliably identify damage across all motor speeds?

1.4 Research Objectives

Objective 1: Compare three different ways of representing vibration data frequency spectra, raw waveforms and scalar statistics to find out which one gives the most

reliable results for classifying bearing damage severity across the full motor speed range.

Objective 2: Build and test a deep learning model that works directly on raw vibration signals from all 3 sensor axes and compare its performance against simpler frequency-based classifier to see which one handles unseen motor speeds better.

Objective 3: Use Explainable AI algorithms to look inside the trained models and understand which of the sensor axes radial X, axial Y or radial Z is most important for detecting bearing damage and whether these changes depending on how fast the motor is spinning.

Objective 4: Build three different anomaly detection models that are trained only on healthy bearing data and examine how well each one detects bearing faults across the full motor speed.

1.5 Scope and Limitations:

The scope is specifically confined to developing and evaluating machine learning and deep learning models for bearing health classification. Currently, temperature and acoustic monitoring methods are not considered in this research study. The focus of this work is on understanding how much good information can be extracted from sensor data alone.

The following are the limitations of the thesis:

- Single fault mechanism: All experimental data comes from one specific fault induction method that is metallic dust contamination. This gives a clean and controlled environment to label the data.
- Dataset asymmetries: The longitudinal dataset contains temporal and speed range asymmetries between recording sessions. Due to the nature

of long-term experimental data collection which is not possible to have a perfect balance data for every recording across years.

- Low-speed windowing limitation: There is also a physical limitation with the 300-sample windowing strategy at a 1660 Hz sampling rate. No model can extract meaningful information rotational features from a window that does not contain full rotation. This affects model reliability at low operating speeds.
- Data confidentiality: Finally, in accordance with the data governance framework agreed with ABB Oy. This study is limited to presenting class-level statistical aggregates instead of Dataset.

1.6 Thesis Structure:

The Thesis is organized into 6 chapters. Chapter 1 covers the industrial background, the research problem and the research questions and objectives. Chapter 2 deals with the relevant literature including bearing fault signal analysis, machine learning and deep learning approaches for fault classification, the variable-speed generalization problem, explainable AI methods and unsupervised anomaly detection. Chapter 3 explains the methodology, the experimental testbed, the dataset, the preprocessing pipeline, feature extraction, model architecture and the evaluation protocol. Chapter 4 presents the results for all the supervised classification and anomaly detection models. Chapter 5 discusses results of the research questions and situates them within the broader literature. Finally, Chapter 6 discusses the conclusion and suggests directions for future work.

2 Literature review

This chapter reviews the existing literature that forms the foundation of this thesis as follows:

1 Bearing fault mechanisms and classical signal analysis methods.

2 Machine learning and deep learning approaches for bearing fault classification.

3 Variable-speed generalization challenges.

4 Explainable AI for bearing fault diagnosis.5 unsupervised anomaly detection for motor health monitoring.

6 identified research gaps that this thesis directly addresses.

2.1 Bearing Fault and Signal Analysis:

Bearings are one of the most frequently failing components in any rotating machine. When a bearing starts to develop a fault, it does not just break suddenly. It leaves behind patterns in the vibration signal first. The way this works is that when a small defect on the surface of a rolling element or raceway passes through the loaded zone, it will create a short sharp impact. That impact repeats itself at a specific rate depending on two main things. First, the geometry of the bearing itself and second according to authors McFadden & Smith, 1984, How fast the motor shaft is spinning.

That is why, if the bearing geometry is already known we can calculate exactly which frequencies to look for. The most important characteristic defect frequencies are the following:

1. Ball Pass Frequency Outer Race (BPFO). It tells us if the fault is on the outer raceway
2. Ball Pass Frequency Inner Race (BPFI). It tells us if the fault is on the inner raceway
3. Fundamental Train Frequency (FTF). It is related to the cage rotation
4. Ball Spin Frequency (BSF). It is related to the rolling element itself

Each one of them tells us something different about where the fault might come from. The idea is very simple that just track these frequencies in the vibration spectrum and see which ones show stronger than they should.

Back in 1984, the authors McFadden and Smith were among the first to properly describe this kind of frequency domain thinking. Later, the research authors Randall and Antoni (2011) expanded on it and their work became pretty much the go to reference in this area. This foundation is important because:

- Traditional expert-based rule systems rely on it to set fault thresholds
- Modern machine learning approaches use it to label and interpret features

Both depend on understanding what normal fault looks like first before anything unusual pattern can be spotted. There are some classical signal processing methods that have been widely used for bearing fault detection.
Two of the most important ones are:

1. Spectral Kurtosis: It basically measures how impulse a signal is in the frequency domain. The idea is that a healthy bearing produces a relatively smooth signal and while a faulty one creates sharp impacts that show up clearly.
2. Envelope Analysis: This works differently. Instead of looking at the raw signal, it demodulates the high-frequency part of the signal. That way, it can recover the low-frequency repetition pattern that tells us something is wrong.

Both methods have been around for a long time and are considered reliable starting points for bearing diagnosis. Randall and Antoni (2011) showed that kurtosis is one of the best indicators for catching early-stage faults. That is why it became so popular.

However, there is a problem with kurtosis that is important to mention:

- It works well when the fault is still small and localized just like a single pit on the raceway.
- But as the damage spreads and becomes more distributed across the surface. The signal stops being impulse and becomes more broadband.
- When that happens, the kurtosis starts dropping. Sometimes significantly making it much harder to detect the fault.

That is exactly what it was also noticed in this thesis results. Under LOSO evaluation across 57 folds, the kurtosis did not achieve reliable results. This directly confirms what Randall and Antoni (2011) warned that a single scalar impulse statistic simply cannot discriminate between different levels of contamination type fault severity, especially when the shaft speed is also changing. (Randall & Antoni, 2011).

2.2 Machine Learning for Bearing Fault Classification:

In 2012 the Case Western Reserve University (CWRU) bearing dataset was made publicly available by the author Loparo (2012). Because of this public dataset the research really started to move fast in terms of machine learning applications for bearing fault diagnosis. Researchers started applying different methods to this benchmark dataset and the results were promising. Some of the early approaches included:

- Support Vector Machines (SVM): This model uses hand crafted time-domain and frequency-domain features, and these models managed to achieve classification accuracy typically in the range from 88-93% on the CWRU benchmark which is depending on the feature extraction process employed dataset (Lei et al., 2020).

- Lei et al. (2009) proposed a fault classification method that combined wavelet packet transform, empirical mode decomposition and a radial basis function network. The method was applied to mechanical fault diagnosis. The results demonstrated that feature-based approaches could reliably distinguish between different fault states.
- Random Forest: This model brought further improvements because it naturally handles high-dimensional feature spaces without needing strict assumptions about data distribution (Breiman, 2001).

However, both classical approaches share the same fundamental problem. The features must be defined manually by someone which is expert in this field and who already understands the machine. That means if the machine configuration changes and the features might not capture the right fault signatures anymore. That is why researchers started looking for something better.

Deep learning offered a solution to this. Instead of relying on hand crafted features, deep learning models learn useful representations directly from the raw data. The authors Zhang et al. (2017) states that a 1D convolutional neural network operating directly on raw vibration waveforms could match or even beat traditional machine learning models like SVM with engineered features all without any manual feature extraction. Around the same time, Ince et al. (2016) applied a similar 1D-CNN approach to motor current signals. The authors stated that models could learn temporal filters that captured bearing fault signature.

More recently, researchers have moved towards multi scale architecture. The idea behind these is straightforward:

- Instead of looking at the signal at only one resolution if processes the data at multiple temporal scales.
- This gives the model a more complete picture of what is happening in the signal.

This is one of the main motivations behind the multi-scale residual architecture that is developed later in this thesis.

2.3 The Variable Speed Generalization:

Generalization to variable speed is a major limitation of models trained and tested at single speed, yet this issue receives limited attention in the bearing fault diagnosis literature. Most research studies only evaluate their models at a single speed or within a very narrow speed range. A good example is the CWRU dataset, the most widely used benchmark in this field, which is collected at only four discrete speeds. A model trained and tested on this dataset is never required to classify a bearing fault at a speed it has not previously seen a requirement that routinely arises in in real-world applications, especially when the motor is connected to a variable-speed drive. The physical reason behind this problem is that a model that has learned to associate a fixed frequency pattern with a fault class will simply fail at any speed that was not in its training data.

In this thesis, speed conditioning was chosen for all classifiers. This makes it easier for the model to make sense of what it is seeing. More recently, the research authors Zhong et al. (2023) and Saha et al. (2024) proposed transfer learning methods to tackle the cross-speed generalization problem. These are interesting approaches, but they still require some overlap between the training and test speed distributions. That means they are not truly testing the model at a completely unseen speed.

That is exactly where the Leave-One-Speed-Out (LOSO) protocol used in this thesis is different and stricter:

- Each fold completely withholds one speed point that was never seen during training phase.
- The model is then evaluated at a genuinely novel operating condition.
- This is repeated across 57 speed folds in total.

Speed generalization evaluation reported in the bearing fault diagnosis literature.

2.4 Explainable AI for Bearing Fault Diagnosis:

Nowadays, deep learning models are being used in industrial applications where the stakes are high That is why interpretability has become a serious concern. The problem is simple that a model that achieves high accuracy is good, but it cannot explain why it made a certain decision is very hard to trust. Maintenance engineers cannot just accept a black box answer without understanding what is behind it. That is why there has been growing interest in explainability methods for bearing fault diagnosis.
One of the most well-known frameworks for this is SHAP. It was introduced by the author Lundberg and Lee (2017) and the idea behind it comes from cooperative game theory. The way it works is as follows:

- Each input feature gets assigned an attribution value.
- That value tells us how much that specific feature contributed to the model's prediction
- Features with higher attribution had more influence on the final classification decision.

Several studies have applied SHAP-based attribution to bearing fault diagnosis. The authors Sanakkayala et al. (2022) applied explainability methods to bearing prognosis using deep learning. These studies typically apply attribution to a single architecture at a fixed operating speed. DeepSHAP is a specialized variant of this framework designed to leverage the internal structure to compute attribution more efficiently. Based on the literature review of this thesis no prior study has performed a cross-architecture DeepSHAP comparison across different speed points. This thesis used DeepSHAP which is applied to two CNN models that is a raw waveform CNN and an FFT-based CNN at 10 representation speed folds enabling a cross-architecture comparison.

2.5 Unsupervised Anomaly Detection for Motor Health Monitoring:

One of the biggest practical problems with most fault detection approaches is that they require labelled fault data for training. In real industrial settings this is often not available. Motors are usually expected to run in a healthy condition for long periods which means fault examples are rare or maybe the data is not collected. That is why unsupervised anomaly detection has become an attractive alternative and it only needs normal-condition data to train and then flags anything that looks different as a potential fault.

The most important research paper approach for doing this is the autoencoder. Back in 2014, The authors Sakurada and Yairi (2014) demonstrated that an autoencoder trained only on normal data could be used for anomaly detection.

The idea behind it is as follows:

- The autoencoder learns to reconstruct normal patterns well.
- When a faulty or unusual sample is passed through it the autoencoders model reconstruction error is much higher.
- That reconstruction error becomes the anomaly score.

This approach has since been validated on motor vibration data by several researchers. The authors Principi et al. (2019) applied it to motor signals. Another approach that works quite differently is Isolation Forest which was introduced by the author Liu et al. (2008). Instead of learning to reconstruct data it works by randomly partitioning the feature space.

The core logic of this model is as follows:

- Normal data points are densely packed together and take many partitions to isolate.
- Anomalies sit in sparse regions and get isolated much faster.

- Points that are isolated quickly through shorter paths and they will receive higher anomaly scores.

The advantage of this approach is that it does not require any distance or density calculation. Based on the literature review of this thesis, no prior study has evaluated different autoencoder approaches on the same variable-speed dataset, making direct performance comparison impossible from existing literature. So, in this thesis, 3 anomaly detectors are evaluated on the same dataset and speed folds: Isolation Forest, FT Autoencoder and CNN Autoencoder. Based on the literature reviewed, this is the first direct comparison of these types of anomaly detectors for bearing condition monitoring across a wide variable-speed range. (Liu et al., 2008 and Sakurada & Yairi, 2014).

2.6 Identified Research Gaps:

Based on the reviewed literature, four major gaps were identified. They represent areas where existing research has not sufficiently addressed real-world operating conditions, particularly variable-speed conditions.

The four gaps are as follows:

Gap 1. LOSO evaluation at scale: Leave-One-Speed-Out evaluation across a wide speed range is the one of the important protocols that actually proves whether a model can generalize to unseen speeds. However, based on the literature reviewed, no prior fault diagnosis study has ever been applied at a scale of 57 unique speed. Most existing work either tests at a single speed or uses a very narrow range, which does not reflect real variable speed operating conditions.

Gap 2. Longitudinal validation: Based on the literature reviewed, no prior study validated spanning multiple recording years from the same physical testbed. This is important because real machines change over time through environmental variations, temperature

drift and cumulative wear and tear. A model validated on one session cannot account for these effects.

Gap 3. DeepSHAP comparison: Several studies have applied SHAP based attribution methods to individual bearing fault diagnosis. However, based on the literature reviewed prior study applied DeepSHAP and a direct comparison across architecturally different CNN models at multiple speed points. This is important because different architectures have different inductive biases and it is not yet clear how that affects what features they pay attention to.

Gap 4. Anomaly detector comparison:
Finally, In the bearing fault diagnosis literature reviewed, reconstruction based and isolation-based anomaly detectors have never been evaluated on the same variable speed dataset. Studies tend to pick one approach and test it in isolation as it makes it very hard to draw meaningful conclusions about which detector type is best suited to variable speed operating conditions.

This thesis was designed to address all these four gaps.

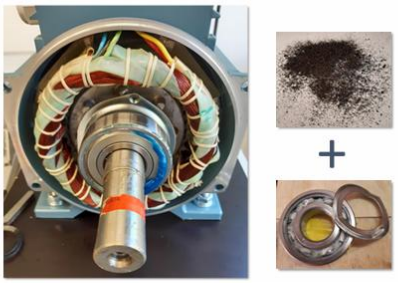
3 Methodology

This section focuses on experiment testbed, data collection, preprocessing pipeline applied to raw vibrational signals, the feature representation, the supervised classifications developed, the interpretability methods and the unsupervised anomaly detection models. The process of evaluation is also explained in detail.

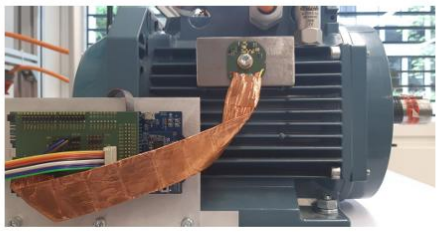
3.1 Data collection

The data used in this study was collected from the ABB Research center Switzerland a health motor monitoring research project described by Goyal et al. (2022).The testbed has three ABB M3AA asynchronous 3-phase induction motors rated at 2.2 kw, 400V each representing a distinct bearing health state that was established and maintained throughout the four years of data collection. The motor is powered by a standard 3-phase AC supply, and its speed is regulated by an ABB ACS355 Variable Speed Drive (VSD) which allows precise control of the motor operating conditions during data collection. Bearing faults were physically induced by introducing controlled amount of metallic dust into motor load side bearing which produces three distinct health states.

- Motor 1 operates with an undamaged drive end bearing designated as a Healthy class.
- Motor 2 operates with a lightly contaminated bearing produced by adding 0.2g of metallic dust particles into the front bearing before sealing which was designated as Lightly Damaged class.
- Motor 3 operates with a highly contaminated bearing produced by the same method but with 1g of dust contamination and designated as Heavily Damaged class. This controlled protocol ensures that class labels are deterministic and free from label noise.



(a) Controlled bearing damages by adding measured quantities of hardened metallic dust into the front motor bearing



(b) Accelerometer mounted on the side of the motor and interfaced to a Nordic nRF5340 development board

Figure 3. Test bed of Motor fault prediction (Goyal et al. (2022), p.3).

Figure 3 is Experimental testbed showing the 3 ABB M3AA motors with sensor placement at the motor terminal block. Vibration data were recorded using a sensor mounted on the motor terminal block. The X-axis and Z-axis measure radial vibration in two orthogonal directions which were perpendicular to the shaft. The Y-axis measures axial vibration parallel to the shaft. The sensor placement at the motor terminal block was selected to maximize signal to noise ratio for bearing fault signatures consistent with established in vibration-based bearing health monitoring (Ayankoso et al., 2024).

3.1.1 Longitudinal Data Structure

Data was collected and saved across four years spanning from 2022 to 2025. The consolidated motor vibration dataset is used in this research work consists of six primary columns which are floating point variables X, Y and Z representing physical sensor readings and three integer-based columns for motor type, operational speed, timestamp and the collection year. The 2025 data collection was conducted through a personal visit to the ABB Research center in Switzerland. This visit allowed access to the experimental testbed. The data is distributed across three distinctive motor types based on the health of bearing Motor1, Motor2 and Motor 3 which totaled 21 million rows with operational speeds varying from 0 to 1500 and 57 unique speeds. The four-year temporal structure of this

dataset is unique in the bearing fault diagnosis literature, and all existing public benchmarks are single session collections.

The ten representative speeds reported in this study are drawn from all four recording years. The year 2022 contributes 50, 500, 800, and 1500 RPM. The year 2023 contributes 535, 994, and 1500 RPM. The year 2024 contributes 600, 900, 1200, and 1500 RPM and the year 2025 contributes 50 and 100 RPM. The speed 1500 RPM appears across three consecutive years making it the primary reference point for longitudinal stability comparison. In contrast, the irregular-interval speeds 535 and 994 RPM appear in a single year only. These speeds serve as pure generalization stress-test points, with no temporal averaging available to support the model. This makes the models' evaluation more realistic and more representative of the challenges that would be encountered in a real industrial setting.

3.1.2 Class Balance and Label Assignment:

Labels were assigned based on the known physical state of each motor at the time of data collection. The healthy motor is labelled as 0, Lightly damaged motor as 1 and heavily damaged motor as 2. This deterministic labelling eliminates label noise and across all recorded years each motor contributes approximately 33 percent of the total row count with each recording session providing three class balance at approximately 1:1:1.

3.1.3 Data Governance and Ethics:

The vibration data were collected within the ABB research infrastructure which operates under the data residency and privacy governance framework. In accordance with the ABB data governance agreement and this thesis does not reproduce raw vibrational signal data in any form and does not disclose any operational data and presents all results as class level statistical aggregates. All model performance metrics reported in

the following chapters are computed over class level predictions and do not expose individual samples or time stamps of the raw data. The author of this thesis has agreed to confidentiality terms of ABB Oy.

3.2 Data Preprocessing

The vibration signals were subjected to a preprocessing pipeline before any feature extraction or model training must be performed. This pipeline is designed to remove DC offset, standardize the representation of signals collected across four recording years and 57 rotational speeds, and reduce spectral leakage during frequency-domain analysis. Section 3.2.1 provides an overview of the pipeline structure Section 3.2.2 through 3.2.7 describes each processing step in detail including time-domain segmentation, Hanning windowing. DC offset removal, FFT feature extraction, global Z-score normalization and stratified class alignment sampling.

3.2.1 Preprocessing Pipeline Overview

A preprocessing pipeline was applied to all raw vibration data before feature extraction or model training. The same pipeline was applied identically across all four recording years and all 57 LOSO folds. The preprocessing consists of 6 sequential stages applied to vibration data:

1. Time-domain segmentation
2. Hanning Windowing
3. DC offset removal
4. Feature extraction
5. Global Z-score normalization
6. Stratified class aligned sampling

Each stage is described in detail in Sections 3.2.2 to 3.2.7.

3.2.2 Time-Domain Segmentation:

Raw vibration data was segmented into non overlapping windows of 300 samples corresponding to 180.7 milliseconds at 1660 Hz. This window length balances two competing requirements that is sufficient duration to capture multiple fault impulse cycles at mid-to-high speeds while keeping the FFT frequency resolution and computational cost tractable. For example, at 1500 RPM a 300-sample window captures 4.52 shaft rotations approximately and at 100 RPM it captures 0.30 rotations. The number of rotations captured per window is the primary determinant of classification reliability at low speeds.

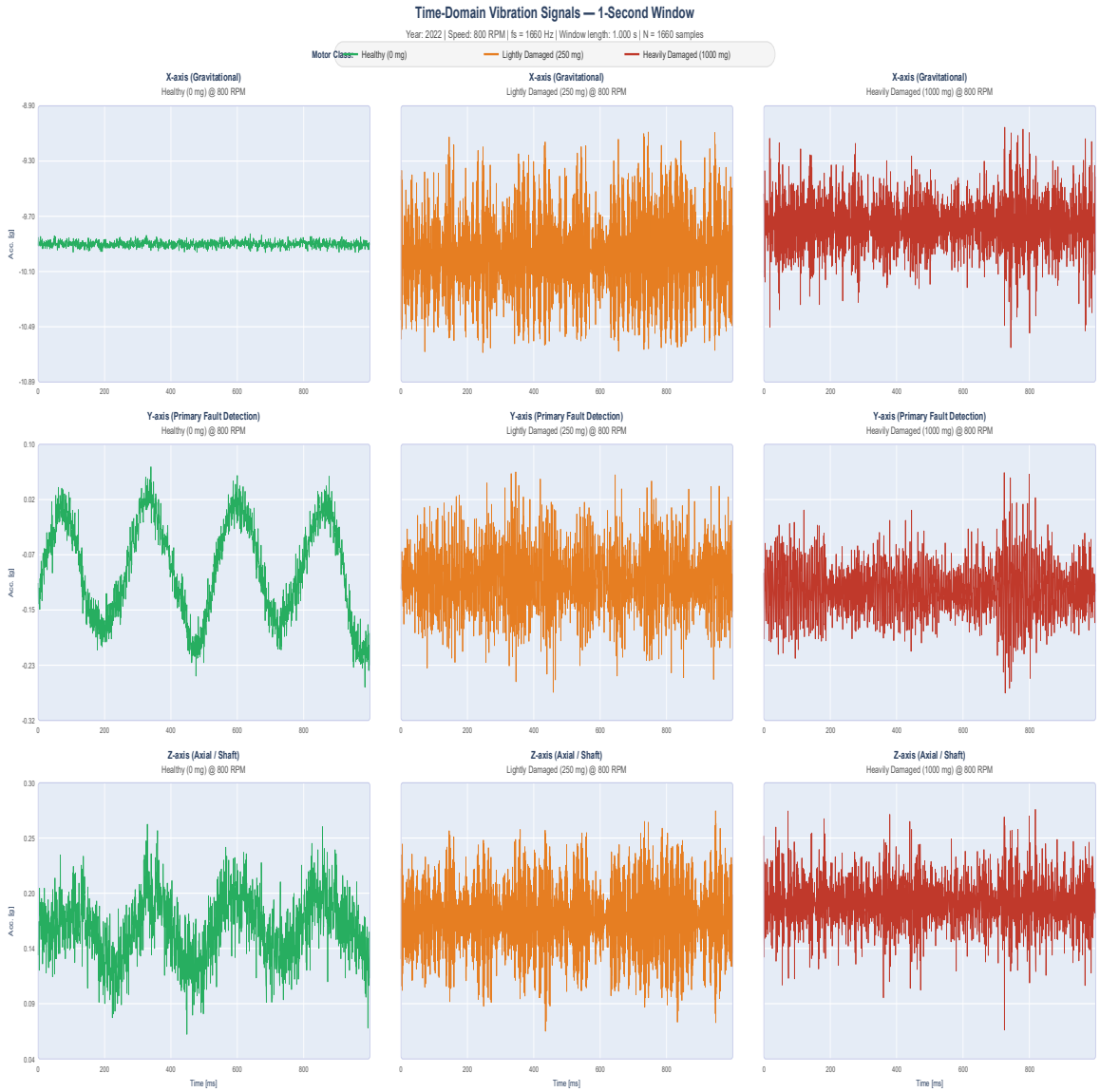


Figure 4. Time domain vibration signals at speed 800.

Figure 4 illustrates that class differences green color is healthy motor, orange color is lightly damaged and red color is strongly damaged motor. The Y axis where healthy motor produces a clean low amplitude sinusoidal pattern in contrast to high frequency noise of both lightly and severely damaged motors. The difference is visually apparent.

3.2.3 DC Offset Removal:

The arithmetic-mean was subtracted from each window independently before applying any spectral transformation. This DC offset removal eliminates the static gravity component which appears as a non-zero mean in the acceleration signal which is due to the orientation of the sensor relative to the gravitational field as well as to eliminate any sensor bias.

3.2.4 Hanning Window Application

Prior to FFT computation each window was multiplied element wise by a Hanning window function of the same length. The Hanning window is a smooth taper that reduces spectral leakage.

3.2.5 Frequency-Domain Feature Extraction

The real valued FFT was applied to each windowed segment per axis producing 151 complex frequency bins spanning 0 Hz to 830 Hz. The DC bin at index zero was excluded from feature vector and bins 1 through 150 were retained providing a frequency resolution of 5.53 Hz per bin and a coverage range of 5.53 Hz to 829.7 Hz. FFT magnitudes were scaled by 2/N, where N is the window length, to convert from raw transform units to physical acceleration in units of enabling physically interpretable amplitude values.

The FFT magnitude spectra from all 3 axes were concatenated into a single 450-dimensional feature vector which means 150*3 axes. Rotational speed was appended as a scalar conditioning variable producing a final 451-dimensional input vector for the classifiers. This speed conditioning variable allows the classifier to contextualize the spectral representation relative to the know operating speed. It is enabling to distinguish spectral peaks that are speed dependent mechanical resonances from those that scale the bearing defect frequencies.

3.2.6 Global Z-score Normalization

Global Z-score normalization was applied to standardize the FFT feature vectors prior to model training. The normalization parameters mean and standard deviation were computed exclusively from the training windows of each LOSO fold and the same parameters were then applied identically to the corresponding validation and test windows. This fold of local normalization prevents the amplitude statistics of held out speed points from contaminating the normalization parameters which would constitute a form of data leakage.

3.2.7 Pre-Training Class Alignment via Stratified Sampling:

To ensure that the model was not dominated by any year's data collection a stratified sampling strategy was implemented during training data preparation. Equal numbers of windows were sampled from each recording year, and the natural class balance of approximately 33 per cent per class was preserved within each year stratum. For binary class normalization task an initial 1:2 Healthy to light plus heavily unhealthy ration was evaluated. This was found to be systematically bias the binary decision boundary to majority class.

To rectify this strict 1:1 binary subsampling protocol was implemented. This protocol draws equal numbers of samples for the Healthy that is class 0 and the merged unhealthy that is class 1 and class 2 targets. This process ensures that the loss function is equally sensitive to both healthy and damaged states. All binary classification results reported in Chapter 4 use the strict 1:1 subsampling protocol.

3.3 Feature Analysis:

Following preprocessing, three distinct feature representations were extracted from the vibration signals and exploration data analysis is done prior to model training. The purpose of this analysis is to know the discriminative structure of each representation across the 3 bearing health states and the full operational speed range.
The three representations are:
(1) Root Mean Square (RMS)
(2) Pearson kurtosis and
(3) the FFT magnitude spectrum.

T-SNE visualization of the FFT feature space is presented in Section 3.3.4 to examine the cluster separability prior to classification. The feature analysis results reported in this section motivate the supervised model design choices described in Section 3.4.

3.3.1 RMS Analysis:

RMS features can extract from a vibration signal is the Root Mean Square (RMS) value which gives a measure of the signal's average energy over a given window. the RMS is computed as:

$$\boldsymbol{RMS} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{x}[\boldsymbol{n}]^{\mathbf{2}}}$$

RMS was computed separately for each of the three accelerometer axes per window giving three scalar values per sample: RMS_X, RMS_Y, and RMS_Z. The basic idea behind using RMS for bearing health assessment is that intuitively a damaged bearing tends to generate more vibration energy than the RMS value to increase with worsening damage severity. This has been widely reported in the bearing fault diagnosis literature [Althubaiti et al., 2022 and Tiboni et al., 2022].

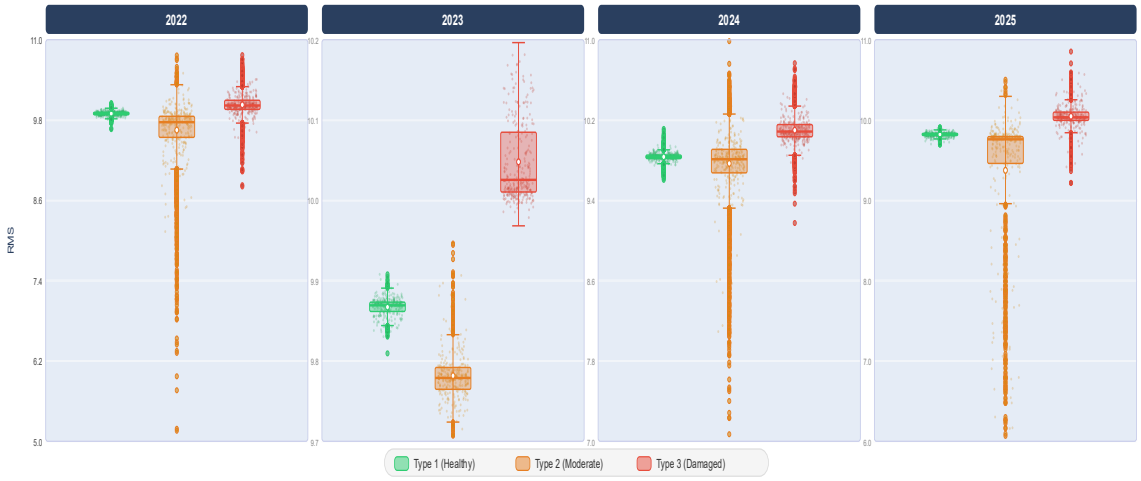


Figure 5. RMS Distribution from 2022 to 2025.

Figure 5 represents the Box plot distribution of RMS for Healthy, Lightly Damaged and Heavily Damaged bearing classes across 4 years. In 2022, 2024 and 2025 the Heavily Damaged motor consistently produces the highest median RMS which aligns with the expected monotonic relationship between damage severity and vibration energy. However, the Lightly Damaged motor exhibits a wider interquartile range in all 4 years with considerable distribution overlaps with both the Healthy and Heavily Damaged classes. The 2023 recording presents a more extreme departure from the expected ordering as the Lightly Damaged motor produces a lower median RMS than the Healthy motor, a cross-year inversion that directly contradicts the monotonic damage energy assumption.

The inversion can be explained in the specific damage mechanism of metallic dust contamination. In the early stage of contamination at 250 mg individual metallic particles act as sharp abrasives that excite structural resonances in the bearing housing. This pro-

duces periodic high-amplitude impulses that elevate radial RMS beyond what the damage mass alone would predict. As contamination progresses to the heavy damage state at 1000 mg the bearing surfaces become uniformly degraded and the sharp particle impacts transition to sustained broadband friction which produces a lower peak amplitude but a higher and flatter spectral noise floor. This mechanism explains why the Lightly Damaged motor does not consistently exist between the Healthy and Heavily Damaged motors in the RMS distribution. It exposes the fundamental limitation of RMS as a three-class severity discriminator. As a result, RMS cannot distinguish between the spectral morphologies that characterize the three damage states in this dataset as a shortcoming that directly motivates the evaluation of frequency domain features.

3.3.2 Kurtosis Analysis:

Pearson kurtosis, the fourth standardized moment of the amplitude distribution which is equal to 3.0 for a Gaussian signal was computed per axis per window. On the Y-axial and Z-radial axes the ordering Healthy below Lightly Damaged below Heavily Damaged held consistently across recording years confirming that metallic contamination introduces measurable non-Gaussian impulses. However, the X-radial axis showed inconsistent ordering across years, and the 2024 recording revealed an anomalous Healthy-class X-axis kurtosis of 11.72 is exceeding both damaged classes, which is most likely a structural resonance excited during that campaign. This cross-year inconsistency, combined with the fundamental limitation of kurtosis as a severity discriminator once damage transitions from impulse to distributed, foreshadows the model evaluation results presented in Chapter 6.

$$K = \frac{\frac{1}{N}\sum_{n=1}^{N}(x[n]-\mu)^4}{\left(\frac{1}{N}\sum_{n=1}^{N}(x[n]-\mu)^2\right)^2}$$

A Gaussian signal yields a kurtosis of exactly 3.0 values above 3.0 indicate impulse non-Gaussian behavior which is theoretically expected when a bearing develops a localized defect that generates periodic impact impulses (Antoni, 2006). Kurtosis was computed per axis, that is K_X, K_Y and K_Z.

Figure 6. Pearson Kurtosis by health class and recording year 2022-2025.

Figure 6 shows the kurtosis of the vibration magnitude vector, computed as the square root of $(x^2+y^2+z^2)$ across all three health classes and four recording years. The dashed reference line at 3.0 marks the Gaussian baseline. Results for kurtosis are inconsistent, with no clear trend over the years. The Heavily damaged motor has a notably higher kurtosis in three of the four years, while for the lightly damaged motor the kurtosis is even slightly below the baseline for three of the four years. In 2025, the ordering inverts as the Healthy motor produces higher kurtosis than the Lightly damaged motor.

This inconsistency has a straightforward explanation. Kurtosis is theoretically optimal for detecting early stage localized defects that produce sharp isolated impulses (Randall & Antoni, 2011). Once damage becomes distributed as is the case

with 1000 mg of metallic dust contamination, the signal transitions from impulse to broadband friction and kurtosis decreases rather than increases. Combined with the cross-year instability visible in Figure 6, this confirms that kurtosis alone is not a reliable feature for three class severity discrimination across a longitudinal dataset.

3.3.3 FFT Feature Analysis:

Unlike RMS and kurtosis, which reduce the vibration signal to a single scalar value the FFT decomposes the entire signal into its constituent frequency components producing a spectral magnitude vector that preserves the full frequency domain structure of the bearing vibration. For a discrete signal x[n] of length N, The FFT is defined as:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi kn/N}$$

where k= 0,1, ...N-1

Figure 7. Power spectral density curves by health class and recording year (2022 to 2025).

Figure 7 represents the Power Spectral Density curves for all three health classes across the four recording years 2022 to 2025. A consistent pattern is visible across 2022, 2024,

and 2025 as the Healthy Motor Green Color maintains a substantially lower broadband power floor sitting approximately 20 dB below the damaged classes across the full 0 to 800 Hz range. The Lightly Damaged motor orange color and Heavily Damaged motor which is indicated as red color both show elevated power levels, but their profiles differ in shape. The Lightly Damaged motor tends to show sharper localized peaks at characteristic resonant frequencies while the Heavily Damaged motor shows broader spectral elevation with higher broadband noise. The morphological difference that is completely invisible to both RMS and kurtosis but directly learnable from the FFT magnitude vector.

The 2023 recording is again an exception where all three classes converge more closely and consistent with the cross-year variability observed in kurtosis and RMS sections. However, even in 2023 the Healthy motor maintains a visibly lower power floor than both damaged classes confirming that frequency-domain separation between healthy and damaged bearings is more stable across years than the scalar features examined previously.

These spectral differences are that is a near zero noise floor for healthy bearings, sharp resonant peaks for lightly damaged bearings and broad elevated noise for heavily damaged bearings form the physical basis for the superior classification performance of FFT-based models.

3.3.4 t-SNE Exploratory Analysis of FFT Features:

Before training any classifier a t-SNE, exploration analysis was performed on the FFT feature space to visually assess whether the 450-dimensional feature vectors carry sufficient structure to separate the three health classes and to understand how operating speed organizes the feature space. t-SNE is a nonlinear dimensionality reduction technique that projects high dimensional data into two dimensions while preserving local neighborhood structure and making it well suited for visualizing cluster separation in complex feature spaces (van der Maaten and Hinton, 2008).

The analysis was performed on FFT feature vectors extracted from sampled windows across all four recording years, three health classes, and four representative speed points per year. The 450-dimensional FFT vectors were first standardized and reduced to 50 dimensions using PCA, retaining the dominant variance structure, before t-SNE was applied to produce the final two-dimensional embedding.

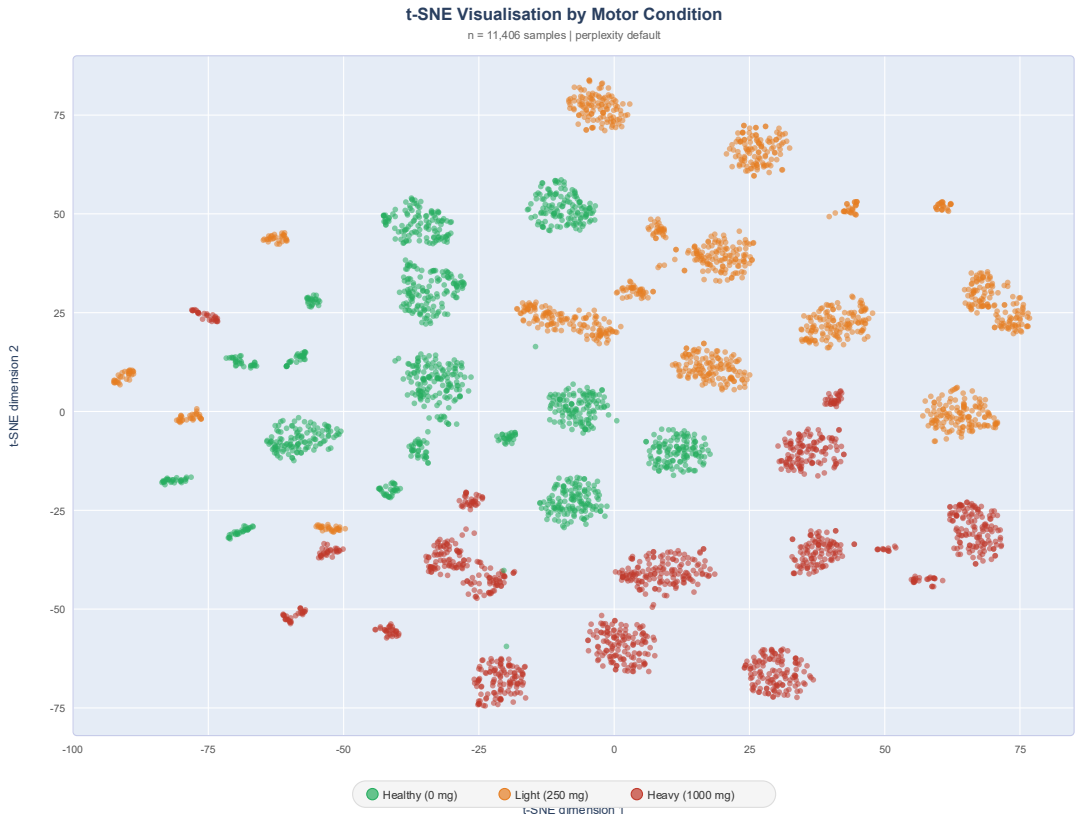


Figure 8.t-SNE projection of FFT features colored by damage class.

Figure 8 presents t-SNE plot of the three health classes Healthy (green), Lightly Damaged (orange) and Heavily Damaged (red) form a largely separate cluster territories across the embedding space. This spatial separation provides direct visual evidence that the 450-dimensional FFT vector. It encodes damage related spectral differences that are learnable by a downstream classifier.

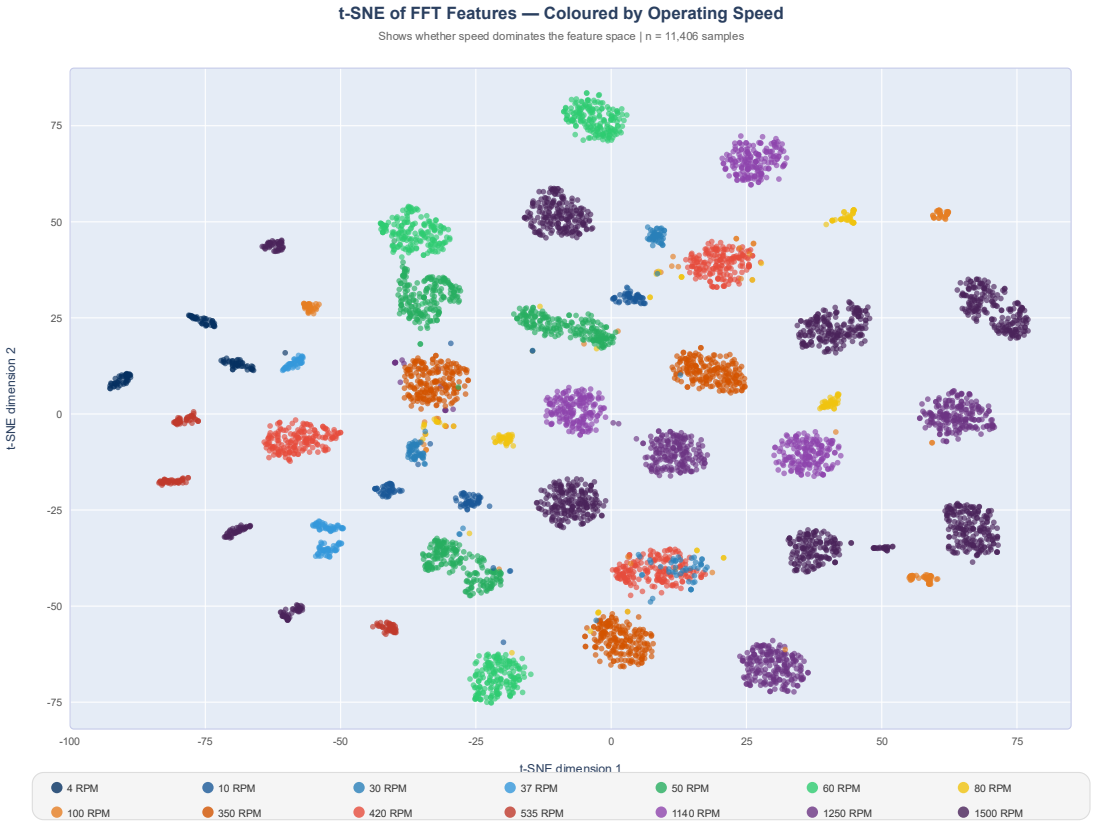


Figure 9. t-SNE projection of FFT features colored by operating speed (RPM).

Figure 9 presents the same embedding colored by operating speed. The same compact islands visible in the class-colored plot now reveal a clear speed-dependent structure. Each island corresponds predominantly to a single speed point with low-speed windows (blue) and high-speed windows (purple) occupying entirely different regions of the embedding space. This dual structure and class-separated globally separated speed locally which has a direct methodological implication.

The t-SNE projections show that when the points are colored based on damage class (Figure 8). The 3 healthy states from largely distinct cluster in the FFT feature space. This gives us strong visual confirmation that the 450-D FFT vector carries enough information

to support good accurate classification. The separation may not be perfect. But it is clear enough to explain why the FFT-based classifiers perform well.

3.4 Supervised Classification Models

This section describes the 4 supervised classification architectures evaluated in this thesis. They are Decision Tree, Random Forest, CNN-FFT and multi-scale Residual Raw CNN. All models were evaluated under the same LOSO cross validation protocol with 57 folds.

The model selection for the bearing health classification is a critical step after data preprocessing and feature extraction. In this study two supervised classical machine learning models were selected they are computationally fast, reliable and interpretable. They are applied to all three feature representations. So, the two classical supervised classifiers were selected for this study:

- Decision Tree
- Random Forest.

Both models were chosen for two complementary reasons. First both are inherently interpretable, and a trained Decision Tree can be rendered as a visual flowchart where every split node, threshold value and leaf class label is directly readable. It allows the diagnostic behavior of each feature representation to be inspected without requiring post-hoc explainability methods. Random Forest additionally exposes feature importance scores that rank the relative contribution of each input variable to the ensemble decision.

Second, both models are computationally lightweight and make no assumptions about input dimensionality or distribution. It makes them equally applicable to the scalar inputs of RMS, kurtosis and the FFT vector under identical experimental conditions.
Both models were evaluated on two classification tasks:

- Binary classification that is Healthy versus Unhealthy and
- Three class classification is Healthy, Lightly Damaged and Heavily Damaged classes.

3.4.1 Decision Tree:

A Decision Tree is a supervised learning algorithm that splits the input feature space into regions by using a series of threshold-based rules. Each internal node of the tree represents a decision based on a feature value. Each branch represents the outcome of that decision, and each leaf node represents a predicted class label. The model is trained by recursively selecting the feature and threshold which best separates the classes according to the Breiman et al., 2017 authors Gini impurity criterion formula is defined as:

$$Gini(t) = 1 - \sum_{i=1}^{C} p(i|t)^2$$

A lower Gini value indicates a purer node with better class separation.
The key configuration parameters used in this study are:

- **Criterion:** Gini impurity
- **Random seed:** 42 randomly selected
- **Class weight:** balanced to handle class imbalance
- **Maximum depth:** varies per feature representation

3.4.2 Random Forest:

Random Forest is an ensemble learning method that builds a collection of Decision Trees and aggregates their predictions task by using majority voting. Each tree in the forest is trained on a randomly sampled subset of the training data. At each split only a random subset of features is considered. According to author Breiman, 2001 this double randomization reduces overfitting and improves generalization compared to a single Decision Tree. The key configuration parameters used in this study are:

- **Maximum depth: varies** per feature representation
- **Random seed: 42** randomly selected
- **Class weight:** 3-class balanced
- **Criterion:** Random ensemble method (Gini impurity)

Classification models Pipeline Evaluation Protocol:

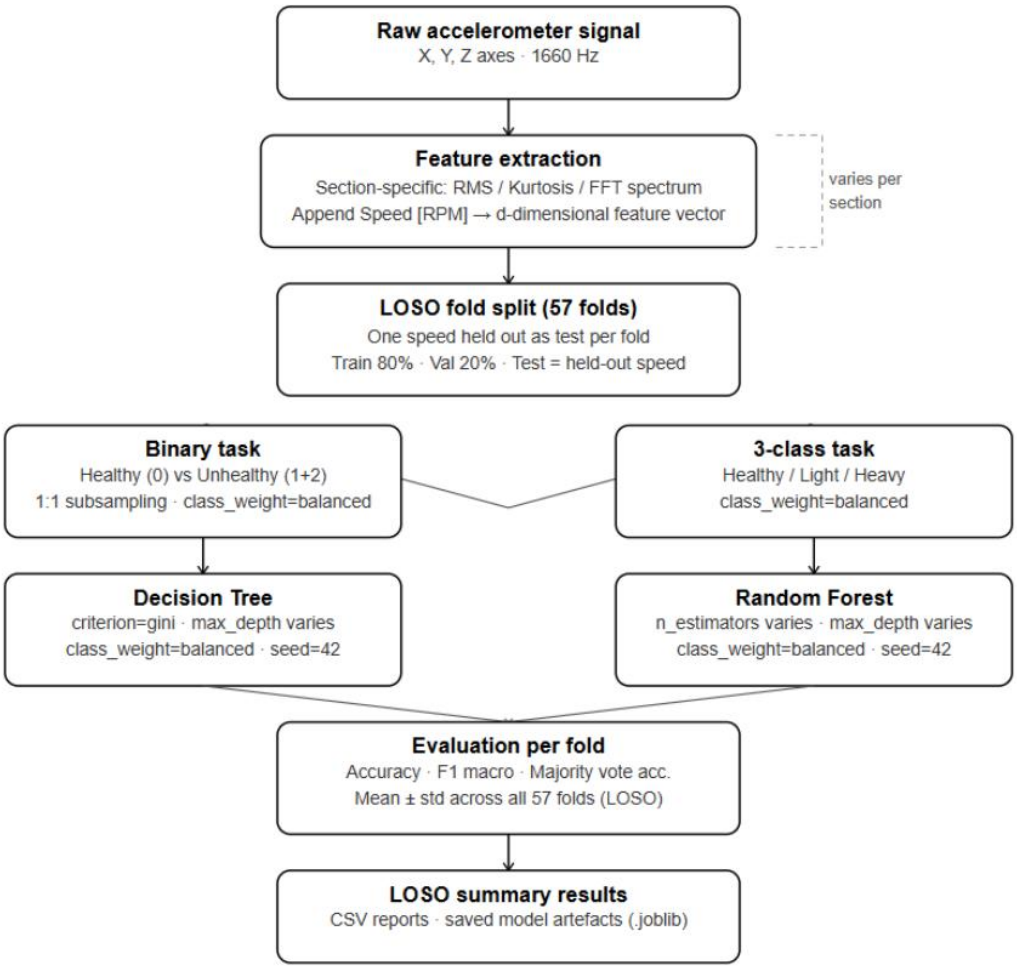


Figure 10. Classical Supervised Machine Learning Pipeline.

Figure 10 illustrates the common pipeline applied identically across all three feature representations. Raw tri-axial accelerometer signals at 1660 Hz were first passed through

feature extraction and it produces either the three scalar RMS values, three scalar kurtosis values or the 450-dimensional FFT magnitude vector depending on the section and the normalized rotational speed appended as a conditioning scalar in all cases. The resulting feature vectors were then split using Leave-One-Speed-Out (LOSO) cross-validation across 57 folds as that much unique speeds in the dataset where one speed point was held out as the test set per fold and the 80 percent of the remaining data was used for training and 20 percent for validation.

Each fold was evaluated on two tasks independently. In the binary task the three health states collapsed into Healthy that is class 0 versus Unhealthy, which is formed by merging Lightly Damaged and Heavily Damaged, that is classes 1 and 2 with strict 1:1 subsampling applied to prevent majority class bias on the decision boundary. In the three-class task all three health states which are Healthy, Lightly Damaged and Heavily Damaged were treated as independent targets with class-weight balancing applied.

Evaluation Protocol:

All classifiers that are both classical machine learning and deep learning models were evaluated under the same Leave-One-Speed-Out cross-validation protocol across 57 speed folds. The evaluation protocol is explained as below:

- In each fold one speed point was held out as the test set and the model was trained at all remaining 56 speeds.
- Within each fold 80 percent of training data was used for model fitting and 20 percent for validation.
- The binary task applied strict 1:1 subsampling to prevent majority class bias on the Unhealthy class.
- The 3-class task used class-weight balancing across Healthy, Lightly Damaged and Heavily Damaged.
- The same performance metrics were applied consistently across all models reported in this thesis.

The metrics are defined as follows:

Accuracy measures the overall proportion of correctly classified windows across all classes:

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

Precision measures how many of the predicted True positive instances were correct:

$$\Pr e\,cision = \frac{TP}{TP + FP}$$

Recall measures of how many of the actual positive instances were correctly identified:

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

F1 Score is the harmonic mean of Precision and Recall. It is particularly important when the dataset is imbalanced because it penalizes models that achieve high accuracy by favoring the majority class:

$$F1 = \frac{2 \times \Pr e\,cision \times Recall}{\Pr e\,cision + Recall}$$

The macro-averaged F1 score extends this to multi-class problems by computing the F1 score independently for each class and taking the unweighted mean across all classes:

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C} F1_i$$

Where C is the total number of classes.

The macro-average F1 score is the primary ranking metric used throughout this thesis for both classical and deep learning models because it weighs all classes equally. It is robust to the class imbalance present in the binary task (Sokolova & Lapalme, 2009). Mean

and standard deviation of all metrics across all 57 LOSO folds are reported as the primary summary statistics. The F1 score reported in all results tables throughout this thesis refers to the macro-averaged F1 score.

3.4.3 CNN FFT

The CNN FFT model has a 1-D CNN that works on FFT frequency spectra instead of raw vibration signals. The input to the model is the same 451-dimensional vector. This vector contains 150 FFT bins from each of the 3 axes giving 450 values.

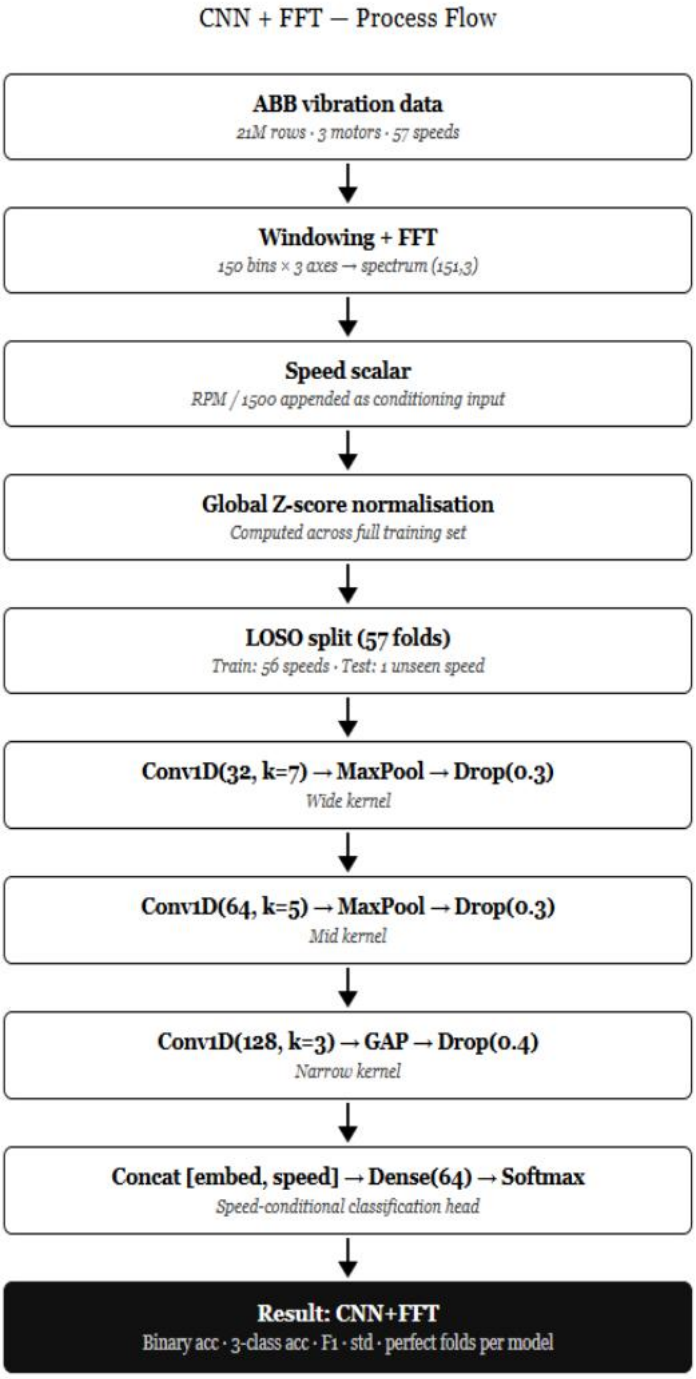


Figure 11. CNN FFT Classification Pipeline.

Figure 11 explains about the CNN FFT architecture and process flow. The network is built from 3 sequential convolutional blocks with decreasing sizes that are wide, mid and narrow to capture spectral patterns. Each layer is supported by Max-pooling to reduce dimensionality and drop out parameter to prevent overfitting. The Global Average pooling is used for feature maps. After the three convolutional blocks the fixed-length embedding vector is joined together with the normalized speed value.

Training Configuration

The model was trained using the Adam optimizer. Categorical cross-entropy was used as the loss function for both tasks. To avoid overfitting early, stopping was applied with a patience of 10 epochs. The same Leave-One-Speed-Out cross-validation protocol was used across all 57 speed folds that is one speed point was held out as the unseen test set and 80% of the remaining data was used for training and 20% for validation.

Table 1. The CNN-FFT model hyperparameters.

Parameter	**Value**
Input shape	(151, 3) FFT magnitude spectrum
Frequency bins	150 per axis
Feature vector	450-dimensional
Speed conditioning	RPM / 1500 normalized scalar
Optimizer	Adam
Learning rate	5*10^-4
Loss function	cross-entropy
Batch size	64
Early stopping patience	3 epochs
Early stopping monitor	Validation loss
Normalization	Global Z-score
Cross-validation	Leave-One-Speed-Out (LOSO)
Number of folds	57
Training speeds per fold	56

Test speed per fold	1 unseen speed
Random seed	42

CNN to FFT input Analysis:

Using a CNN on the pre-computed FFT spectrum rather than the raw waveform has one advantage. The FFT reduces the input from 300 raw time domain samples down to 150 frequency bins per axis. It makes the model lighter and faster to train. At the same time the FFT still preserves the spectral shape differences between healthy and damaged bearings.

3.4.4 Multi-Scale Residual Raw CNN:

Raw CNN is very important deep learning model in this research study. Unlike the FFT-based classifiers it accepts raw tri-axial vibration data directly without any precomputed spectral transformation.

The architecture is a multi-scale residual 1D CNN that processes the raw waveform through 3 parallel convolutional branches.

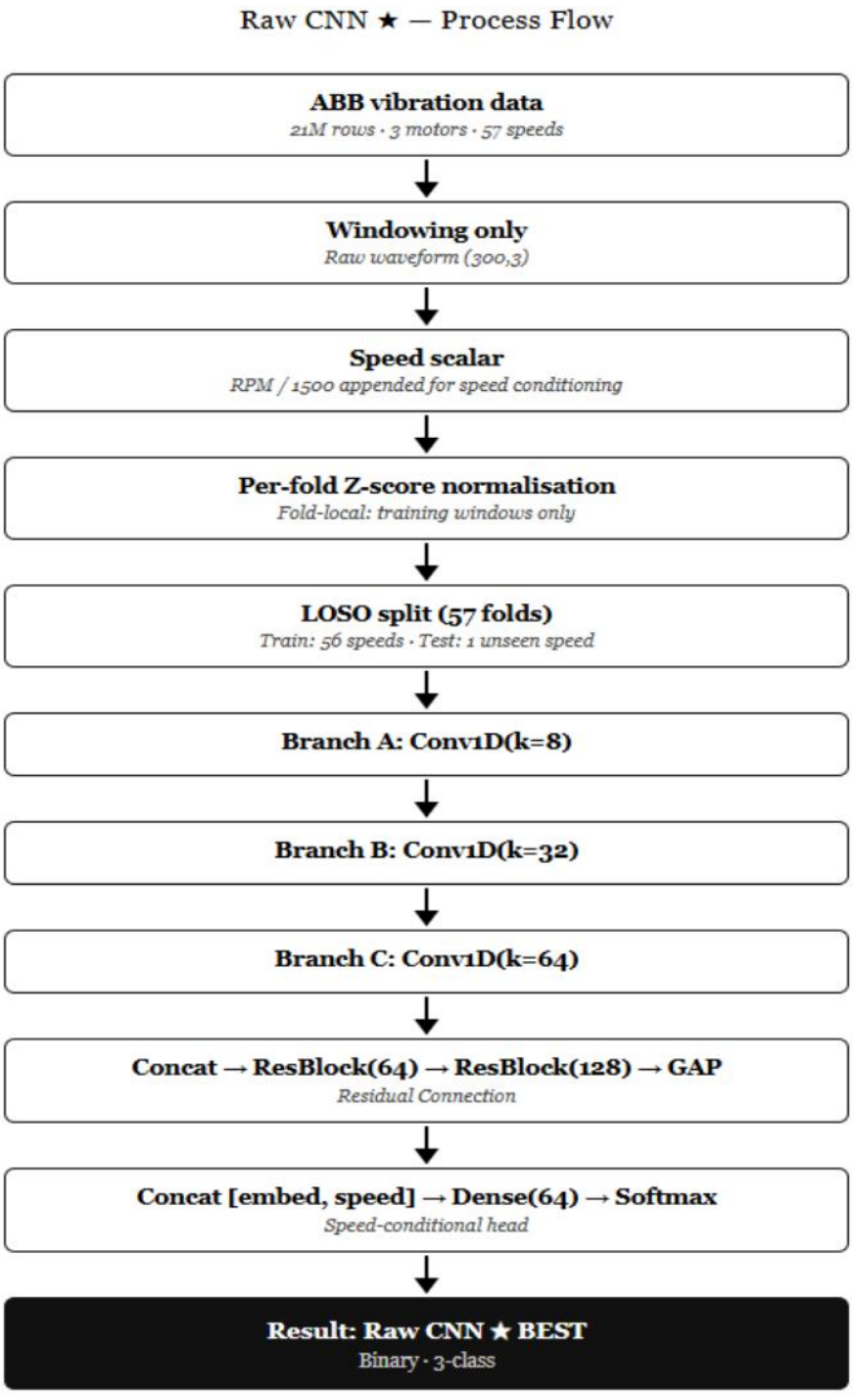


Figure 12. Multi-scale Residual Raw CNN architecture diagram.

Figure 12 explains the Raw CNN architecture network and process flow. The network branches have several kernel sizes such as k=8 for fine-grained local patterns where k=32

for intermediate signals and k=64 for long-term temporal dependencies. It ensures a comprehensive analysis of the data. The resulting features are concatenated and refined through sequential residual blocks and global average pooling. Finally, the model integrates a normalized speed scalar. Finally, it is validated through a Leave-One-Speed-Out cross-validation approach. The training protocol is same as CNN FFT model.

Table 2. Raw CNN model hyperparameters.

Parameter	**Value**
Optimizer	Adam
Learning rate	10^-3
Loss function	sparse categorical cross-entropy
Batch size	64
Early stopping patience	5 epochs
Early stopping monitor	Validation accuracy
Normalization	Per-fold Z-score
Cross-validation	Leave-One-Speed-Out (LOSO)
Number of folds	57
Training speeds per fold	56
Test speed per fold	1 unseen speed
Random seed	42

Confidence-Threshold Rejection:

All the supervised classification models in this thesis output a probability vector of 3 healthy classes. At Inference time a confidence threshold rejection gate was implemented. If the maximum class probability exceeds 0.80 the predicted class label that is healthy or lightly healthy or unhealthy is returned. If the maximum falls at or below 0.80 the model returns an Inspection decision. This is especially important at low and transitional operating speeds where the model tends to be uncertain. This mechanism directly fulfils the Recommend stage of the ABB Detect-Predict-Recommend framework.

3.5 Interpretability of Classification Models

Any ML or DL models that cannot explain which signal features or sensor axes are used for decisions are difficult for maintenance engineers to trust and audit that system. So, achieving high classification accuracy alone is not enough for deployment in industrial application.

This section describes the interpretability methods applied to both the deep learning models and classical machine learning models which are evaluated in this study. The classical models that can be explained are achieved by Tree visualization and feature importance scoring. DeepSHAP attribution analysis applied to both CNN architectures to interpret the deep learning models. This section describes the interpretability methods applied to both machine and deep learning models evaluated in this study.

3.5.1 Decision Tree and Random Forest Interpretability:

A decision Tree visualization exposes every split decision as a readable rule. It shows which feature was selected at each node and the threshold value applied. It is very easy to audit whether the model is relying on the speed conditioning variable or vibration signal content to reach its classification decision.

3.5.2 DeepSHAP Analysis for CNN Models

DeepSHAP was applied independently to two trained Deep learning model architectures:

- The multi-scale Raw CNN and
- CNN+FFT operates on frequency-domain spectra.

This analysis was carried out across 10 representative speed folds that cover the full operating range from 0 to 1500 RPM. The raw 3-axial waveform tensor (300, 3) and normalized speed scalar which tells the model how fast the shaft is spinning at that moment are passed into the model.

For each speed fold, the process worked as follows:

- A background reference set of 100 windows was randomly sampled from the held-out data.
- A fixed random seed of 42 was used to make sure this sampling is reproducible.
- Attribution values were computed for 50 test windows drawn from the same held out speed fold

The background set size of 100 windows was selected to provide stable baseline estimate. Each attribution value tells us how much a specific input feature pushed the model's prediction away from what it would have predicted. In other words, it shows us what the model found unusual or important about a particular test window compared to a typical one. (Lundberg & Lee, 2017). The same process protocol was applied to CNN based FFT model.

To make the results easier to interpret physically a relative axis importance was calculated. It will give the axis' total summed attribution as a percentage of the combined attribution across all three axes. An additional metric was also computed to help understand what the attribution results mean in practice. Class separability delta measure shows us how strongly each axis discriminates between the two most extreme health states. To check whether the models are using the vibration signal or speed to make their decisions. That is why the normalized speed scalar was appended to both CNN architectures.

3.6 Unsupervised Anomaly Detection Models

Unsupervised anomaly detection addresses a different problem from supervised classification. An anomaly detector is trained exclusively on healthy bearing data and learns to recognize normal behavioral inference time any test data that deviates from the learned healthy distribution is flagged as anomaly. As in real time environment without requiring labelled fault data during training this approach directly maps to the ABB Detect stage of the Detect-Predict-Recommend framework identifying deviations from normal behavior.

Three complementary anomaly detection models were evaluated in this study:
1 Isolation Forest — tree-based anomaly scoring on FFT features.
2 FFT Autoencoder — spectral reconstruction error-based detection.
3 CNN Autoencoder — raw waveform reconstruction error-based detection.

Each model was trained exclusively in Healthy class windows data and evaluated on both Lightly Damaged and Heavily Damaged data.
The Performance was measured using three metrics:

- Accuracy: overall proportion of correctly classified windows
- Detection Rate (DR): proportion of actual fault windows correctly flagged as anomaly
- False Alarm Rate (FAR): proportion of healthy windows incorrectly flagged as anomaly

The key tradeoff in anomaly detection is between DR and FAR. A model that flags everything as anomalous achieves 100% DR but also 100% FAR. A good model anomaly detector must achieve high DR while keeping FAR low.

3.6.1 Isolation Forest

Isolation Forest is a tree based unsupervised anomaly detection algorithm that identifies anomalies. According to author Liu et al., 2008 anomalies are detected by isolating observations through random feature splitting. An anomaly score is assigned to each sample based on the average path length required to isolate it. The shorter paths indicate more anomalous samples.

The key advantages of Isolation Forest for this study are:

- It requires no assumptions about the distribution of the data
- It is computationally efficient and scales well to high dimensional inputs
- It naturally handles the 450-dimensional FFT feature vector without dimensionality reduction
- It provides a continuous anomaly score that can be threshold at different levels

Table 3.The Isolation Forest model hyper parameters.

Parameter	**Value**
Input feature	FFT magnitude spectrum
Number of estimators	200 trees
Contamination	Auto
Threshold	5th percentile of healthy training scores per fold
Training data	Healthy class windows only
Test data	Lightly Damaged and Heavily Damaged windows
Cross-validation	Leave-One-Speed-Out (57 folds)
Random seed	42

Isolation Forest Architecture:

The model takes only healthy class windows as input where no fault data is seen during training. Each window is transformed into a 450-D FFT feature vector using windowing and spectral extraction. The Isolation Forest, with 200 estimators, is then trained on

these healthy features and assigns an anomaly score to each window based on isolation depth. The shorter isolation paths indicate more anomalous behavior.

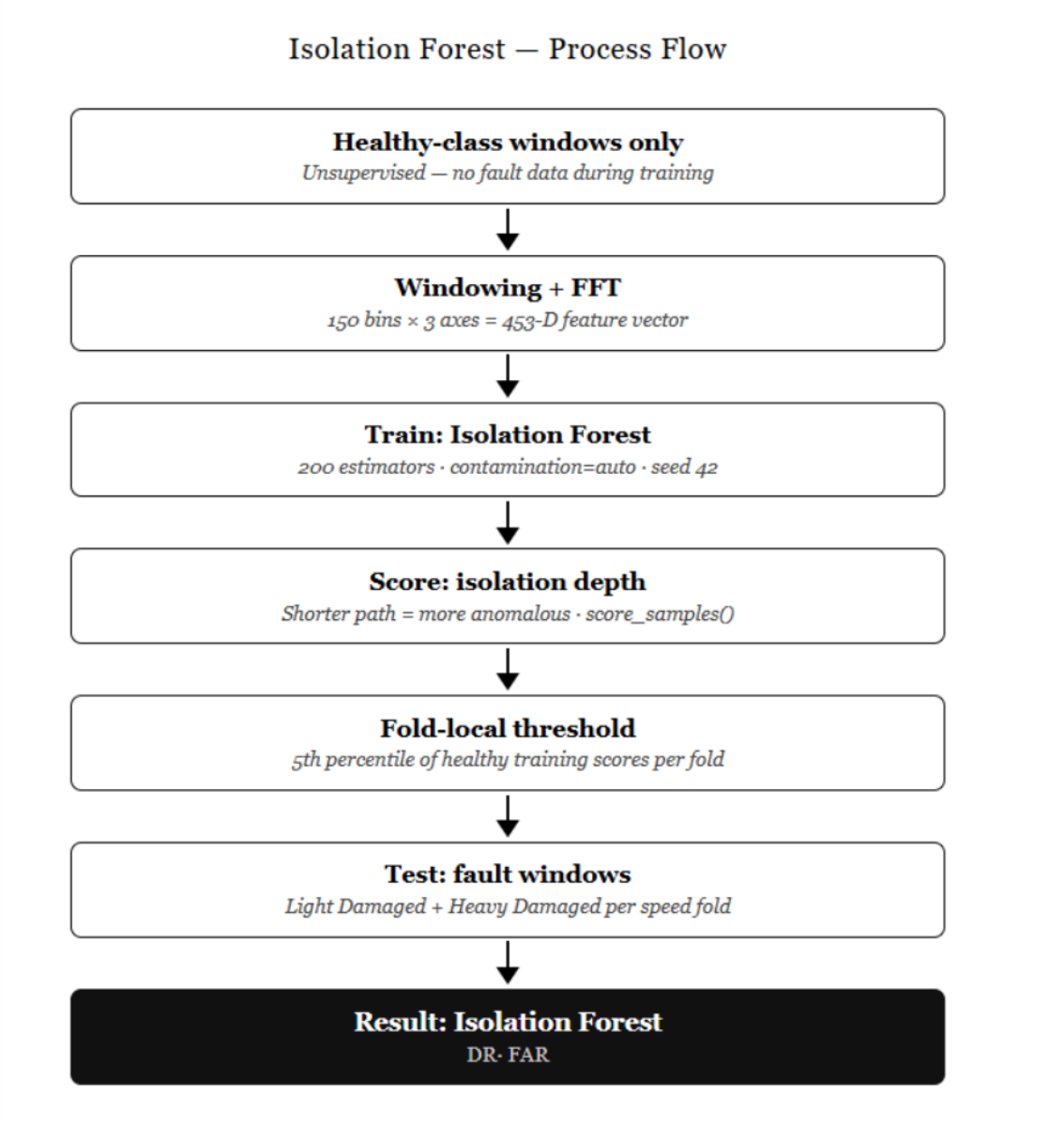


Figure 13. Isolation Forest process flow.

Figure 13 illustrates the Isolation Forest process pipeline. A fold local threshold set at the 5th percentile of healthy training scores determines the anomaly decision boundary.

At testing phase Lightly Damaged and Heavily Damaged windows from the held-out speed fold are scored against this threshold and classified as either normal or anomalous. Final performance is reported as Detection Rate and False Alarm Rate per speed fold.

3.6.2 FFT Feature Autoencoder

The FFT Autoencoder is a convolutional autoencoder that teaches to reconstruct the FFT magnitude spectrum of healthy bearing vibration signals. According to authors Goodfellow et al., 2016 an autoencoder consists of two components, first an encoder that compresses the input into a low-dimensional latent representation and second a decoder that reconstructs the original input from that compressed representation. The anomaly detection principle is a model trained exclusively on healthy spectral patterns, and it will try to learn to reconstruct healthy windows accurately. It will produce high reconstruction errors only for damaged bearing windows whose spectral morphology deviates from the learned healthy distribution.

The key advantages of the FFT Autoencoder for this study are:

- It operates in the frequency domain.
- Reconstruction error provides a continuous anomaly score
- It requires no fault labels during training

Table 4. FFT Autoencoder model hyper parameters

Parameter	**Value**
Input shape	(151, 3), FFT magnitude spectrum
Frequency bins	150 per axis
Normalization	Global, spec_mean and spec_std from healthy data
Encoder	Conv1D layers
Decoder	Conv1D layers
Anomaly score	Mean Squared Error, a spectral reconstruction error
Threshold type	Global
Threshold value	95th percentile of pooled healthy training errors
Training data	Healthy class windows only

Test data	Lightly Damaged + Heavily Damaged per speed fold
Cross-validation	Leave-One-Speed-Out (57folds)
Random seed	42

FFT Autoencoder Architecture:

The architecture consists of a Convolutional 1D encoder that compresses the input FFT spectrum through a Lambda resize layer into a bottleneck representation. It is followed by a Convolutional 1D decoder that reconstructs the original spectrum.

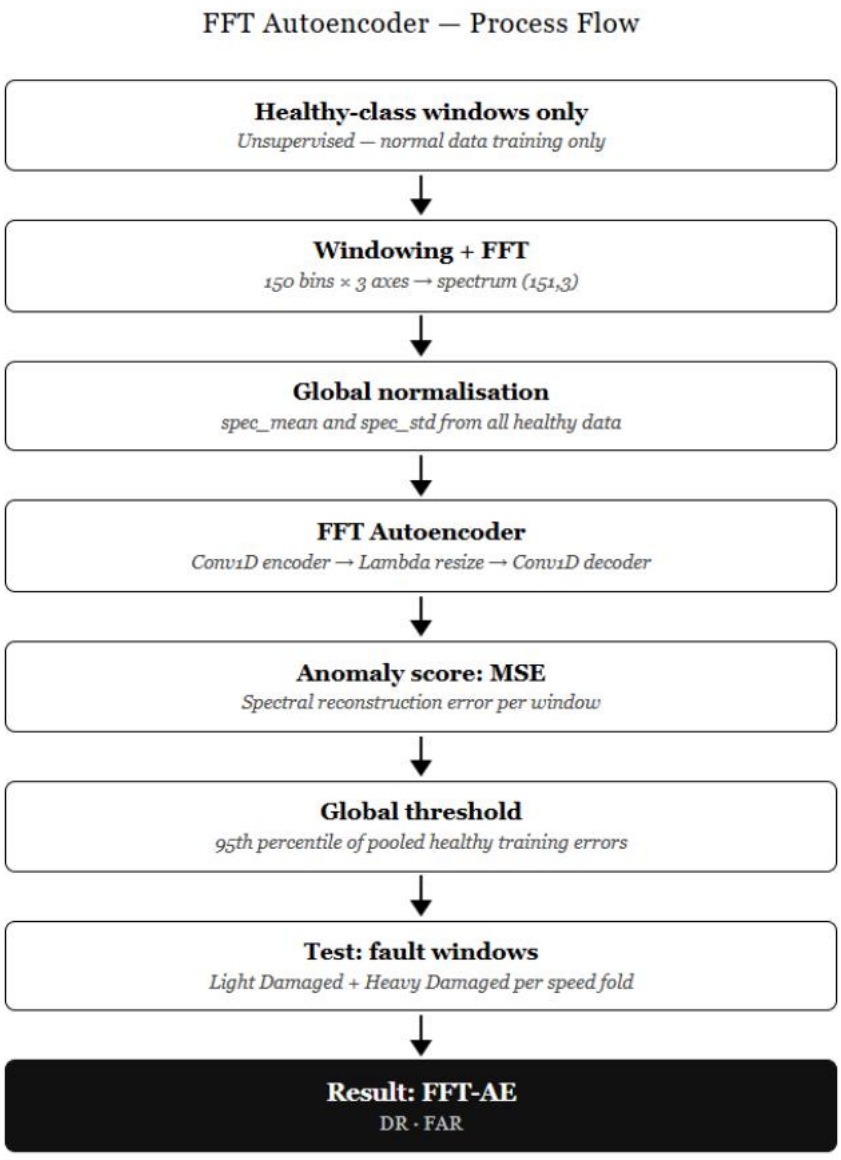


Figure 14. FFT Feature Autoencoder process flow.

Figure 14 displays the FFT autoencoder pipeline process flow. The anomaly score for each window is the Mean Squared Error (MSE). A global threshold set at the 95th percentile of pooled healthy training reconstruction errors. It determines the anomaly decision boundary windows exceeding this threshold are classified as anomalous.

3.6.3 CNN Autoencoder

The CNN Autoencoder is a convolutional autoencoder that learns to reconstruct raw 3-axial vibration waveforms of healthy bearing signals directly. Unlike the FFT Autoencoder which operates on pre-computed on frequency spectra the CNN Autoencoder processes the raw waveform tensor of shape (300, 3). This allows the model to learn waveform morphology and temporal patterns of healthy bearing patterns. It may not be fully captured by frequency domain representations alone.

The key advantages of the CNN Autoencoder for this study are:

- It operates directly on raw waveforms.
- It does not require any hand-crafted feature extraction.
- Reconstruction error provides an interpretable anomaly score

CNN Autoencoder Architectures

The architecture consists of a Convolutional 1D encoder that compresses the raw waveform through two convolutional layers. The max pooling will convert these vectors into a Dense bottleneck of 64 units. The decoder reconstructs the original waveform shape using Dense layers. It is followed by two Convolutional 1D layers producing the final (300, 3) output.

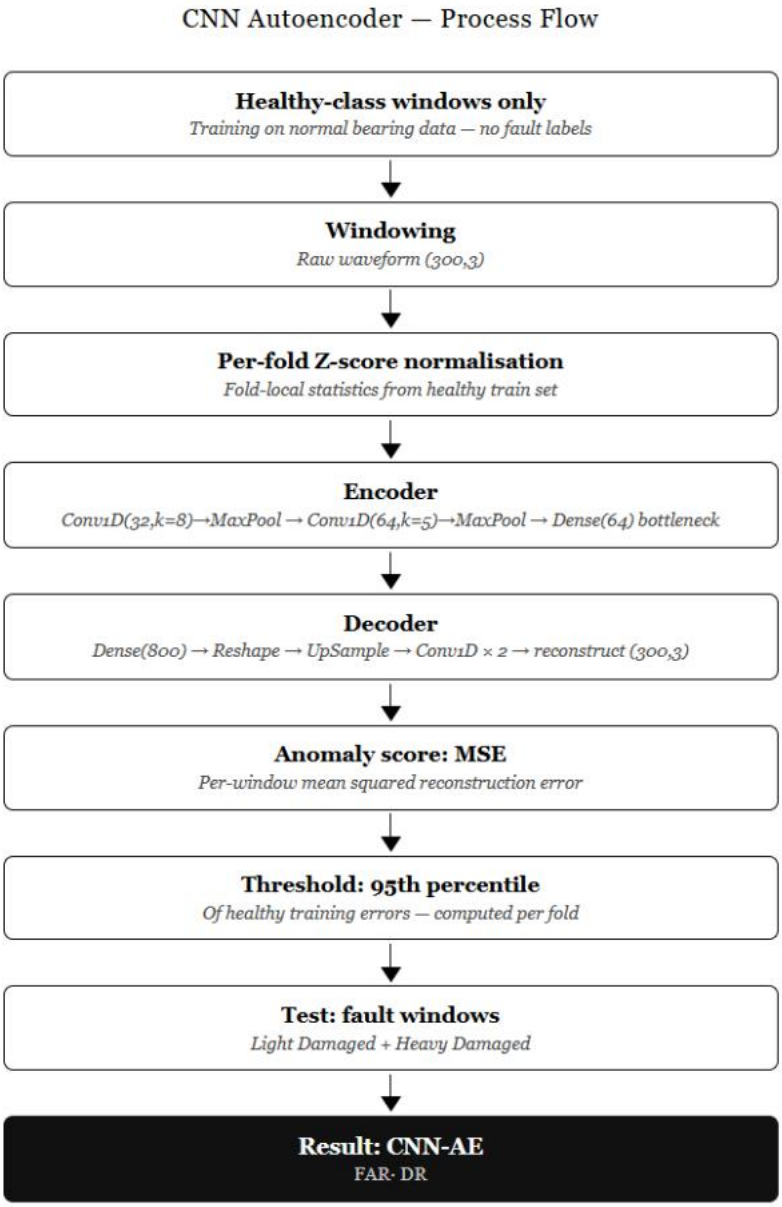


Figure 15.CNN Raw Autoencoder process flow.

Figure 15 explains CNN Raw Autoencoder architecture. The anomaly score for each window is the Mean Squared Error (MSE). A per-fold threshold set at the 95th percentile of healthy training reconstruction errors. These errors determine the anomaly decision boundary which ensures the threshold adapts to the waveform amplitude of each speed point.

4 Results

This chapter presents the experimental results of all supervised classification models and unsupervised anomaly detectors evaluated in this study. Results are reported in the following order: Sections 4.1 to 4.3 cover the three classical feature-based classifiers which are kurtosis, RMS, and FFT. They are evaluated under both Decision Tree and Random Forest architecture. Sections 4.4 and 4.5 report the two deep learning classifiers, CNN-FFT and the multi-scale Residual Raw CNN respectively. Section 4.6 presents the Deep-SHAP interpretability results for both CNN architectures. Section 4.7 reports the unsupervised anomaly detection results for all three detectors. Section 4.8 provides an overall comparative performance summary. All models were evaluated under the Leave-One-Speed-Out (LOSO) cross-validation protocol across 57 speed folds as described in Section 3.4. Mean accuracy and macro-averaged F1 score across all folds are the primary reported metrics.

4.1 Kurtosis Classification Results:

Kurtosis Binary class Analysis by Decision Tree and Random Forest:

Binary class by Decision Tree:

The below Table 5 presents the selected per-speed Decision Tree binary classification results for the kurtosis feature under LOSO evaluation across 10 operational speeds.

Table 5. Kurtosis binary classification results – Decision Tree (selected speeds).

Speed	Accuracy	F1 Score
50	51.22%	50.92%
100	29.28%	29.28%
500	71.44%	69.75%

535	17.94%	17.91%
600	74.33%	72.84%
800	56.33%	56.29%
900	52.56%	52.38%
994	42.97%	39.94%
1200	89.22%	88.64%
1500	63.78%	54.52%

The results show high variability across speed folds. The binary classifier performs best at 1200 RPM reaching 89.22% accuracy but collapses to 17.94% at 535 RPM. It indicates that the kurtosis feature at this speed is actively misleading the classifier. These results confirm that kurtosis is not a good feature for binary bearing health classification.

Binary class by Random Forest:

Table 6 below presents the per-speed Random Forest binary classification results for the kurtosis feature.

Table 6. Kurtosis binary classification results – Random Forest (selected speeds)**.**

Speed	**Accuracy**	**F1 Score**
50	59.33%	55.44%
100	52.30%	42.97%
500	71.33%	68.50%
535	42.37%	41.65%
600	73.78%	70.35%
800	66.00%	63.58%
900	62.67%	60.56%
994	51.33%	50.24%
1200	91.78%	91.22%
1500	62.67%	50.51%

The Random Forest improves marginally over the Decision Tree. The best performance again occurs at 1200 RPM at 91.78% while the weakest remains at 535 RPM at 42.37%. The limited improvement from Decision Tree to Random Forest confirms that the problem lies in the kurtosis feature itself rather than classifier capacity. The ensemble aggregation cannot recover discriminative structure that is not present in the input features.

3-class Classification by Decision Tree:

Table 7 presents the per-speed Decision Tree three-class classification results for kurtosis.

Table 7. Kurtosis three-class classification results – Decision Tree (selected speeds).

Speed	**Accuracy**	**F1 Score**
50	38.56%	38.64%
100	37.17%	37.72%
500	65.78%	65.53%
535	19.08%	21.09%
600	66.56%	66.40%
800	66.56%	66.26%
900	51.78%	51.73%
994	38.02%	26.44%
1200	60.67%	57.81%
1500	52.33%	50.70%

The three-class task exposes the complete failure of kurtosis as a severity discriminator. Performance is highly inconsistent across speeds at 500 and 600 RPM. The classifier reaches 65–66% while at 535 RPM it collapses to 19.08% and at 100 RPM to 37.17%. This erratic pattern across neighboring speed points confirms that kurtosis does not encode a stable relationship with damage severity across the operational speed range.

3-class Classification by Random Forest:

Table 8 below presents the per-speed Random Forest 3-class classification results for kurtosis.

Table 8. Kurtosis three-class classification results – Random Forest (selected speeds).

Speed	Accuracy	F1 Score
50	38.89%	38.72%
100	46.38%	46.10%
500	60.22%	59.18%
535	34.73%	25.41%
600	64.67%	64.57%
800	61.78%	62.35%
900	55.56%	55.82%
994	44.11%	35.64%
1200	71.67%	70.42%
1500	52.22%	51.62%

The Random Forest three-class accuracy and F1 score shows negligible improvement over the Decision Tree. Across all 57 LOSO folds the Random Forest produced zero majority-vote wins on the three-class task and no single held-out speed fold has a correct majority class prediction when distinguishing between all three health states. This is the strongest empirical evidence that kurtosis is fundamentally unsuitable as a three-class severity discriminator for this dataset.

4.2 RMS Classification Results

Binary Classification by Decision Tree:

Table 9 below presents the per-speed Decision Tree binary classification results for the RMS feature under LOSO evaluation across 10 representative speed points.

Table 9.RMS binary classification results – Decision Tree (selected speeds).

Speed	Accuracy	F1 Score
50	28.11%	30.56%
100	88.16%	88.28%
500	99.67%	99.52%
535	95.42%	95.48%
600	97.56%	99.55%
800	99.67%	98.15%
900	98.89%	98.39%
994	98.10%	98.09%
1200	98.89%	98.28%
1500	78.00%	76.28%

The RMS Decision Tree binary results represent a dramatic improvement over kurtosis. At all speeds above 100 RPM the classifier achieves between 88% to 99 % accuracy approximately. The results confirm that RMS amplitude provides strong energy separation between the Healthy and Unhealthy classes across the mid to high-speed range. The only notable weakness occurs at 50 RPM where accuracy drops to 28.11%. At very low rotational speeds the bearing generates minimal mechanical excitation and the energy difference between healthy and damaged states collapses which is making RMS an unreliable binary discriminator at this speed point.

Binary Classification by Random Forest:

Table 10 below presents the per-speed Random Forest binary classification results for the RMS feature.

Table 10.RMS binary classification results – Random Forest (selected speeds)**.**

Speed	Accuracy	F1 Score
50	57.33%	51.59%
100	95.72%	95.09%
500	100.00%	100.00%
535	98.86%	98.72%
600	100.00%	100.00%
800	100.00%	100.00%
900	100.00%	100.00%
994	100.00%	100.00%
1200	100.00%	100.00%
1500	100.00%	100.00%

The Random Forest binary results for RMS have strong performance. One major improvement when compared to decision tree is at 994 RPM the Random Forest reaches 100% accuracy compared to 96.20% for the Decision Tree. At 1500 RPM the Random Forest achieves 100% compared to 98.33%. This confirms that ensemble aggregation provides a small but consistent improvement at speeds where a single tree finds marginal class boundaries.

3-class classification by Decision Tree:

Table 11 below presents the per-speed Decision Tree three-class classification results for the RMS feature

Table 11. RMS three-class classification results – Decision Tree (selected speeds).

Speed	Accuracy	F1 Score
50	58.89%	52.68%
100	95.39%	94.70%
500	100.00%	100.00%
535	98.86%	98.72%
600	100.00%	100.00%
800	100.00%	100.00%
900	100.00%	100.00%
994	96.20%	95.82%
1200	100.00%	100.00%
1500	98.33%	98.15%

The RMS Decision Tree three-class results have strong performance across the mid to high range speeds maintaining an accuracy of 95.39% to 100%. The decision boundary that separates Healthy from Unhealthy in the binary task also separates the three classes cleanly in the three-class task at mid-to-high speeds. The low-speed weakness at 50 RPM remains at 58.89% states that it is not good bearing excitation at very low rotational velocity.

3-class Classification by Random Forest:

Table 12 below presents the per-speed Random Forest three-class classification results of the RMS feature.

Table 12.RMS three-class classification results –Random Forest (selected speeds).

Speed	Accuracy	F1 Score
50	41.78%	44.05%
100	85.20%	85.47%
500	99.89%	99.89%

535	89.31%	88.99%
600	99.44%	99.44%
800	99.33%	99.33%
900	99.44%	99.44%
994	100.00%	100.00%
1200	99.67%	99.67%
1500	83.89%	83.06%

The Random Forest three-class results reveal the key limitation of RMS as a severity discriminator. At 50 RPM accuracy drops sharply to 41.78% and to 85.20% at 100 RPM. At 1500 RPM accuracy falls to 83.89% compared to 98.33% for the Decision Tree binary task. This three-class degradation at speed extremes is the fundamental limitation that motivates the evaluation of the full FFT spectrum.

4.3 FFT Classification Results:

Binary Classification by Decision Tree:

Table 13 below presents the per-speed Decision Tree binary classification results for the FFT feature under LOSO evaluation across 10 representative speed points.

Table 13.FFT binary classification results – Decision Tree (selected speeds).

Speed	**Accuracy**	**F1 Score**
50	66.27%	66.26%
100	90.13%	88.38%
500	98.67%	98.49%
535	55.34%	54.79%
600	100.00%	100.00%
800	99.07%	98.94%
900	98.67%	98.49%
994	67.68%	67.67%
1200	99.73%	99.70%

1500	97.07%	96.74%

The FFT Decision Tree binary results show a distinctly different pattern compared to both kurtosis and RMS. At mid-to-high speeds the classifier performs strongly and achieves 100% at 600 RPM and above 97% at 1200 and 1500 RPM. However, two speed points stand out as challenging 535 RPM where accuracy drops to 55.34% and 994 RPM where it falls to 67.68%. These are genuine physical challenges rather than featuring limitations. Importantly at 50 RPM the FFT Decision Tree achieves 66.27% when compared to kurtosis at 51.22% and RMS at 58.89% at the same speed demonstrating that spectral shape information provides better low-speed discrimination than scalar energy statistics even for a simple single tree classifier.

Binary Classification by Random Forest:

Table 14 below presents the per-speed Random Forest binary classification results for the FFT feature.

Table 14.FFT binary classification results – Random Forest (selected speeds).

Speed	**Accuracy**	**F1 Score**
50	98.40%	98.20%
100	99.67%	99.63%
500	100.00%	100.00%
535	95.04%	94.60%
600	100.00%	100.00%
800	100.00%	100.00%
900	100.00%	100.00%
994	100.00%	100.00%
1200	100.00%	100.00%
1500	100.00%	100.00%

The FFT Random Forest binary results are the strongest of all classical models evaluated in this thesis. The transition from Decision Tree to Random Forest produces the largest

relative gain of any feature as at 50 RPM accuracy jumps from 66.27% to 98.40%. At 535 RPM accuracy rises from 55.34% to 95.04%. This improvement confirms that the 450-dimensional FFT feature space contains rich discriminative information that a single decision tree cannot reliably do partition. At 7 of the 10 representative speed points the FFT Random Forest achieves 100% binary accuracy. The 50 RPM have 98.4% where both kurtosis and RMS struggled significantly.

3-class Classification by Decision Tree:

Table 15 below presents the per-speed Decision Tree three-class classification results for the FFT feature.

Table 15. FFT three-class classification results – Decision Tree (selected speeds)**.**

Speed	**Accuracy**	**F1 Score**
50	73.07%	72.92%
100	90.13%	90.14%
500	98.93%	98.93%
535	59.16%	54.92%
600	100.00%	100.00%
800	99.87%	99.87%
900	98.00%	97.99%
994	84.79%	84.39%
1200	97.33%	97.33%
1500	97.60%	97.60%

The FFT Decision Tree 3-class results demonstrate a critical advantage over RMS and kurtosis. The FFT Decision Tree maintains strong three-class task performance across the full speed range. At 50 RPM the FFT Decision Tree achieves 73.07% 3 class accuracy compared to only 58.89% for RMS Decision Tree at the same speed.

3-class Classification by Random Forest:

Table 16 below presents the per-speed Random Forest 3 class classification results for the FFT feature.

Table 16. FFT three-class classification results – Random Forest (selected speeds)**.**

Speed	Accuracy	F1 Score
50	99.07%	99.07%
100	98.03%	98.02%
500	100.00%	100.00%
535	66.03%	53.40%
600	100.00%	100.00%
800	100.00%	100.00%
900	100.00%	100.00%
994	100.00%	100.00%
1200	100.00%	100.00%
1500	100.00%	100.00%

The FFT Random Forest 3 class results represent the best overall performance among all classical supervised models in this study. At 50 RPM the classifier achieves 99.07% three-class accuracy when compared to only 41.78% for RMS Random Forest at the same speed. At 9 of the 10 representative speed points the FFT Random Forest achieves above 98% three-class accuracy, which proves that FFT outperformed compared to remaining features that are kurtosis and RMS.

4.4 CNN-FFT Classification Results:

CNN_FFT Binary Classification:

Table17 below presents the per-speed CNN_FFT binary Classification under LOSO evaluation across 10 representative speed points.

Table 17.CNN-FFT binary classification results (selected speeds).

Speed	Accuracy	F1 Score
50	40.76%	37.11%
100	97.04%	96.70%
500	100.00%	100.00%
535	91.60%	91.03%
600	100.00%	100.00%
800	100.00%	100.00%
900	100.00%	100.00%
994	97.72%	97.47%
1200	100.00%	100.00%
1500	95.69%	95.45%

The CNN FFT binary results show a clear and consistent pattern across the speed range. At mid-to-high speeds the model performs very strongly. At 500, 600, 800, 900 and 1200 RPM it achieves perfect 100% accuracy and F1. At 100 RPM and 994 RPM it still performs very well at 97.04% and 97.72%. At 535 RPM the model reaches 91.60% accuracy. The only speed point where the binary model clearly fails is 50 RPM where accuracy drops to 40.76% and F1 falls to 37.11%. At 1500 RPM the model achieves 95.69% accuracy, which is good but slightly lower than the mid-range speeds.

CNN_FFT 3-class Classification:

Table 18 below presents the per-speed CNN_FFT 3-class Classification under LOSO evaluation across 10 representative speed points.

Table 18. CNN-FFT three-class classification results (selected speeds).

Speed	Accuracy	F1 Score
50	41.89%	33.80%
100	98.03%	98.04%
500	100.00%	100.00%
535	45.42%	36.74%
600	100.00%	100.00%
800	99.81%	99.81%
900	100.00%	100.00%
994	55.89%	47.03%
1200	100.00%	100.00%
1500	97.61%	97.77%

The 3-class results reveal an interesting and unexpected pattern when compared to the binary results above. At 535 RPM accuracy falls from 91.60% in the binary task down to 45.42% in the 3-class task. At 994 RPM accuracy drops from 97.72% down to 55.89%. These two speed points stand out as genuine weak spots for the 3-class CNN FFT model.

At 50 RPM the 3-class model again fails with 41.89% accuracy and 33.80% F1. It achieves 100% accuracy at 500, 600, 900 and 1200 RPM and above 97% at 100 and 1500 RPM. The model successfully distinguishes all three health states including the finer boundary between lightly and heavily damaged bearings at most of the speed range.

4.5 Raw CNN Classification Results

CNN Raw data Binary Classification:

Table 19 below presents the per-speed Raw CNN binary classification results under LOSO evaluation across 10 representative speed points.

Table 19. Raw CNN binary classification results (selected speeds).

Speed	Accuracy	F1 Score
50	98.95%	98.81%
100	100.00%	100.00%
500	100.00%	100.00%
535	99.24%	99.14%
600	100.00%	100.00%
800	100.00%	100.00%
900	100.00%	100.00%
994	100.00%	100.00%
1200	100.00%	100.00%
1500	99.95%	99.95%

The Raw CNN binary results are the strongest of all models evaluated in this thesis. At 9 of the 10 representative speed points the classifier achieves 100% accuracy. The only two speeds below 100% are 50 RPM,1500 RPM at 98.95% and at 99.95% but both of which still exceed all classical ML models at the same speeds. Most critically at 50 RPM where kurtosis achieved only 59.33%, RMS achieved 57.33% and FFT Random Forest achieved 98.40%, the Raw CNN achieves 98.95% which is the the highest binary accuracy at this challenging low speed point across all models. At 535 RPM which was the most challenging speed for all previous models, the Raw CNN achieves 99.24%accuracy. This model is the most reliable binary bearing health classifier across the full 0-1500 RPM operational range.

CNN 3-class classification:

Table 20 below presents the per-speed Raw CNN three-class classification results under LOSO evaluation.

Table 20. Raw CNN three-class classification results (selected speeds)**.**

Speed	Accuracy	F1 Score
50	97.89%	97.89%
100	100.00%	100.00%
500	100.00%	100.00%
535	96.56%	96.56%
600	100.00%	100.00%
800	100.00%	100.00%
900	100.00%	100.00%
994	100.00%	100.00%
1200	100.00%	100.00%
1500	90.21%	90.82%

The Raw CNN 3 class results confirm it as the best performing model across all classifiers evaluated in this study. At 50 RPM the Raw CNN achieves 97.89% three-class accuracy compared to only 41.78% for RMS and Random Forest and 99.07% for FFT Random Forest at the same speed. The 3-class performance remains above 96% at all speed points except 1500 RPM where it drops to 90.21%. At 535 RPM Raw CNN achieves 96.56% three-class accuracy which is the highest of any model at this challenging speed.

4.6 Interpretability of Classification of models:

4.6.1 Decision Tree FFT Visualisation

Below are the two visualizations figures of FFT algorithms

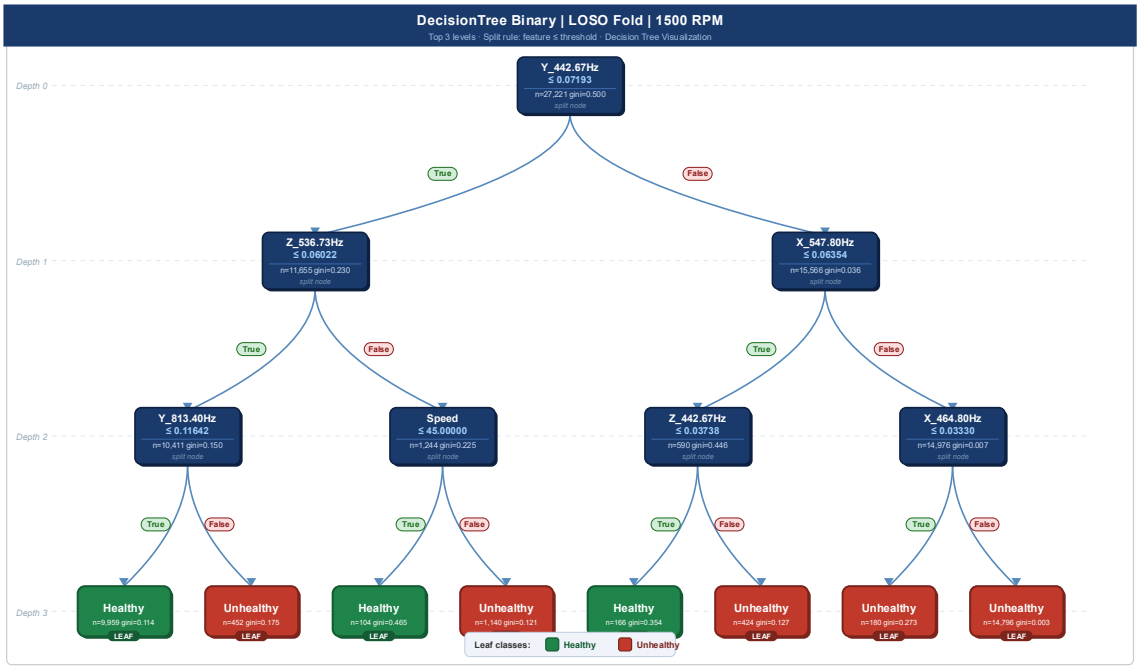


Figure 16. Decision Tree FFT Binary classification at 1500 RPM.

Figure 16 displays the decision Tree FFT binary classification at 1500 RPM. The FFT Decision Tree binary structure at 1500 RPM roots on Y 442.6 Hz which is a mid-range frequency bin on the Y-axial axis. It indicates that axial spectral content at this frequency provides the strongest single class separation between healthy and unhealthy. One of the key observations is speed appears only at depth 2 a threshold of 40 RPM. In the diagram, which has top 3 levels of flow of the classification.

4.6.2 Random Forest FFT 3 classification

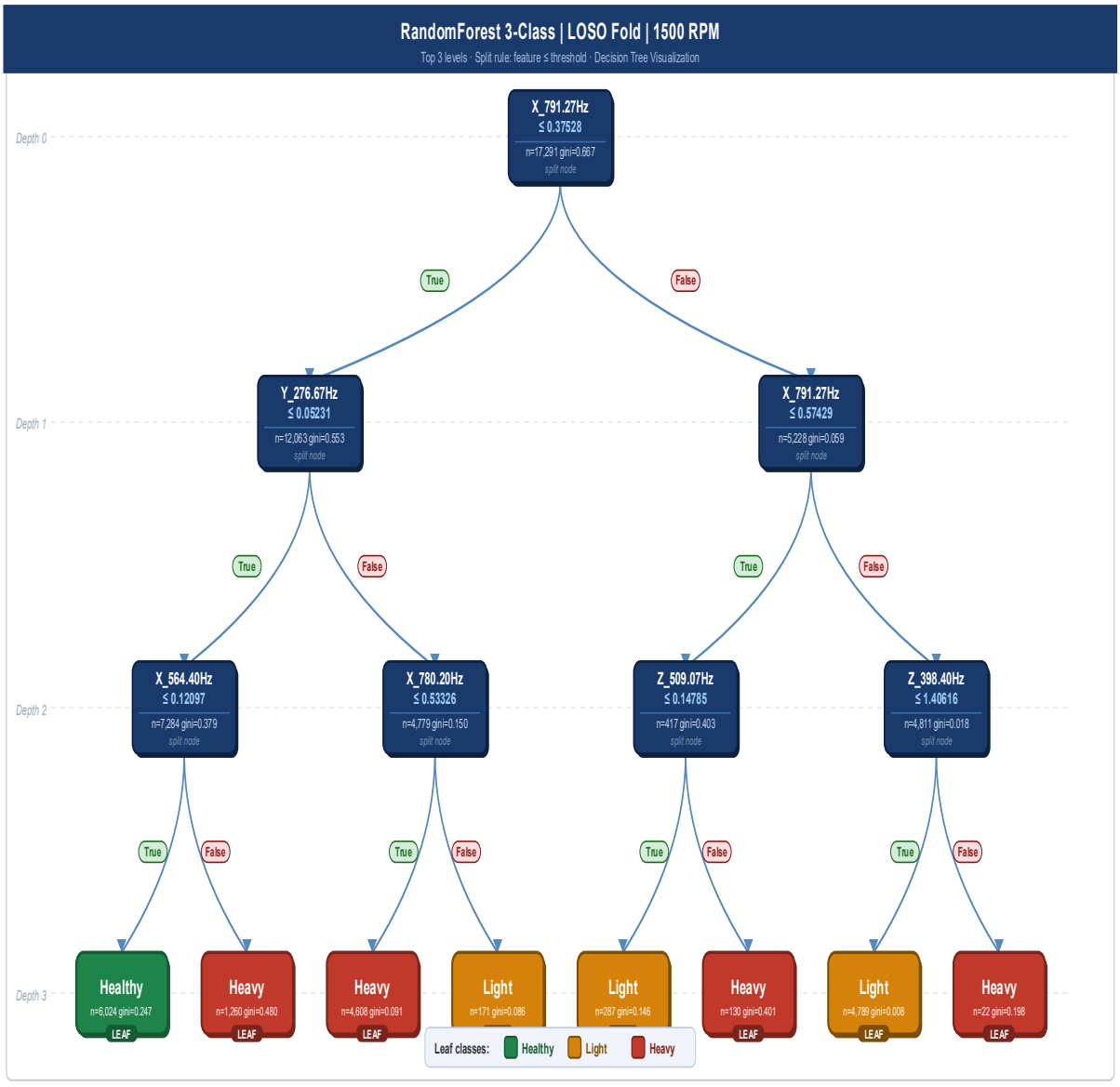


Figure 17. Random Forest FFT 3-class classification at 1500 RPM.

Figure 17 displays the Tree visualization of the Random Forest FFT 3-class classification at 1500 RPM. The FFT Random Forest 3-class tree of Top 3 level depth at 1500 RPM shows a single estimator structure rooted on X 791.2 Hz. A high frequency bin on the X-

radial axis with a Gini impurity of 0.667 at the root shows the balanced 3-class distribution across the full training set. Speed does not appear at any level. The complete absence of speed from the visualized tree levels confirms that the FFT spectral representation provides good class discriminative structure.

4.6.3 DeepSHAP CNN FFT:

As shown in Table 21, the X-radial and Z-radial axes share the attribution much more equally. X averages 40.82% and Z averages 37.85%. It means neither axis clearly dominates on its own. The Y-axial axis contributes 21.33 per cent on average. At four of the 10 speed folds the Z-radial axis is the dominant contributor.

Table 21. DeepSHAP three-class axis analysis - CNN FFT model (selected speeds).

Speed (RPM)	X imp. (%)	Y imp. (%)	Z imp. (%)	Sep. Δ X	Sep. Δ Y	Sep. Δ Z
50	39.44	9.57	50.99	0.2429	0.0155	0.196
100	46.63	20.08	33.28	0.0451	0.0597	0.0818
500	47.65	16.1	36.25	0.0499	0.0146	0.0307
535	32.94	30.11	36.94	0.0046	0.0237	0.0243
600	43.39	22.42	34.19	0.0063	0.0104	0.0333
800	46.33	19.25	34.42	0.0161	0.0048	0.0412
900	46.68	17.91	35.41	0.0223	0.0162	0.061
994	24.27	29.08	46.65	0.1304	0.0306	0.2373
1200	41.51	23.7	34.79	0.0132	0.0025	0.0358
1500	39.4	25.04	35.57	0.0688	0.0279	0.1007
Mean	**40.82%**	**21.33%**	**37.85%**	**0.06**	**0.0206**	**0.0842**

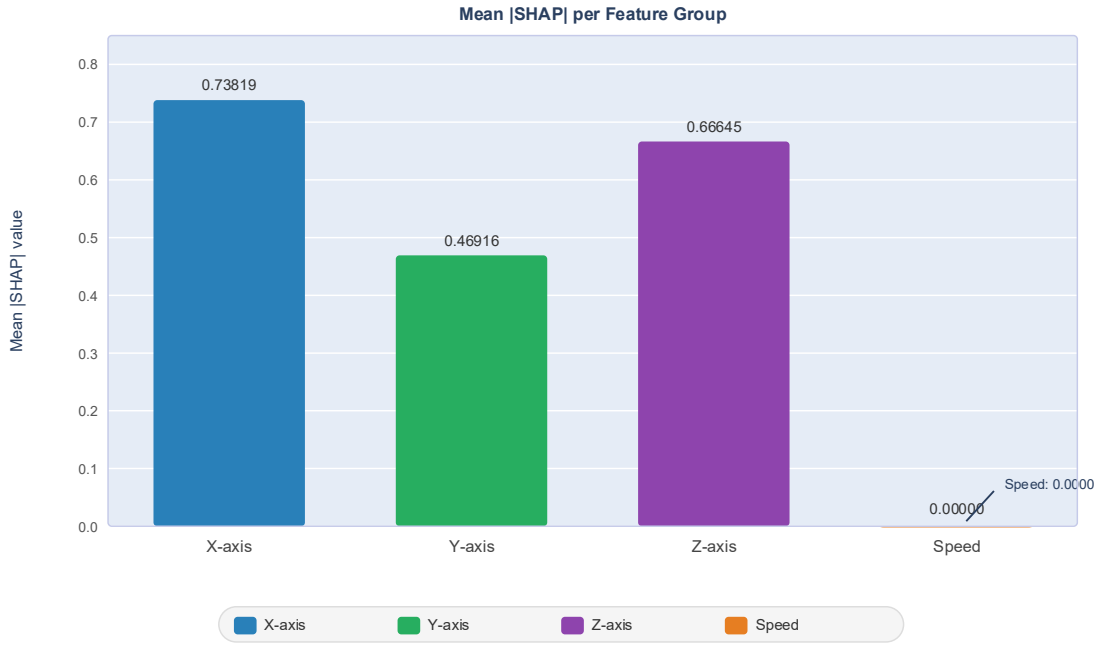


Figure 18. DeepSHAP means attribution of CNN-FFT model 3-class classification at 1500 RPM.

Figure 18 shows the mean absolute DeepSHAP attribution values at 1500 RPM across all 3 axes for the CNN-FFT model. At 1500 RPM the X-radial axis contributes 39.40% of attribution, the Y-axial axis contributes 25.04% and the Z-radial axis contributes 35.57%. The overall pattern confirms that the CNN-FFT model decision is based on bearing fault spectral signatures not on the speed-dependent artefact.

4.6.4 DeepSHAP Raw CNN

Looking at the DeepSHAP results for Raw CNN the first thing that stands out is that the X-radial axis is clearly the dominant contributor across all 10 speed folds. Table 22 shows the full breakdown of axis importance and class separability delta values. On average, the X-radial axis accounts for 50.88 per cent of the total attribution which is higher than the other 2 axes.

Table 22. DeepSHAP three-class axis analysis – Raw CNN model (selected speeds).

Speed (RPM)	X imp. (%)	Y imp. (%)	Z imp. (%)	Sep. Δ X	Sep. Δ Y	Sep. Δ Z
50	57.78	11.28	30.95	0.0309	0.0082	0.0204
100	46.26	30.33	23.41	0.0164	0.0219	0.0118
500	50.88	20.59	28.53	0.0191	0.0174	0.0135
535	49.92	29.24	20.84	0.0226	0.0131	0.015
600	57	23.52	19.48	0.0178	0.0171	0.0149
800	50.97	26.42	22.61	0.0188	0.0175	0.0137
900	50.11	27.13	22.76	0.026	0.0153	0.0161
994	50.35	30.04	19.61	0.0172	0.0157	0.0075
1200	46.48	34.74	18.78	0.0138	0.0281	0.011
1500	49.06	32.72	18.22	0.0153	0.0285	0.0162
Mean	**50.88%**	**26.60%**	**22.52%**	**0.0198**	**0.0183**	**0.014**

In Table 22 there is an interesting pattern in the separability delta column. From 50 to 800 RPM, the X-radial axis provides the highest-class separability delta. But at 1200 RPM and 1500 RPM speed points the Y-axial axis produces the highest separability delta values.

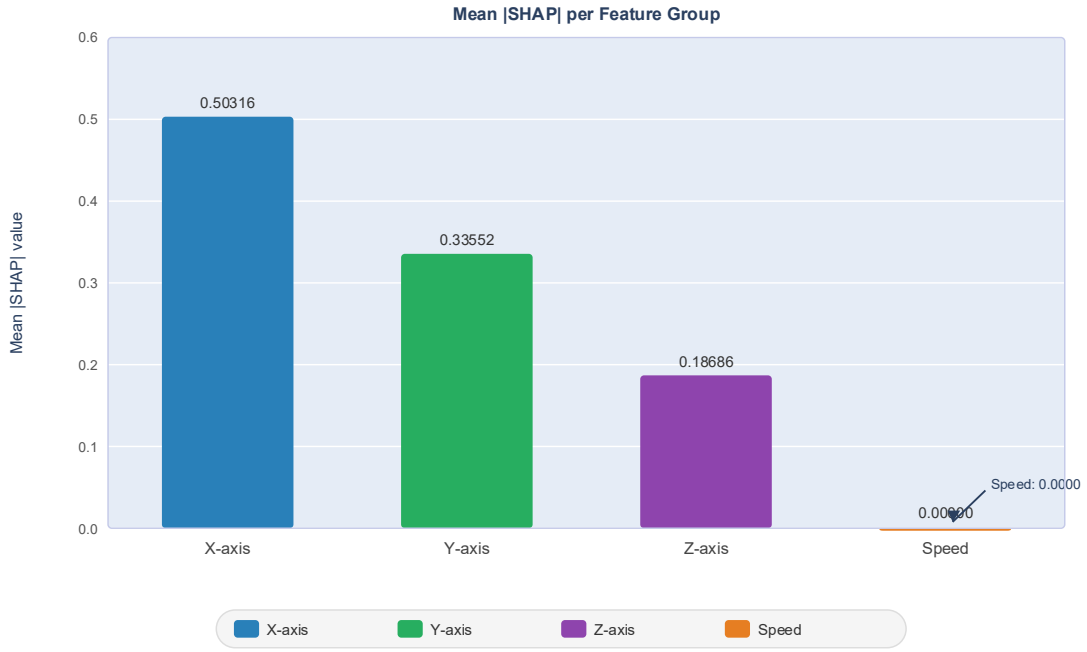


Figure 19. DEEPSHAP means attribution of the Raw CNN model at 1500 RPM.

Figure 19 displays the mean absolute DeepSHAP attribution values at 1500 rpm speed across all 3 raw-waveform axes for the Raw CNN model. This model concentrates its attributions on the X-radial channel at 49.06, the Y-axial axis is at 32.72% and the Z-radial axis at 18.22%. Attribution peaks are distributed across the full window length, not confined to a narrow temporal segment. It is indicating that the multi-scale architecture captured fault-related impulses. Speed scalars are not given importance in this model also.

4.6.5 Cross Architectural comparison

By comparing the two raw CNN and CNN FFT, it has 3 important findings:

Table 23. Cross architecture DeepSHAP axis attribution comparison (Raw CNN vs CNN-FFT).

Speed	Raw CNN dom.	CNN+FFT dom.	Agreement	Y-axis high-speed	Speed scalar
50–800	X	X or Z	Partial	Negligible	0 both
1200–1500	X	X	Yes	Y sep. both	0 both
All speeds	X dominant	X/Z competitive	X always top-2	Y>X sep. at 1200–1500	Negligible both

Table 23 has first one, whether the model is working with raw time-domain or FFT spectra, The X-radial is always dominant axis and always the top two contributors at every speed fold. It confirms that radial direction carries the dominant fault severity information for this data and bearing configuration.

The second finding is more interesting because it was noticed in both models surprisingly. The Y-axis separability delta increases at high rotational speed in both models. The CNN FFT as well as Raw CNN shows that Y separability is high at 1200 and 1500 RPM. The two models do not share any parameters. They learn their representation in completely different input domains, but both came to one conclusion about the Y-axis at high speed.

The third finding is that the speed scalar which was included as an input for both models receives negligible SHAP attribution at every speed. That means neither model is using speed as a shortcut to learn the data patterns. So, both models learned the fault representation that are inherently speed robust. This reliability is important for industrial applications.

4.7 Anomaly Detection Results

4.7.1 Isolation Forest

Table 24. isolation Forest anomaly Detection (selected speeds).

Speed	Accuracy	DR	FAR
50	71.58%	80.38%	45.91%
100	96.71%	97.54%	4.95%
500	100.00%	100.00%	0.00%
535	100.00%	100.00%	0.00%
600	100.00%	100.00%	0.00%
800	74.25%	100.00%	76.83%
900	100.00%	100.00%	0.00%
994	99.62%	100.00%	1.15%
1200	99.25%	100.00%	2.26%
1500	71.33%	100.00%	78.94%

Table 24 displays the Isolation Forest results reveal a highly speed-dependent performance pattern that can be divided into 3 distinct groups. At mid-range speeds which are 500, 535, 600 and 900 RPM, the model achieves perfect performance with 100% accuracy, 100% DR and 0% FAR. This confirms that the FFT feature space at these speeds produces good separated healthy and anomalous distributions. At 100, 994 and 1200 RPM performance remains strong with accuracy above 96%, DR above 97% and FAR below 5%. It indicates reliable fault detection with very few healthy windows. The most challenging speed points are 50, 800 and 1500 RPM where a critical tradeoff between DR and FAR and Low FAR are detected at this operating speed.

4.7.2 FFT autoencoder

Table 25.FFT autoencoder anomaly detection (selected speeds).

Speed	Accuracy	DR	FAR
50	94.69%	99.62%	15.09%
100	95.39%	99.51%	12.87%
500	100.00%	100.00%	0.00%
535	67.18%	100.00%	98.85%
600	100.00%	100.00%	0.00%
800	82.35%	100.00%	52.66%
900	100.00%	100.00%	0.00%
994	99.62%	100.00%	1.15%
1200	100.00%	100.00%	0.00%
1500	65.63%	100.00%	94.63%

Table 25 displays the FFT Autoencoder results show a very good consistent pattern across all speed folds The Detection Rate reaches almost 100% or near 99 % at every single speed point. At 50 RPM the FFT Autoencoder achieves 99.62% DR compared to 80.38% for the Isolation Forest. At 100 RPM DR reaches 99.51% with a FAR of only 12.87% indicating reliable detection with a good false alarm rate.

However, the FFT Autoencoder introduces a new challenge at several speed points which has extremely high FAR values. At 535 RPM the FAR reaches 98.85% which means almost every healthy window is incorrectly flagged as anomalous despite 100% DR. At 1500 RPM FAR reaches 94.63% and at 800 RPM 52.66%. It is producing high reconstruction errors even for healthy windows and triggering excessive false alarms.

4.7.3 Raw CNN auto encoder

Table 26.CNN auto anomaly Detection (selected speeds).

Speed	Accuracy	DR	FAR
50	45.18%	43.99%	52.45%
100	96.05%	94.09%	0.00%
500	100.00%	100.00%	0.00%
535	93.89%	90.86%	0.00%
600	100.00%	100.00%	0.00%
800	100.00%	100.00%	0.00%
900	100.00%	100.00%	0.00%
994	100.00%	100.00%	0.00%
1200	100.00%	100.00%	0.00%
1500	100.00%	100.00%	0.00%

Table 26 above presents the CNN Autoencoder anomaly detection results across all 10 representative speed points. The most important characteristic of these results is the near zero False Alarm Rate across 9 out of 10 speed points. At all speeds above 100 RPM the FAR is exactly 0.00%. It means the model correctly identifies every healthy window as normal without generating a single false alarm. This is the lowest FAR score of all three anomaly detection models at 5.25%.

The weakness of the CNN Autoencoder is concentrated entirely at 50 RPM where accuracy collapses to 45.18%, DR drops to 43.99% and FAR rises to 52.45%. It means at 50 RPM the CNN Autoencoder misses more than half of all faulty windows. This is in direct contrast to the FFT Autoencoder. It achieved 99.62% DR at 50 RPM by operating in the frequency domain.

At 100 RPM the CNN Autoencoder recovers strongly, achieving 96.05% accuracy, 94.09% DR and 0.00% FAR. From 500 RPM onwards the model achieves perfect performance across all remaining test speeds.

4.8 Overall Performance Results Summary

Classical Supervised Machine Learning Algorithms:

This Table 27 below summarizes the key findings before the detailed discussion that follows.

Table 27. Overall supervised model performance summary.

Feature / Model	Task	Accuracy	F1 Score
Kurtosis (Decision Tree)	Binary	64.88%	61.82%
	3-Class	53.61%	52.04%
Kurtosis (Random Forest)	Binary	68.64%	62.94%
	3-Class	54.46%	52.83%
RMS (Decision Tree)	Binary	95.91%	94.61%
	3-Class	91.48%	90.86%
RMS (Random Forest)	Binary	97.14%	96.28%
	3-Class	94.06%	93.67%
FFT (Random Forest)	Binary	97.33%	97.06%
	3-Class	95.31%	94.32%
FFT (Decision Tree)	Binary	91.36%	90.75%
	3-Class	89.25%	87.92%
Raw CNN	Binary	99.81%	99.79%
	3-Class	98.47%	98.53%
CNN FFT	Binary	92.28%	91.78%
	3-Class	83.87%	82.23%

After evaluating all the models across multiple feature representations and architectures, Table 4 has a clear performance hierarchy. Starting from the scalar impulses which are represented by kurtosis simply do not work well enough for this task. It has an accuracy of only 68.64% and fails to provide reliable discriminative boundaries for both tasks. Moving to RMS amplitude performs much better for the binary case and reaching up to 97.14% accuracy. However, for 3-class severity discrimination its performance starts to drop. The classical frequency-domain approach does considerably better.

FFT features with a Random Forest classifier have a highly stable baseline:

- 3-class accuracy of 95%.
- Consistent performance across speed folds.

This model is lightweight and good for edge device deployment. The deep learning models results can be summarized as follows:

1. Multi-scale Raw CNN: This is the best performing model. It achieves good binary accuracy of 99.81% and a 3-class accuracy of 98.47%. Multi-scale architecture clearly helps the model to capture faulty signatures across all speeds.
2. FFT-based CNN: This model is also strong, but evaluation metrics are slightly below when compared to Raw CNN.

The Raw CNN model is not only the most accurate but the most consistent model across all operating speeds. These are the most important criteria for deployment in industries. The Raw CNN model is good for Cloud infrastructure which has high operational capacity.

Anomaly Autoencoder:

Table 28.unsupervised anomaly detection model performance summary.

Anomaly Models	**Accuracy**	**DR**	**FAR**
CNN Autoencoder	93.51%	92.89%	5.25%
FFT Autoencoder	90.49%	99.91%	27.53%
Isolation Forest	91.27%	97.79%	21.01%

Table 28 displays all unsupervised anomaly detection summary results. On the unsupervised side, the simple isolation forest is also having good detection rate of 97.79% but false alarm is 21.01% which is difficult in industry and unnecessary maintenance time scheduling. The FFT Autoencoder catches almost everything with a detection rate of 99.91%. But that comes at the cost of the false alarm rate is 27.53% which is worser than Isolation Forest. However, The CNN Autoencoder has detection rate of 92.89% and false alarms are very low at just 5.25%. The choice of model depends on the use case of customer, for edge device Isolation Forest and for cloud analytics CNN Autoencoder are best.

5 Discussion

This chapter interprets the experimental results presented in Results section. Section 5.1 addresses each research question and section 5.2 explains the comparison table of published studies related and this thesis. Section 5.3 maps the machine learning pipeline developed in this thesis onto the ABB Detect–Predict–Recommend operational framework. Section 5.4 examines the limitations of the findings. Throughout the discussion connects empirical observations to the theoretical foundations established in Chapter 2 and identifies where the results confirm or challenge existing knowledge in the bearing fault diagnosis literature.

5.1 Addressing the Research Questions

Each of the four research questions defined in Section 1.3 are answered below. The answers are based directly on the quantitative results reported in Chapter4.They are also connected to the relevant literature reviewed in Chapter 2.

5.1.1 RQ1: Feature Representation and Speed-Robust Classification

Which feature representation gives the most reliable and accurate bearing damage classification across a full 0-1500 RPM operational speed range?

FFT features combined with Random Forest provide the best baseline results. It achieved 95.31% 3-class and 97.33% accuracy for binary classification tasks. Kurtosis fails uniformly across all speed folds. RMS breaks down on 3-class classification tasks.

The results section show that kurtosis failed uniformly across all LOSO folds. The authors Randall and Anton's (2011) research paper state that a scalar impulse statistic simply cannot discriminate between fault severity levels. RMS also collapsed on 3-class task but does better for binary classification reaching 97.14% accuracy. That is because RMS did not have spectral structure and cannot differentiate between the sharp resonant peaks

of a lightly damaged bearing and heavily damaged. That is why FFT features with Random Forest have become the best option. The 450-dimensional FFT vector achieves 95.31% for 3 -class accuracy across all folds. Adding the normalized speed scalar as a conditioning input helps the classifier separate actual bearing defect harmonics from speed-dependent mechanical resonances. This is recommended for edge deployment where computational simplicity matters.

5.1.2 RQ2: Deep Learning Generalisation to Unseen Speeds

Can a deep learning model trained on raw vibrational signals generalize to motor speeds it has never seen during training?

Yes. The Raw CNN maintained above 96% accuracy for 3-class tasks across nearly all 57 held out operational speed under LOSO evaluation including the challenging speeds like 535 RPM. The multi-scale architecture processes the waveform through 3 parallel branches at kernel sizes of 8, 32, and 64. Because of this fault signatures get captured regardless of how speed shifts the impulse interval. Different kernel sizes recognize different temporal phenomena simultaneously. This makes the model does not lose track of fault signatures even when the operating condition changes. The DeepSHAP results show that the speed scalar received negligible attribution across all folds. That means the model is not cheating by using speed as a proxy label. It is genuinely learning fault structure from the waveform itself.

5.1.3 RQ3: Axis-Dependent Attribution and Speed Effects

Do the three sensor axes contribute equally to bearing fault detection and does their contribution change at different motor speed?

No, the 3-sensor axis contribution is not equal for bearing fault detection. The X-radial axis dominates in the Raw CNN of about 50.88% of mean DeepSHAP attribution across all speed folds and as well as for CNN-FFT with X-axis dominance of 40.82%. However,

above 1200 RPM something changes in both architectures. After this speed the Y-axial axis starts providing the highest-class separability delta. So, it is suggested that Y-axial vibration carries more discriminative severity information at higher rotational speeds. The speed scalar is almost negligible which shows that CNN RAW is speed independent model.

5.1.4 RQ4: Unsupervised Anomaly Detection Across the Speed Range

Can a faulty detection system train only on healthy bearing data reliably identify damage across all motor speeds?

Unsupervised anomaly detection trained exclusively on healthy bearing data can reliably identify bearing faults across the full 0-1500 speed range. Although no single detector provides both low false alarm rate and reliable detection across the full speed range. But among all autoencoders models of this thesis the CNN Autoencoder achieves near zero false alarms at 9 of 10 speed folds and a good 92.89% detection rate. On the other hand, The FFT Autoencoder has DR at 99.91% and FAR above 94% at 1500 RPM which is operationally not suitable. The Isolation Forest has a middle ground result at 97.79% DR and 21.01% FAR which is also good option for edge devices. In all these models CNN Auto encoder has best candidate for the ABB detect stage because of balanced results.

5.2 Comparison of Related Research work

Table 29. Comparison of related research studies.

Study	Data Set	Operating Conditions	Evaluation Protocol	Fault Type	XAI	Anomaly Detec-tion	Tem-poral Span
Saha et al. (2024)	CWRU	4 loads	Transfer Learning	Bearing	No	No	<1 day
Zhang et al. (2017)	CWRU	3 loads	Cross Domain	Bearing	No	No	<1 day
Lessmeier et al. (2016)	Pader-born	4 conditions	Cross condition	Bearing	No	No	<1 day
Hendriks et al. (2022)	CWRU	4 loads	Leave bearing out	Bearing	No	No	<1 day
Present work	ABB	57	LOSO (57 folds)	Bearing Contamina-tion (3-class)	Yes	3	4 years

As shown in the comparison table 29 above, the 57-fold LOSO protocol used in this thesis is unique when compared to prior research studies. But beyond the evaluation protocol, there are three things this thesis does are the following:

- Four-year longitudinal validation on the same physical testbed.
- DeepSHAP comparison at multiple speed points.
- Direct comparison of multiple anomaly detector classes on the same variable-speed dataset.

No single prior study combines all three of those. This thesis study is about demonstrating that the results hold up under conditions that reflect industrial reality.

5.3 Positioning Within the ABB Detect-Predict-Recommend Framework

The CNN Autoencoder fulfils the Detect stage. It flags deviations from healthy behavior. When it raises an alarm, something is defective in the motor. The Raw CNN takes care of the Predict stage. It is classifying fault severity across the unseen speeds under LOSO conditions. For situations like where computational simplicity matters more such as for edge deployment, the FFT Random Forest serves as a good lightweight alternative model.

The confidence-threshold handles the Recommend stage in a way that is honestly more useful than a forced decision. When model confidence falls below 0.80, the system returns an Inspect decision instead of taking a potentially wrong diagnosis. In practice, the Inference testing results of Random Forest FFT model at 1500 RPM and 535 RPM confirmed this behavior. The binary classifier remained highly confident across all tested windows. The 3-class classification model correctly withheld the diagnosis on ambiguous severity windows, and it returned a Inspect instead of forced label.

It also found that DeepSHAP analysis plays a bigger role. By confirming that predictions are grounded in genuine fault signatures rather than speed proxies. It provides the kind of physical evidence that maintenance engineers need before they can trust a system recommendation. In a safety-critical context, that trust is not optional, it is a prerequisite for the whole pipeline to be operationally useful. (Lundberg & Lee, 2017)

5.4 Research Limitations

The following are the limitations of this thesis before drawing any broader conclusions:

- All data from metallic dust contamination only. The performance on other faults such as fatigue spalling or electric discharge damage remains unvalidated.
- The 300-sample window at 1660 Hz does not capture a full shaft rotation.so the results at low speed reflect a physical acquisition constraint as much as model behavior.
- The DeepSHAP axis attribution findings are specific to this bearing geometry and sensor placement. Without running the same analysis again, they should not be generalized to other motor configurations.

6 Conclusion

Bearing failures are the main causes of unexpected downtime in industries. Although data-driven fault diagnosis has advanced significantly, most existing research remains limited to single-session and narrow speed range laboratory benchmarks. These conditions do not reflect the variable-speed, time-varying nature of real industrial deployments .This thesis was motivated by four specific gaps in the existing literature: the absence of Leave-One-Speed-Out evaluation , the lack of longitudinal multi-year validation, the absence of cross-architecture DeepSHAP comparison at multiple speed points and the lack of side-by-side evaluation of diverse anomaly detector classes on variable-speed data. Each of these gaps is addressed in this work.

To address these gaps a comprehensive machine learning pipeline was developed and evaluated. Four supervised classification models were included in the evaluation: Decision Tree, Random Forest, CNN-FFT and a multi-scale Residual Raw CNN. Each model was evaluated on both binary and three-class classification tasks. In addition, three unsupervised anomaly detectors were evaluated alongside the supervised models. All models were tested under a 57-fold Leave-One-Speed-Out protocol which ensured that each speed point was treated as an unseen test condition during evaluation. To improve the interpretability of the results, DeepSHAP attribution analysis was applied to both CNN architectures. This allowed a direct comparison of how each architecture responds to different input features at varying motor operational speeds.

The results provide clear and consistent answers to all four research questions. For RQ1, the FFT magnitude spectrum combined with a Random Forest classifier provided the most reliable speed-robust feature representation. Kurtosis failed uniformly across speed conditions and RMS showed degraded performance on the three-class task at speed extremes. This confirms that frequency domain features carry stronger generalizable information than simple statistical time domain indicators. For RQ2, the multi-scale Residual Raw Class the highest performance of any model evaluated and reached

99.81% binary and 98.47% three-class mean accuracy across all 57 held-out folds. This demonstrates that a multi-scale temporal architecture can learn speed-invariant fault representations directly from raw waveforms, without requiring manual feature extraction. For RQ3, the three sensor axes were found to contribute unequally to fault detection. The X-radial axis consistently provided the dominant fault attribution, accounting for 50.88% of the mean DeepSHAP share in Raw CNN. Above 1200 RPM, the Y-axial axis became the primary class discriminator. This finding was replicated independently across both CNN architectures which strengthens confidence in the result. For RQ4, no single anomaly detector achieved both high detection rate and low false alarm rate across the full speed range. The CNN Autoencoder provided the best overall balance, achieving a detection rate of 92.89% with a false alarm rate of only 5.25%. This makes it the most operationally suitable detector for the ABB Detect stage.

The contributions of this thesis extend beyond the individual model results. The 57-fold LOSO evaluation protocol reported in the bearing fault diagnosis literature. The four-year longitudinal dataset and its consistent results across recording years provide evidence that the proposed models are robust to the environmental variation and cumulative wear and tear in long-term industrial deployments. The cross-architecture DeepSHAP analysis provides actionable sensor-axis prioritization guidance that can inform future hardware and feature engineering decisions. These contributions demonstrate that reliable, interpretable and speed-robust bearing health monitoring is achievable within the ABB Detect–Predict–Recommend operational framework.

The findings of this thesis have several limitations which are documented in detail in Section 5.4. The most important limitation is that the study was restricted to a single fault mechanism specifically metallic dust contamination. A physical windowing constraint at very low rotational speeds also affected the analysis. These limitations mean that further validation on additional fault types and motor configurations is necessary before the proposed pipeline can be considered broadly generalizable.

Nevertheless, within the scope of the experimental conditions studied, the results are consistent and practically meaningful. This thesis establishes a foundation for the development of bearing health monitoring systems that are both trustworthy and ready for deployment in variable-speed industrial environments.

6.1 Future Work

The following are the good directions to follow on this research:

1. FFT and RMS ensemble: Combining these two representations, which fail at different and complementary speed points, could potentially achieve reliable classification across the complete 0–1500 RPM range.
2. Temporal degradation characterization: The performance trajectory across the 4-year period needs systematic study to understand how often retraining would be required as sensor calibration and bearing surface conditions evolve.
3. Other fault mechanisms: Validation on fatigue spalling and electrical discharge damage would establish whether the FFT based and CNN based approaches generalize beyond contamination type faults.
4. Cross-motor and cross generalization: All models in this thesis were trained and evaluated on a single testbed with a fixed bearing geometry and sensor placement. Validating the proposed pipeline on motors with different bearing configurations, shaft sizes and sensor mounting positions would help whether the findings can generalize beyond this setup.

Acknowledgement:

I would like to thank ABB Oy and ABB employees for their continuous support throughout this project. Working with a global technology leader in automation and predictive maintenance was honestly one of the most valuable parts of this whole experience. This research study bridges the gap between academic theory and industrial application.

References

ABB Ability Digital Powertrain Enabling Hardware. (2026). Abb.com. https://search.abb.com/library/Download.aspx?DocumentID=9AKK108471A9286&LanguageCode=en&DocumentPartId=&Action=Launch

ABB's condition monitoring services help Mokrá cement plant save $210K, while increasing operational efficiency. (2021, July 12). News. https://new.abb.com/news/detail/80450/abbs-condition-monitoring-services-help-mokra-cement-plant-save-210k-while-increasing-operational-efficiency

Antoni, J. (2006). The spectral kurtosis: A useful tool for characterising non-stationary signals. Mechanical Systems and Signal Processing, 20(2), 282–307. https://doi.org/10.1016/j.ymssp.2004.09.001

Albrecht, P. F., Appiarius, J. C., McCoy, R. M., Owen, E. L., & Sharma, D. K. (1986). Assessment of the Reliability of Motors in Utility Applications - Updated. IEEE Transactions on Energy Conversion, EC-1(1), 39–46. https://doi.org/10.1109/TEC.1986.4765668

Althubaiti, A., Albadrani, M., & Al-Duais, F. S. (2022). Bearing fault diagnosis based on mel frequency cepstral coefficients and support vector machine. Journal of Sensors, 2022, 1–10. https://doi.org/10.1155/2022/8074847

Ayankoso, S., Dutta, A., He, Y., Gu, F., Ball, A., & Pal, S. K. (2024). Performance of vibration and current signals in the fault diagnosis of induction motors using deep learning and machine learning techniques. Structural Health Monitoring, 23(4), 2541–2562. https://doi.org/10.1177/14759217241289874

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification And Regression Trees. Routledge. https://doi.org/10.1201/9781315139470

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 1–32. https://doi.org/10.1023/A:1010933404324

Dilda, V., Mori, L., Noterdaeme, O., & Schmitz, C. (2017, August 14). Manufacturing: Analytics unleashes productivity and profitability | McKinsey. Www.mckinsey.com. https://www.mckinsey.com/capabilities/operations/our-insights/manufacturing-analytics-unleashes-productivity-and-profitability.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. https://www.deeplearningbook.org

Goyal, T., Huang, P., Sutton, F., Maag, B., & Sommer, P. (2022). SMiLe: Automated end-to-end sensing and machine learning co-design. In A. Ferscha, MC Chan, SS Kanhere, & RRV Prasad (Eds.), Proceedings of the 2022 International Conference on Embedded Wireless Systems and Networks (EWSN 2022) (pp. 12–23). Junction Publishing/ACM.https://doi.org/10.5555/3578948.3578950

Hendriks, J., Dumond, P., & Knox, D. A. (2022). Towards better benchmarking using the CWRU bearing fault dataset. Mechanical Systems and Signal Processing, 169, 108732.

Ince, T., Kiranyaz, S., Eren, L., Askar, M., & Gabbouj, M. (2016). Real-time motor fault detection by 1-D convolutional neural networks. IEEE Transactions on Industrial Electronics, 63(11), 7067–7075. https://doi.org/10.1109/TIE.2016.2582729

McFadden, P. D., & Smith, J. D. (1984). Vibration monitoring of rolling element bearings by the high-frequency resonance technique — a review. Tribology International, 17(1), 3–10. https://doi.org/10.1016/0301-679X(84)90076-8

Mordor Intelligence. (2026). Predictive maintenance market size, share and trends analysis report 2026–2031. https://www.mordorintelligence.com/industry-reports/predictive-maintenance-market

Lei, Y., He, Z., & Zi, Y. (2009). Application of an intelligent classification method to mechanical fault diagnosis. Expert Systems with Applications, 36(6), 9941–9948. https://doi.org/10.1016/j.eswa.2009.01.065

Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing, 138, 106587. https://doi.org/10.1016/j.ymssp.2019.106587

Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016, July). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In PHM society European conference (Vol. 3, No. 1).

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, 413–422. https://doi.org/10.1109/ICDM.2008.17

Loparo, K. A. (2012). Bearings vibration dataset. Case Western Reserve University. https://csegroups.case.edu/bearingdatacenter/home

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4766–4777.

Principi, E., Roli, A., Squartini, S., & Piazza, F. (2019). Unsupervised electric motor fault detection by using deep autoencoders. IEEE/CAA Journal of Automatica Sinica, 6(2), 441–451. https://doi.org/10.1109/JAS.2019.1911393

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—A tutorial. Mechanical Systems and Signal Processing, 25(2), 485–520. https://doi.org/10.1016/j.ymssp.2010.07.017

Saha, D., Hoque, M. E., & Chowdhury, M. E. H. (2024). Enhancing bearing fault diagnosis using transfer learning and random forest classification. IEEE Access, 12, 5986–6000. https://doi.org/10.1109/ACCESS.2023.3347345

Sakurada, M., & Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. Proceedings of the MLSDA 2014 Workshop on Machine Learning for Sensory Data Analysis, 4–11. https://doi.org/10.1145/2689746.2689747

Sanakkayala, D. C., Varadarajan, V., Kumar, N., Karan, Soni, G., Kamat, P., Kumar, S., Patil, S., & Kotecha, K. (2022). Explainable AI for Bearing Fault Prognosis Using Deep Learning Techniques. Micromachines, 13(9), 1471. https://doi.org/10.3390/mi13091471

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sustainability | ABB. (2025). ABB Group. https://www.abb.com/global/en/company/sustainability

Tiboni, M., Remino, C., Bussola, R., & Amici, C. (2022). A review on vibration-based condition monitoring of rotating machinery. Applied Sciences, 12(3), 972. https://doi.org/10.3390/app12030972

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

Value of Reliability: ABB Survey Report 2023 Industry's perspective on maintenance and reliability LET'S GO →.(n.d.). https://new.abb.com/docs/librariesprovider19/default-document-library/abb_survey-report-2023.pdf

Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2017). A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors, 17(2), 425. https://doi.org/10.3390/s17020425a

Zhong, Z., Liu, H., Mao, W., Xie, X., & Cui, Y. (2023). Rolling bearing fault diagnosis across operating conditions based on unsupervised domain adaptation. Lubricants, 11(9), 383. https://doi.org/10.3390/lubricants11090383