Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies

FIIT-182905-102927

Bc. Miroslav Hájek

Machinery Vibrodiagnostics with the Industrial Internet of Things

Master's Thesis

Thesis Supervisor: Ing. Marcel Baláž, PhD.

May 2024

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies

Reg. No. FIIT-182905-102927

Bc. Miroslav Hájek

Machinery Vibrodiagnostics with the Industrial Internet of Things

Master's Thesis

Study programme:	Intelligent Software Systems
Study field:	Computer Science
Training workplace:	Institute of Computer Engineering and Applied Informatics
Thesis supervisor:	Ing. Marcel Baláž, PhD.
Departmental advisor:	Ing. Jakub Findura
Consultant:	Ing. Lukáš Doubravský

May 2024

Slovak University of Technology in Bratislava
Institute of Computer Engineering and Applied Informatics

Faculty of Informatics and Information Technologies
Academic year: 2023/2024
Reg. No.: FIIT-182905-102927

MASTER THESIS TOPIC

Student:	**Bc. Miroslav Hájek**
Student's ID:	102927
Study programme:	Intelligent Software Systems
Study field:	Computer Science
Thesis supervisor:	Ing. Marcel Baláž, PhD.
Head of department:	Ing. Katarína Jelemenská, PhD.
Consultant:	Ing. Jakub Findura

Topic: **Machinery vibrodiagnostics with the industrial internet of things**

Language of thesis: English

Specification of Assignment:

Monitorovanie prevádzkového stavu rotačných strojov za účelom včasného odhalenia poškodení je dôležité pre plynulý priebeh priemyselných procesov bez náhleho zlyhania kľúčového technického vybavenia. Nadmerné vibrácie alebo graduálna či náhla zmena ich charakteru sú spoľahlivými indikátormi opotrebenia dielcov. V mnohých prípadoch bývajú zavedené iba pravidelné pôchodzkové merania s následným vyhodnotením časových a frekvenčných priebehov kvalifikovaným personálom. Kontinuálna diagnostika a prediktívna údržba rozširujúca sa so zariadeniami IIoT spôsobuje enormný nárast objemu zaznamenaných dát. Sledovanie výchyliek operátorom a manuálna identifikácia súčiastok vyžadujúcich údržbu v celom závode sa tak stáva prakticky nerealizovateľná. Preskúmajte spôsoby zisťovania bežných poškodení strojov z vibračných signálov a analyzujte algoritmy na redukciu množstva posielaných dát zo senzorov vzhľadom na osobitosti aplikačnej domény. Navrhnite reprezentáciu údajov na základe typických čŕt signálu, ktorá zníži výpočtové nároky na zvyšok komunikačného reťazca. Zvolený spôsob predspracovania má zároveň umožniť diagnostiku poškodení zvoleného stroja. Implementuje vaše riešenie s ohľadom na možné nasadenie na prostriedkami limitovanú senzorovú jednotku. Následne posúďte efektívnosť, porovnajte dosiahnuté presnosti diagnostiky a verifikujte voči zaužívaným postupom.

Deadline for submission of Master thesis:	17. 05. 2024
Approval of assignment of Master thesis:	25. 04. 2024
Assignment of Master thesis approved by:	prof. Ing. Vanda Benešová, PhD. – study programme supervisor

Declaration of Honour

I hereby declare on my honour that I wrote this thesis independently under the supervision of Dr. Marcel Baláž, after consultations and with use of cited literature.

Bratislava, 17 May 2024 .....

Bc. Miroslav Hájek

Wir müssen wissen.

Wir werden wissen.

— David Hilbert

Poďakovanie

Napriek tomu, že tento text je len kvapka v mori záverečných prác, pre mňa bol proces tvorby omnoho viac než zachytávajú slová na stránkach. Srdečná vďaka patrí školiteľovi Ing. Marcelovi Balážovi, PhD. a konzultantovi Ing. Lukášovi Doubravskému z firmy R-DAS, s.r.o. Obaja boli otvorení mojim všemožným nápadom a neochvejne ma podporovali v ich realizácii aj cez mnohé ťažkosti. Za praktický expertný pohľad na vibrodiagnostiku vďačím prof. Ing. Stanislavovi Žiaranovi, CSc. a Ing. Ondrejovi Chlebovi, PhD. zo Strojníckej fakulty STU. Cením si ústretovosť Bratislavskej vodárenskej spoločnosti, a.s., konkrétne Ing. Petra Csóku a Ing. Petra Kmeťka, v sprístupnení čerpadiel na merania a súvisiacich podkladov. Rovnako ďakujem za ochotu firme VNET, a.s., a síce Michalovi Országovi a Mgr. Vladimírovi Kupčovi za odporúčania možných prvkov na merania v klimatizáciach pre dátové centrá a sprístupnenie kompresorov. Prácu by som rád venoval rodine a kolegom-kamarátom, ktorí stáli pri mne na „šialenej akademickej dráhe" a podnetné diskusie s nimi prispeli aj k môjmu pohľadu na odborné problémy. Nuž a v konečnom dôsledku som mal niekedy len viac šťastia ako rozumu.

Annotation

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies

Degree course:	Intelligent Software Systems
Author:	Miroslav Hájek
Master's Thesis:	Machinery Vibrodiagnostics with the Industrial Internet of Things
Thesis supervisor:	Dr. Marcel Baláž
Departmental advisor:	Jakub Findura
Consultant:	Lukáš Doubravský

May 2024

Master's thesis focuses on condition monitoring and predictive maintenance of rotating machinery by employing feature discovery and statistical classification.

Trend indicators are derived from vibration signals to form summarizing numerical attributes. The similarity of features to a dependent variable that describes fault type is assessed by supervised feature selection metrics individually and in the ensemble. The accuracy of fault classification using the k-nearest neighbour algorithm is evaluated on the best feature subsets. The goal is to conserve the data rates of the monitoring solution. The incremental learning models are compared to their batch counterparts when true labels are delayed or missing. An extendible software toolkit of vibrodiagnostics is implemented as a Python package.

Classification metrics are evaluated on the MaFaulDa dataset, and lossy compression ratios are expressed. Our accelerometer data logger collects a novel dataset of machinery behavior for compressors and pumps in the industrial environment. The contribution of this work lies in identifying the viability of using a few quantities to describe not just the presence of the machine defect but also its cause.

Anotácia

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Študijný program:	Inteligentné softvérové systémy
Autor:	Bc. Miroslav Hájek
Diplomová práca:	Vibrodiagnostika strojov s priemyselným internetom vecí
Vedúci diplomovej práce:	Ing. Marcel Baláž, PhD.
Pedagogický vedúci:	Ing. Jakub Findura
Konzultant:	Ing. Lukáš Doubravský

Máj 2024

Diplomová práca sa zaoberá monitorovaním stavu rotačných strojov a ich prediktívnou údržbou prostredníctvom techník výberu atribútov a štatistickej klasifikácie.

Z vibračných signálov sú odvodené trendové indikátory, ktoré tvoria sumarizujúce číselné atribúty. Podobnosť atribútov so cieľovou premennou popisujúcou typ poruchy sa posudzuje individuálne a súborovo metrikami výberu atribútov v učení s učiteľom. Presnosť klasifikácie porúch strojov algoritmom k-najbližších susedov sa vyhodnocuje na podmnožinách najlepších atribútov. Tým sa znižujú požiadavky na objem prenosu dát v monitorovacom riešení. Modely postupného učenia sú porovnané s učením v dávkach, v situáciach kedy je expertné označenie prevádzkového stavu oneskorené alebo chýbajúce. Rozšíriteľná sada funkcií pre vibrodiagnostiku je implementovaná ako balík v jazyku Python.

Metriky klasifikácie sú vyhodnotené v scenároch na dátovej sade MaFaulDa a tiež sú vyjadrené stratové kompresné pomery. Nami vytvorený data logger nazhromaždil novú dátovú sadu z kompresorov a čerpadiel v priemyselnom prostredí. Prínos tejto práce spočíva v overení možnosti použitia iba pár veličín na opis nielen prítomnosti poruchy, ale aj jej príčiny.

Contents

1 Introduction **1**

2 Problem analysis **3**

2.1 Condition monitoring . . . 3

2.1.1 Maintenance strategies . . . 3

2.1.2 Vibration fault types . . . 6

2.1.3 Technical standards . . . 9

2.2 Signal preprocessing . . . 12

2.2.1 Detrending . . . 13

2.2.2 Time synchronous averaging . . . 14

2.3 Feature extraction . . . 14

2.3.1 Time-domain features . . . 14

2.3.2 Frequency-domain features . . . 16

2.3.3 Features in harmonic frequencies . . . 17

2.3.4 Time-frequency domain features . . . 18

2.3.5 Wavelet domain features . . . 19

2.4 Feature transformations . . . 23

2.4.1 Feature normalization . . . 23

2.4.2 Principal components analysis . . . 24

2.5 Feature selection . . . 24

2.5.1 Filtering methods . . . 25

2.5.2 Feature importance ranking . . . 25

2.6 Diagnostics techniques . . . 27

2.6.1 Novelty detection . . . 27

2.6.2 Classification . . . 28

2.6.3 Incremental learning . . . 31

2.7 Datasets of machinery faults . . . 32

2.7.1 Machinery Fault Database . . . 32
2.7.2 CWRU bearings dataset . . . 34
2.7.3 Unbalance of the rotating shaft . . . 35

3 Design **37**
3.1 Research questions . . . 37
3.2 Machine learning pipeline . . . 38
3.3 Exploratory data analysis of MaFaulDa . . . 43
3.4 Accelerometer data logger . . . 48
3.5 Industrial equipment . . . 50
3.6 Data collection methodology . . . 52
3.7 Data volume savings . . . 54

4 Implementation **57**
4.1 Data analysis . . . 57
4.2 Firmware . . . 58

5 Evaluation **61**
5.1 Fault classifiction in MaFaulDa . . . 61
5.1.1 Complete feature sets . . . 61
5.1.2 Feature subset combinations . . . 63
5.1.3 Feature selection techniques . . . 64
5.1.4 Incremental learning . . . 68
5.2 Industrial dataset analysis . . . 72
5.2.1 Data logger verification . . . 72
5.2.2 Signal waveforms . . . 73

6 Conclusion **77**
6.1 Future work . . . 79

7 Resumé **81**
7.1 Úvod . . . 81
7.2 Sledovanie prevádzkového stavu . . . 81
7.3 Extrakcia a výber atribútov . . . 82

7.4 Diagnostické prístupy 83
7.5 Výskumné otázky 84
7.6 Návrh spracovania pre MaFaulDa 84
7.7 Zber vibrácií v priemysle 85
7.8 Vyhodnotenie presnosti diagnostiky 86
7.9 Rozbor dátovej sady z priemyslu 88
7.10 Záver 88

Bibliography **91**

A Technical documentation

B User Guide

C IIT.SRC 2024 Paper

D Work plan

E Digital medium

List of Figures

2.1 P-F curve represents the evolution of the asset's health [7] 5
2.2 Bathtub curve [1] 5
2.3 Complex machinery vibrations [6] 6
2.4 Bearing parts [13] 8
2.5 The transducer linear response and resonance in tolerance intervals [11] 12
2.6 Transfer function of 1st order DC blocker filters [15] 13
2.7 Comparison of time-frequency transform spectrograms [28] 21
2.8 Dyadic filter banks for discrete wavelet transform [17] 22
2.9 DenStream [44] 29
2.10 Nearest neighbours classification algorithm [49] 30
2.11 Incremental learning deployment architecture [57] 32
2.12 Machinery fault simulator for MaFaulDa 33
2.13 CWRU machine apparatus 34
2.14 Motor driving shaft in unbalance measurement [66] 35

3.1 Machine learning pipeline for MaFaulDa dataset 39
3.2 Vibrations from inner bearing (A) for every fault class with the highest fault severity at 2500 rpm 44
3.3 The value ranges of attributes for bearing A, depending on the number of directions that feature is aggregated out of 45
3.4 Pearson correlations of variables from inner bearing (A) in feature sets aggregated out of 3 axes 45
3.5 Correlation of features to rpm over all experiments 46
3.6 PCA of complete feature set into two principal components (bearing A and three axes 47
3.7 PCA loading plots for bearing A 48
3.8 Data logger hardware block diagram 49

3.9 Activity diagram of data logger firmware functions 50
3.10 Frequency spectrum of audio recorded in close proximity to the side of the standing fan 51
3.11 Machines dedicated for vibration measurements 52
3.12 KSB Guard cloud monitoring for pumps 53

4.1 Accelerometer Data Logger 58
4.2 Timing issue of writing samples to SD card captured on oscilloscope . 59
4.3 Data logger on the machines in measurement positions 60

5.1 Confusion matrix for complete sets of features 62
5.2 Accuracy on complete feature sets depending on the k-value 62
5.3 Model accuracy distribution for bearing A and three axis features . . 64
5.4 Model accuracy distribution from bearing A and three axis features after relabeling for high severity faults 65
5.5 Model accuracy statistical distributions with feature selection methods for three predictors (k = 5) 68
5.6 Quality of choice for feature selection methods 69
5.7 Three features in both domains chosen by rank product with prediction accuracy in brackets. Relabeled dataset is marked with (s) . . . 69
5.8 Ordering of faults in dataset according to relative severity levels . . . 69
5.9 Tumbling window of lengths 1, 10, 100 during incremental learning . 71
5.10 Omission of labels during incremental learning with tumbling window of length 10 and gaps of size 0, 10, and 50 samples 71
5.11 Vibrations from the back of a standing fan in the radial direction . . 73
5.12 Wideband frequency waveform of machinery vibartions in Pump dataset 73
5.13 Characteric bearing frequencies of pumps 74
5.14 Vibration rms levels for a period over a year from KSB Guard 75
5.15 Machinery vibration spectrograms (f_s = 26.8 kHz, w = 8 kS) 75
5.16 Vibrations of water pumps during switch over (f_s = 26.8 kHz) 76
5.17 The attribute value ranges of Pump dataset 76

List of Tables

2.1 Expert observed likely vibration causes (based on [6, 5, 11]) 7
2.2 Characteristic frequencies of bearings 8
2.3 ISO 20816 vibration severity chart with typical magnitudes [14] . . . 10
2.4 Time-domain features 15
2.5 Frequency-domain features 16
2.6 Correlation coefficients 26
2.7 Distance metrics for k-NN 30
2.8 MaFaulDa description of columns 33
2.9 CWRU dataset description of columns 35
2.10 "Unbalance on the rotating shaft" dataset description of columns . . 36

3.1 Number of observation in MaFaulDa split by class label according to source bearing 43
3.2 Hardware parameters of accelerometer data logger 49
3.3 Data collection schedule 53

5.1 Feature selection method accuracy and percentile within accuracy distribution of all three member subsets. (bearing = A, dimension = 3, k=5) 66
5.2 Feature selection method accuracy and percentile within accuracy distribution of all three member subsets. (severity, bearing = A, dimension = 3, k=5) 67
5.3 Feature selection methods evaluated in summary over all experimental conditions 67
5.4 The experimental scenarios in which the method is found to be the best 67
5.5 Bearing characteristic harmonic frequencies of pumps and their motors 74

List of Equations

2.0 DC blocker IIR filter of 1st order . . . 13
2.1 Time synchronous averaging . . . 14
2.2 MMS algorithm: Local maxima identification equality . . . 17
2.3 MMS algorithm: Local minima identification equality . . . 18
2.4 Harmonic series search criterion . . . 18
2.5 Teager–Kaiser energy operator . . . 19
2.6 Transient-extracting transform . . . 19
2.7 Mother wavelet function . . . 19
2.8 Continuous Wavelet transform . . . 20
2.9 Representation of components for Synchrosqueezing Wavelet transform . 20
2.10 Wavelet packet coefficient . . . 22
2.11 Feature euclidean norm . . . 23
2.12 Min-max feature scaling . . . 24
2.13 Feature standardization . . . 24
2.14 Singular Value Decomposition . . . 24
2.15 Pearson correlation coefficient . . . 26
2.16 Fisher score . . . 26
2.17 Mutual information . . . 27
3.1 Lossy compression ratio with features . . . 55

List of Abbreviations

AE Acoustic emission

AM-FM Amplitude modulation and frequency modulation

CbM Condition-based maintenance

CIC Cascaded-integrator-comb

CWT Continuous Wavelet transform

DC Direct current signal component

DWT Discrete Wavelet Transform

EMD Empirical mode decomposition

FFT Fast Fourier transform

FIR Finite impulse response

IIR Infinite impulse response

IMF Intrinsic mode function

IoT Internet of Things

k-NN k-nearest neighbours algorithm

kSpS kilo-samples per second

MaFaulDa Machinery Fault Database

MEMS Micro-Electro-Mechanical Systems

MIMOSA Machinery Information Management Open System Alliance

MSE Mean square error

P-F curve Curve of machine degradation from potential failure to failure

p-p Peak-to-peak distance of the signal waveform

PCA Principal Component Analysis

PC Principal component

rms Root mean square

rpm Revolutions per minute

RUL Remaining useful life

SES Squared envelope spectrum

SNR Signal-to-noise ratio

SST Synchrosqueezing Wavelet transform

STFT Short-time Fourier transform

TET Transient-extracting transform

TKEO Teager-Kaiser energy operator

TSA Time synchronous averaging

WPD Wavelet Packet Decomposition

WT Wavelet transform

1 Introduction

The industry is experiencing a shift in traditional asset operational status evaluation and utilization. The rise of Industry 4.0 means greater automation and robotization of the production halls to achieve optimal usage of available resources. The secondary aspect in the enterprises' endeavour, but not less important, is to keep track of the equipment's wear and tear. The corrective action, be it repair or replacement, should be done on time in response to the key indicators.

The goal is to preserve required safety and production efficiency while extending the useful life of rotating mechanical parts. In the factories and logistics where this equipment is vital, there is a rising interest in the ability to watch the machine's health status in real time. Proactive fault diagnosis is imperative to initiate a repair without adding unnecessary costs.

Vibrations are a non-intrusive way to sense and record eventually fatal deficiencies right at the onset. The experts use it to distinguish faulty states and to identify the malfunction's root causes. The precautions leading to regular machinery checkups are already in place in critical circumstances, as is the case for large turbines in power plants. The monitoring solution has to be sufficiently independent, reliable, and accurate to achieve wider acceptance and spread.

The main issue to consider in large-scale machinery monitoring with vibrations is the presence of many uninformative streams of samples not directly usable for the production line operator. The dashboard must aggregate these flows into trend variables with severity levels categorized based on industrial standards. The majority of signals are viewed once at the maximum. Therefore, to store or even transmit them from the edge device in its entirety would be wasteful. The complex overview of the mechanical equipment status is attainable only when agent devices and sensors are cheap enough with a long lifespan powered out of the battery pack. The devices should preferably also remain physically small to reduce the additional clutter in the factory.

Attempted machine learning and deep learning approaches have the crucial impediment that the construction of every single machine is unique to some extent because of tolerances and variable load during regular operation. The model has to be trained for the target environment to achieve the ideal performance. In addition, failures are relatively rare events that usually occur several months apart. In these circumstances, it is difficult to obtain a large enough sample of fault events. Novelty detection is a technique that can be applied in this case.

The master's thesis opens with a chapter on problem analysis 2 where the explanation is given of mechanical maintenance approaches and industry standards on common fault identification. Vibration monitoring starts with a section on preprocessing that summarizes digital filters. The process of transforming raw samples into features and their meaningful selection is covered in automatic fault pattern recognition with machine learning models trained in offline and online contexts. An overview follows of the most common machinery fault databases.

Chapter 3 about solution design poses research questions and establishes goals. The statistical properties and labeling procedure of the MaFaulDa dataset are described on the original samples as well as on extracted attributes. The data collection procedure and accelerometer data logger for industrial water pumps and air compressor is outlined. Data reduction rates with feature selection are derived based on the MaFaulDa dataset. Chapter 4 describes software libraries and tools used in conducting model evaluation and data logger firmware implementation.

The chapter 5 on evaluation validates the classification accuracy of feature sets from MaFaulDa and strategies for their subset choice in the k-nearest neighbour model. The vibration signals in the realistic environment are also analysed and recommendations for practical application and further study are discussed.

2 Problem analysis

In the problem analysis chapter we explore the feature engineering methods and machine learning algorithms for fault diagnostics. The basis we build upon is the domain knowledge of mechanical engineers in vibration signal measurement and its evaluation.

2.1 Condition monitoring

All rotating machinery eventually fails because of the long-term strain on the individual parts, inadequate workmanship, installation, or operational procedures. In the end, these factors cause the equipment not to fulfill its intended functionality. Many instrumentation methods are practiced to reveal evolving faults: vibration and acoustic noise monitoring, electric supply line measurements, thermography, oil and particle analysis, ultrasonic testing, etc. Vibration signals are the preferred tool for rotating machinery monitoring [1].

The defect needs to be repaired or replaced, preferably without significant production downtime, further damage to the other attached elements, or any endangerment of the responsible personnel. The maintenance strategies are chosen according to the machine's importance as a result of potential failure's impact on the system. The guide to set appropriate maintenance procedures is outlined in the IEC 60706-2 standard and involves reliability-centered maintenance analysis [2].

2.1.1 Maintenance strategies

There are three different approaches to maintenance across the industry: **reactive, preventive, and predictive** [3]. In general, the more sophisticated methods are beneficial in a high-stakes environment. The unexpected machine shutdown can have a negative economic impact on the enterprise which results in decreased product

quality and demands for spare parts to be ready in the supply inventory at all times. In certain situations, it suffices to utilize a simpler maintenance program, but predictive maintenance gains attraction in Industry 4.0 to optimize usage of assets [4].

Reactive maintenance allows machinery to run until a complete failure which is the most inappropriate way to maintain the production line, but it is straightforward enough. It requires a large stock of replacement parts on-site. Any breakage inflicts crisis management upon the plant [3]. On-demand repairs are justified when short downtime is acceptable, full and swift replacement of a broken machine with a backup is possible, or there is a negligible threat to the surrounding environment from failure [5].

Preventive maintenance is performed before any issue is detected. Maintenance occurs at regular intervals derived from a predetermined period in the calendar or expected machine running time and its mean time to failure (MTTF). The schedule is crucial but can result in components being replaced in good condition, creating waste. The parts can occasionally stay in operation too long, and the machine breaks as a result. Conservative planning is usually the norm to keep the machines in perfect shape. Therefore, more frequent interventions are required [1].

Predictive maintenance (PdM) known also as condition-based maintenance (CbM), improves the predictability of reactive maintenance and eliminates waste in overall resource utilization of cautious prevention. The machine downtime is scheduled after the detection of unhealthy trends in fault monitoring with sensors and the identification of troublesome components.

A measurable decrease in effectiveness allows us to order necessary parts in advance and organize repairs of several machines at a convenient time. The missed detection leads to increased costs compared to previous methods and raises the expectation that faults are distinguishable among themselves [6].

The *P-F curve* is a widespread representation of equipment's degradation over time based on historical records (Fig. 2.1). Corrective action should be taken between the event of potential failure (P), when the fault detection is activated, and functional failure point (F) in the P-F interval [8]. These division points are not set precisely but have a statistical distribution.

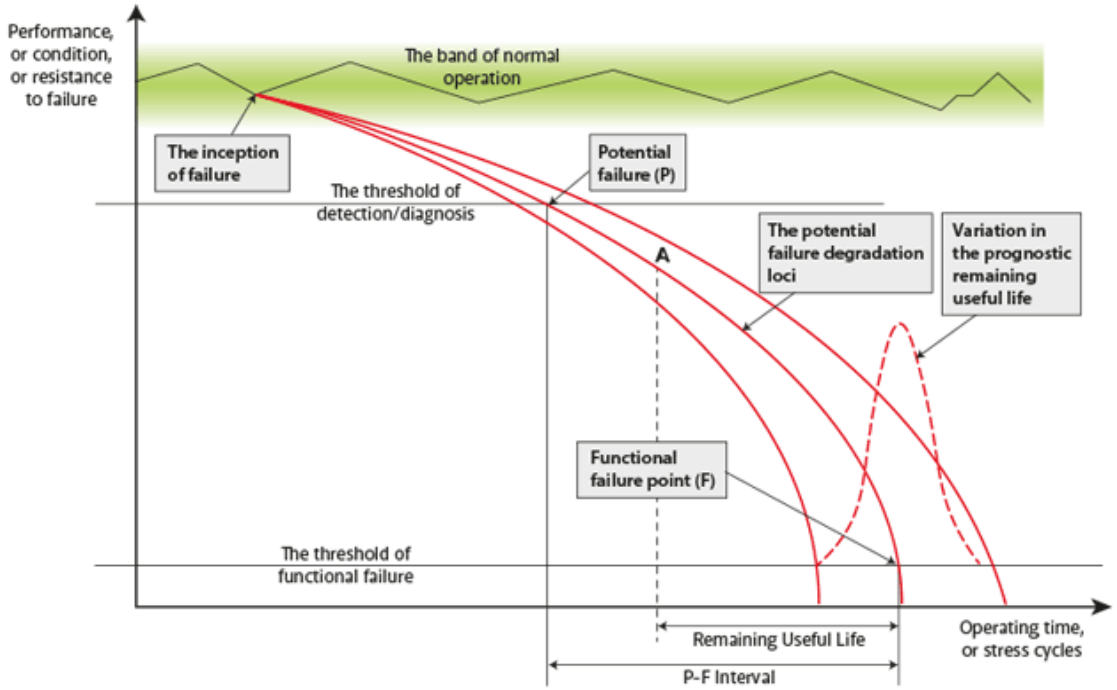


Figure 2.1: P-F curve represents the evolution of the asset's health [7]

The *Remaining Useful Life* (RUL) of the specific running machine in the given instance can be merely estimated analytically, with the survival probabilities of the individual components, based on the model of the "run-to-failure" history and usage parameters [9]. Predictive condition monitoring aims to extend the lifespan to the maximum by predicting expected RUL.

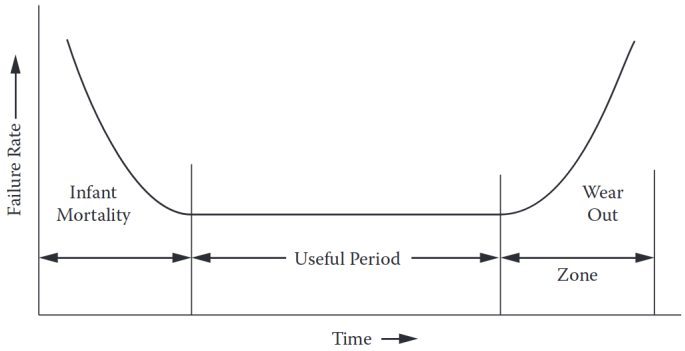


Figure 2.2: Bathtub curve [1]

A high failure rate is present at the worn-out stage when the parts are fatigued or corroded. It can also be registered in the early stages soon after assembly. Manufacturing or material defects, inadequate installation, or improper start-up procedures are all suspected causes. During the stable middle phase, malfunction can occur

after the machine's excessive overload. The time plot to failure rate is known as the bathtub curve (Fig. 2.2).

2.1.2 Vibration fault types

Mechanical problems during machinery operation bring about vibrations in many instances. Therefore, vibroacoustic diagnostics is considered as one of the most important methods in early component fault identification [5].

The cause of vibration comes out of the changing force in its magnitude or direction. The most emerging defects can be encompassed by explaining the deficiencies of the mechanical structure. These defects are broadly categorized as **unbalance, misalignment, looseness, eccentricity, deformation, crack, and influence of the external force** (friction) [6]. It is important to stress that our concern are not the underlying deformities in mechanical parts but the correct fault classification based on the signal waveform.

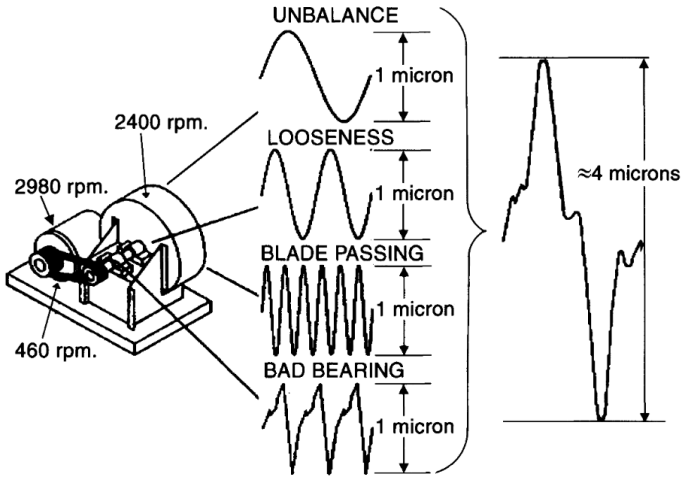


Figure 2.3: Complex machinery vibrations [6]

Rotating machine disorders exhibit frequency signatures at various ranges in the frequency spectrum with supplementary symptoms carried in phase signal. Most of the occurring faults can be tied to the main rotational speed of the component under investigation (Fig. 2.3) [6]. Imbalance, misalignment, and looseness normally appear at frequencies up to 300 Hz. These low-frequency faults are associated with the movement of the shaft and primarily coincide with revolution speed and its harmonics. Bearing and gearbox defects in the late stages of development, show up

in the range between 300 Hz and 1 kHz. Higher frequencies, measured traditionally to a limit of 10 kHz, help notice the faults in bearings even sooner [10].

One of the methods vibration experts utilize in the identification of the damaged part, according to the frequency spectrum, is **order analysis**. The excessive peaks in harmonic frequencies are of interest. Harmonics are integer multiples of fundamental frequency (1x rpm) (Tab. 2.1):

Frequency content		**Likely reason**	**Other causes**
Synchronous	1 x rpm	Imbalance	Eccentric journals Bent shaft / Misalignment (high axial vibration) Bad belt (if rpm of belt)
	2 x rpm	Looseness	Misalignment (high axial vibration) Cracked rotor Bad belt (if rpm of belt)
	3 x rpm	Misalignment	and axial looseness
	Many x rpm	Bad gears Severe looseness	Gear teeth x rpm Fan blade count x rpm
Sub-synchronous	<1 x rpm	Oil whirl	Bad drive belt Background Resonance
Non-synchronous	-	Electrical problems (x 50 Hz) Reciprocating forces Aerodynamic forces Bad antifriction bearings	Rubbing

Table 2.1: Expert observed likely vibration causes (based on [6, 5, 11])

Because of inherent tolerances in machine manufacturing and assembly, the rotational frequency always manifests itself, even in baseline signature [6, 11]. In the most likely scenario, some faults appear as compared to rotational frequency solely in **synchronous, subsynchronous, or non-synchronous components**. The defects can occur in a predictable combination of the ones mentioned. Other common patterns experts look for are modulation sidebands typical for bearings and gears extractable with cepstrum analysis [5]. Procedures relying on elimination narrow down unrelated causes effectively.

The mechanical properties of bearings are instrumental in identifying the characteristic frequencies of their components [1, 5]. Bearings are composed of outer

and inner circular races between which the rolling elements are held inside the cage to avoid mutual contact (Fig.2.4). The developing scratches or pits can appear as increases in amplitude around characteristic frequencies and their harmonics but sometimes this is not enough for accurate diagnostics [12].

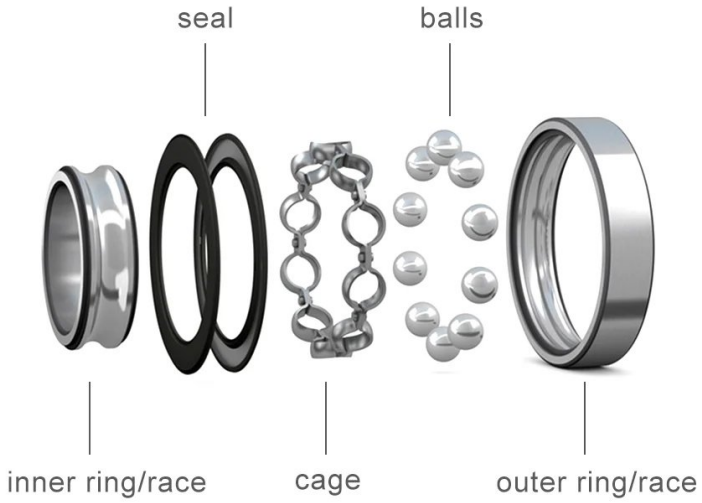


Figure 2.4: Bearing parts [13]

The table 2.2 provides formulas for rolling bearing defect frequencies at rotational speed f_s where n is a number of rolling elements, d is the average of outer and inner ring diameters, and contact angle β. Because the frequencies of components depend on shaft rotational speed the values are supplied as its multiplication factors alongside datasets.

Feature	**Equation**
Outer race fault frequency	$BPFO = f_s \cdot \frac{n}{2} \cdot \left(1 - \frac{d}{D}\cos\beta\right)$
Inner race fault frequency	$BPFI = f_s \cdot \frac{n}{2} \cdot \left(1 + \frac{d}{D}\cos\beta\right)$
Cage rotation frequency	$FTF = f_s \cdot \frac{1}{2} \cdot \left(1 - \frac{d}{D}\cos\beta\right)$
Ball defect frequencies	$BSF = f_s \cdot \frac{D}{d} \cdot \left(1 + \left(\frac{D}{d}\cos\beta\right)^2\right)$

Table 2.2: Characteristic frequencies of bearings

2.1.3 Technical standards

Vibration-based condition monitoring practices adopted in the factory's predictive maintenance management must comply with normative guidelines formalized in ISO international standards. Standards are concerned with each step in the process that originates with transducer placements and data acquisition. They prescribe conventions for setting fault severity levels and provide empirically observed vibration characteristics of common defects. Relevant standards for IoT diagnostics systems are *ISO 20816* (updated from ISO 10816) and *ISO 13373*.

ISO 20816-1:2016 establishes the approaches of vibration measurement and evaluation on non-rotating housing of machinery parts [14]. The measurement units are agreed upon for kinematic quantities of vibrations. Acceleration is measured in meters per second squared (m/s^2), velocity in millimeters per second (mm/s), and displacement in micrometers (μm). It is customary to evaluate broad-band vibration velocity in terms of root mean square value (rms), as it is related to signal energy. No simple direct relationship is expressible among these physical quantities, except in stationary signals.

The vibration severity is the maximum magnitude measured in two radial directions (horizontal, and vertical) or supplemented with a third direction along the shaft on the axial axis. Multiple measurement locations, i.e. on different bearings or couplings, should be assessed independently.

Criteria introduced to judge vibration severity are absolute vibration magnitude, change in the magnitude vector, and rate of change. In terms of maximal magnitudes, the machines of varying sizes are split into four severity zones defined in the chart (Tab. 2.3). The values in this table serve as guidelines for realistic requirements between machine manufacturers and their customers.

Zone A is reserved for newly commissioned machines. Zone B signifies suitability for long-term operation. In zone C the machine is deemed in unsatisfactory condition and corrective action should be taken soon. Finally, in zone D vibrations can cause damage to the machine. The span of acceptable values differs with the machine class from I through to IV and their output power of 15 kW (class I), 75 kW (class II), 10 MW (class III), or greater.

Vibration velocity rms [mm/s]	Class I Small machines	Class II Medium machines	Class III Large machines Rigid supports	Class IV Large machines Flexible support
0.28	Good (A)	Good (A)	Good (A)	Good (A)
0.45				
0.71				
1.12	Satisfactory (B)			
1.8		Satisfactory (B)	Satisfactory (B)	Satisfactory (B)
2.8	Unsatisfactory (C)			
4.5		Unsatisfactory (C)	Unsatisfactory (C)	Unsatisfactory (C)
7.1	Unacceptable (D)			
11.2		Unacceptable (D)	Unacceptable (D)	Unacceptable (D)
18				
28				
45				

Table 2.3: ISO 20816 vibration severity chart with typical magnitudes [14]

The operational limits in the form of *alarms* and *trips* are usually established on the zone boundaries or close to them. Alarms are placed between zones B and C providing a warning about reaching the threshold significant for noticeable change. Trips in between zones C and D urge immediate action or machine shut down. Both limits should not exceed 1.25 times the upper boundary of the lower zones and are initially set based on previous experience with the machine [11].

ISO 13373-1:2002 delves into further nuances of vibration monitoring and expands on procedures outlined in terms of the ISO 20186 vocabulary. According to the standard, the data collection operates in continuous or periodic observation modes that follow an event or interval. Both designs can be permanently mounted. In continuous collection, it is recommended to have a "multiplexing rate sufficiently rapid, so there is no significant data or trends lost" [11]. When channels are scanned in succession with gaps between data points, the system is known as "scanning".

The condition monitoring program is run according to a flowchart adapted from one designed by the standard specifically to best benefit the plant. Those steps can be summarized as follows [11]:

1. Review machinery history and establish failure modes.
2. When vibration monitoring is not applicable, check for other condition-monitoring techniques or resort to preventive maintenance.
3. Select monitoring points and take preliminary vibration measurements.

4. Select vibration monitoring techniques: broadband, frequency analysis, or special techniques, and set parameters of measurement units.
5. Take baseline measurements.
6. Change levels that would warrant investigation.
7. Carry out routine condition monitoring.
8. If an alarm was exceeded, notify appropriate personnel to review data and trends, perform diagnostic evaluation, and repair as necessary. In case a new baseline is needed, continue in the step of taking baseline measurements.
9. Shut down the machine when the trip level is exceeded. Then proceed the same as after the alarm is triggered.

Measurement of vibrations should be accompanied by a description of the machine and its operating conditions. The machine description includes its identifier and type, power source, rated rotation speed and power, configuration (shaft or belt driven), and machine support. Measurement parameters are to be recorded alongside the measurement value itself, such as timestamp, transducer type, sensor location and orientation in **MIMOSA convention** (Machinery Information Management Open System Alliance), measurement units and units qualifier (p-p, rms), and other processing options (filters, number of averages, etc.) [11].

The transducer of choice for condition monitoring is the accelerometer. It measures the acceleration value of the body and velocity after signal integration. However, standards advise against double integrating for displacement. The recommended frequency range for an accelerometer is 0.1 Hz to 30 kHz. The choice of transducer mount significantly influences its resonance frequency. The stud mount and stiff cement mount least lower the effective resonance. The resonance is reduced to around 8 kHz when using the soft epoxy or permanent magnet.

Broadband measurement requires “frequency ranges of 0.2 times the lowest rotational frequency to the highest frequency of interest” [11], not exceeding 10 kHz, with rms velocity 0.1 - 100 mm/s. Bearings and gear diagnosis may push the upper-frequency limit even higher. The tolerances of amplitude and frequency calibrations fall into two types with allowable tolerances of $\pm 5\%$ or $\pm 10\%$ (Fig. 2.5).

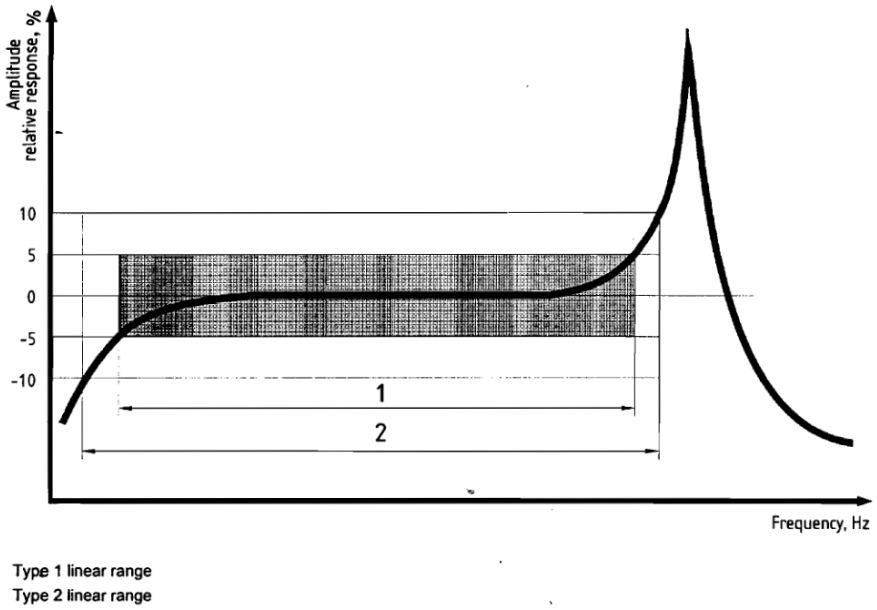


Figure 2.5: The transducer linear response and resonance in tolerance intervals [11]

Equipment's "health" can be mischaracterized when there are significant differences in the machine's normal operating conditions. Baseline measurements in all acceptable conditions are to be acquired to reduce the error in vibration evaluation. According to the bathtub curve (Fig. 2.2) reference signatures should be obtained after the initial part wear-in period. The reference spectral mask of the baseline condition is designed if maximal acceptance amplitudes are different for each significant frequency band [5].

The vibration baseline is defined by broad-band magnitudes and phases of motion vectors, the waveform in the time and frequency domain, the rotational speed of the machine as well as its frequency response to different speeds during start-up and coast-down captured in the Bode plot and waterfall plot. Changes during the machine's operation are then depicted in value trends. Trends can be shown by overall amplitudes or limited to frequency bands.

2.2 Signal preprocessing

The vibration signals in a factory environment are inherently full of disturbances. Nearby equipment operation and heavy objects handling in the surroundings can all contribute to the unwanted chaotic movement in otherwise mostly pure oscillatory

motion. In addition, accelerometers suffer from systematic measurement errors in the form of thermal noise, zero-g offset from slight miscalibration, and bias originating from a constant force of gravity. These unavoidable distortions are somewhat suppressable by digital filters. In the preprocessing stage, we consider trend removal and time synchronous averaging to eliminate external interference.

2.2.1 Detrending

The oscillatory motion should be centered around the zero level for further manipulation. The constant offset is eliminated simply by subtracting the overall mean from the signal. Moreover, the high pass DC blocker infinite impulse response (IIR) filter of 1st order can adjust to shifts of the average value over time (Equation 2.1). The transition band depends upon the choice of corner frequency f_{3dB} (Fig. 2.6).

$$y_k = (1 - \frac{\omega}{2}) \cdot (x_k - x_{k-1}) + (1 - \omega) \cdot y_{k-1}; \quad \omega = 2\pi \cdot \frac{f_{3dB}}{f_s} \tag{2.1}$$

A steeper 3 dB attenuation band is achieved by increasing the order of the filter. Then, the cutoff frequency should be such that filter coefficients are fractions to counteract rounding errors [15].

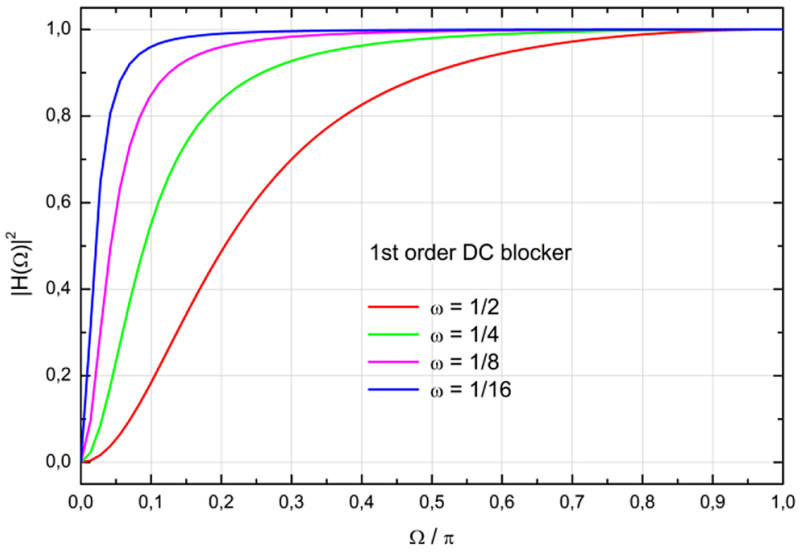


Figure 2.6: Transfer function of 1st order DC blocker filters [15]

The finite impulse response (FIR) filter is not recommended for DC component removal because of the undesirable ripple effect with the small number of taps. Cascaded-integrator-comb (CIC) filters are proposed as an alternative instead [16].

2.2.2 Time synchronous averaging

Time synchronous averaging (TSA) diminishes the impact of vibration sources unrelated to the rotational frequency and its harmonics. TSA averages time-domain waveform over N points and aligns it to a synchronization pulse with period T (Equation 2.2).

$$x_{TSA} = \frac{1}{N} \sum_{n=0}^{N-1} x(t + nT) \tag{2.2}$$

This technique has been successfully applied to the gearbox and bearing fault diagnosis [6, 17].

2.3 Feature extraction

After preprocessing, raw numerical vectors are merely low-level descriptors of the underlying physical phenomena. At first, these incomprehensible sequences of numbers are reduced to summary attributes called features in the process of feature extraction or feature discovery. Features can be hand-crafted (as is our case), learned implicitly within the model representation, or explicitly from an optimization problem solution.

Predictive maintenance has ideal prerequisites for the application of feature engineering because the signal is usually pseudo-stationary, and the trend monitoring variables come out of extensive domain expertise in mechanics. The advantages of the add-in extraction effort as opposed to processing unmodified samples are to gain better classification precision and reduce computational burden and storage capacity downstream with dimensionality reduction [18].

It is important to note that the design of features is not a standalone step in the machine learning pipeline, but it should be performed iteratively to improve the target model. Signal features are computed in the time and frequency domain [12].

2.3.1 Time-domain features

The most widely found features in the literature are rudimentary statistical measures of the central moment: mean, variance, standard deviation, skewness, and

kurtosis (Tab. 2.4). Statistics can be calculated in any domain, but the mean value should not be used in detrended data. The vibration severity metrics out of technical standards are also highly regarded. The amplitude characteristics include root-mean-square and peak-to-peak distance [19]. The waveform shape is captured by zero-crossing rate and average amplitude change [20].

The other significant time-domain attributes are derived as ratios of previous simpler ones. These ratios are crest, clearance, impulse factor, and shape factor (Tab. 2.4) [17]. Many articles have been successful in fault detection of bearings out of transients in impulsive signals with kurtosis, crest, and clearance factor [12]. It is also suggested that the shape factor can signify unbalance and misalignment faults [17].

Feature	**Equation**
Peak-to-peak	$X_{ppv} = \max(x_i) - \min(x_i)$
Zero-crossing rate	$X_{zc} = \frac{1}{N}\sum_{i=1}^{N} \operatorname{sgn}(x_i \cdot x_{i-1})$
Root mean square	$X_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$
Skewness	$X_{sv} = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma}\right)^3$
Kurtosis	$X_{kv} = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma}\right)^4$
Shape factor	$X_{sf} = \frac{X_{rms}}{\frac{1}{N}\sum_{i=1}^{N} \lvert x_i \rvert}$
Crest factor	$X_{cf} = \frac{max(\lvert x_i \rvert)}{X_{rms}}$
Impulse factor	$X_{if} = \frac{max(\lvert x_i \rvert)}{\frac{1}{N}\sum_{i=1}^{N} \lvert x_i \rvert}$
Clearance factor	$X_{mf} = \frac{max(\lvert x_i \rvert)}{\left(\frac{1}{N}\sum_{i=1}^{N} \sqrt{\lvert x_i \rvert}\right)^2}$
Average Amplitude Change (AAC)	$X_{AAC} = \frac{1}{N}\sum_{i=1}^{N-1} \lvert x_{i+1} - x_i \rvert$

Table 2.4: Time-domain features

2.3.2 Frequency-domain features

The mechanical faults present themselves as oscillatory patterns which are combinations of frequencies with various amplitudes. The Fourier transform is one of the most prominent strategies in power spectral density estimation. Experts on vibrodiagnostics utilize it as a primary signal processing technique for data analysis as it is recommended in ISO 13373-2 standard [21].

The inherent symmetries in the Fourier matrix made it possible to implement the Fast Fourier transform (FFT) algorithm with time complexity $O(n \log n)$. The drawback of the plain spectral analysis is the lack of resolution for events that occurred at distant time instants, and so their spectral components might adversely blend in together.

Feature	**Equation**
Spectral centroid	$X_{fc} = \frac{\sum_{i=0}^{N-1} f_i \cdot a(f_i)}{\sum_{i=0}^{N-1} a(f_i)}$
Energy	$E(N) = \sum_{i=1}^{N} a^2(t)$
Energy ratio	$E_r = E_i \ / \ \sum_{i=1}^{N} E_i$
Spectral roll-on	$E(f_c) = 0.05 \cdot E(f_s/2)$
Spectral roll-off	$E(f_c) = 0.85 \cdot E(f_s/2)$
Spectral flux	$X_{\text{flux}} = 1 - \text{corr}(A_{t-1}, A_t)$
Noisiness	$X_{noise} = \text{SNR}(X) = \mu \ / \ \sigma$
Shannon entropy	$H(X) = -\sum_i P(X = x_i) \cdot \ln(P(X = x_i))$
Spectral negentropy	$\Delta I_E(f, \Delta f) = \sum_{k=0}^{N-1} p \cdot \ln(p); p = \frac{\text{SES}^2}{\frac{1}{N}\sum_{k=0}^{N-1} \text{SES}^2}$

Table 2.5: Frequency-domain features

In the frequency domain, we can obtain the usual statistical properties of the distribution which are spectral centroid, skew, and kurtosis. Additionally, spectral roll-on and roll-off, fundamental frequency, entropy, negentropy, spectral flux, signal-to-noise ratio (noisiness), energy in frequency bands, and energy ratio are extracted

(Tab. 2.5) [22].

In geometric terms, the spectral centroid represents the barycenter of the frequency magnitude plot. Spectral roll-off gives a notion about the spectral distribution because it identifies the frequency f_c below which 85% of the signal energy is contained. The roll-on frequency is chosen so that 5% of the signal energy is below this value. According to the definition, spectral flux is normalized cross-correlation between two successive amplitude spectra. A flux value of one means the spectra are the most dissimilar.

"Negentropy measures the inclination of a system to increase its level of organization" [23]. The larger *spectral negentropy* suggests more fault-induced impulses. The squared envelope spectrum (SES) is interpreted as $E(k, f, \Delta f)$ which is an envelope in the Fourier domain: $\mathcal{F}\{|y(k, f, \Delta f)|^2\}$.

2.3.3 Features in harmonic frequencies

If a single principal frequency exists, it can be determined by maximum likelihood estimation. Such frequency would explain the signal spectrum the best [22]. The frequency spectrum is a discrete set of amplitudes where peaks have to be reliably identified to create representative attributes.

The essential peak-finding approaches are based either on magnitude or gradient. All found extrema are commonly filtered with the magnitude of prominences and the width at half prominence. In the magnitude-based method, the middle point x_i is compared to two neighbouring points and the peak is then: $x_{i-1} < x_i > x_{i+1}$. The gradient-based method evaluates the first derivative at the point that is equal to zero if the point is a local maximum, local minimum, or inflection point [24].

A substantial improvement is a robust non-parametric peak identification named *MMS* based on the sum of terms in an arithmetic progression based on maximum, minimum, and sum. MMS max-min finder in the elementary form processes points in the window of length 3, it advances one point and deems its middle point as a local extremum if it satisfies equalities below. Equation 2.3 is for the hill and Equation 2.4 is for the valley. The filtration techniques are incorporated in the adaptations of the MMS algorithm: MMS-WBF, MMS-SG, MMS-LH [24].

$$\frac{a_{max} - a_{min}}{S_3 - a_{min} \cdot 3} = \frac{a_{mid} - a_{min}}{S_3 - a_{min} \cdot 3} \tag{2.3}$$

$$\frac{a_{max} - a_{min}}{a_{max} \cdot 3 - S_3} = \frac{a_{max} - a_{mid}}{a_{max} \cdot 3 - S_3} \tag{2.4}$$

$$v_i^{(r)} = \frac{v_j}{\min |v_j - r \cdot v_i|} \tag{2.5}$$

Multiple harmonic series and the sidebands can be separated into a discrete set of frequency components, each with a central frequency, uncertainty, and amplitude $C_i(v_i, \Delta v_i, A_i)$ by an exhaustive search algorithm. Harmonic family identification is a non-trivial problem because of spectrum estimation errors. The criterion is proposed to select harmonics at the minimal distance from the true fundamental frequency multiple (Equation 2.5). Two series with the same fundamental frequency are merged and thought of as a modulation series [25].

2.3.4 Time-frequency domain features

The **Short-time Fourier transform** (STFT) splits the time-domain signal into equal-length intervals. Individual chunks have a 50% overlap and are multiplied with weights of window function to balance scalloping loss and spectral leakage due to the Fourier transform periodicity assumption. Traditionally, the Hann window is commonly used instead of the rectangular window [5, 21].

In the time-frequency domain, the same features can be derived as in the frequency domain, but in addition, the attributes are time-localized in this way. The STFT has a considerable flaw for implementation in a self-adaptable system and that is fixed resolution. The optimal window size has to be set beforehand or chosen after performing multiple transformations on chunks out of the range of lengths. Welch's method averages multiple consecutive blocks to better estimate the spectrum. We have already researched the suitability of STFT for online detection of constant frequencies [26].

The alternative to isolating weak impacts with high time resolution is a **Teager-Kaiser energy operator** (TKEO) (Equation 2.6). It is a tool for envelope analysis

to demodulate characteristic AM-FM signals present during bearing faults. Energy operator output can be utilized as a standalone feature attribute.

$$\psi[x(n)] = [x(n)]^2 - x(n-1)x(n+1) \tag{2.6}$$

Improved TKEO is necessary to prevent analysis from suffering from the noisy source. The key idea is to perform TKEO after signal decomposition into narrowband components with different center frequencies. The extracted modes are reconstructed with weights assigned based on their correlations to the original signal [27].

Time-frequency spectrum modification preserving localization of abrupt wideband spikes at time t_0 and simultaneously reducing energy smear over the larger region is a goal of **Transient-extracting transform** (TET). Post-processing of STFT window $G(t, \omega)$ involves multiplication of the spectrum with the Transient extracting operator (TEO). This operator is expressed in the form of a Dirac delta function $\delta(t)$ (Equation 2.7).

$$\mathrm{Te}(t, \omega) = G(t, \omega) \cdot \delta(t - t_0) \tag{2.7}$$

TET representation retains non-zero coefficients where the absolute value of the ratio between two STFTs, $G^{tg}[n, k]$ / $G[n, k]$ is less than half the sampling interval T. These two transforms use distinct windows $g[n]$ and $n \cdot g[n]$. Decomposition of a signal with TET is proved to produce significantly larger kurtosis (around 38 in TET, 4 in other methods) and hence better discriminate the transient fault [28].

2.3.5 Wavelet domain features

The bands for lower frequencies should be longer in time than for higher frequencies. The **Wavelet transform** (WT) possesses such a multiscale discrimination property, effectively increasing resolution in time-frequency plain. Wavelet basis functions are constructed for that purpose (Equation 2.8). There are several wavelet families. For example, Haar, Daubechies, Coiflets, Symlets, Morlet, and Meyer [17].

$$\psi_{s,\tau} = \frac{1}{\sqrt{s}} \psi \left(\frac{t - \tau}{s} \right) \tag{2.8}$$

Continuous Wavelet transform (CWT) (Equation 2.9) is done by scaling and translating the mother wavelet ψ picked out of the appropriate family [17]. The scale factor is denoted with s and time position with τ. The choice of wavelet type is data-driven because distinct wavelet shapes have an impact on the response and ultimately contribute to filter length. The decision lies between recognition abilities for impulse-like signals or the inclusion of wider surrounding space.

$$W_{x(t)}(s,\tau) = \frac{1}{\sqrt{s}} \int x(t) \cdot \psi^* \left(\frac{t-\tau}{s} \right) \, \mathrm{dt} \tag{2.9}$$

The CWT is computationally intensive when a highly detailed scale resolution is required because each wavelet scale convolves with the entire signal. The fCWT algorithm allows 100 times higher spectral resolution than previous implementations at the same speed. It increases performance 122 times compared to Wavelib and 34 times in comparison with PyWavelets [29].

The fast CWT algorithm reaches compelling improvement by applying Parseval's theorem to the wavelet transform formula that removes the dependence on the time offset parameter [29]. The convolution takes place with the mother wavelet in the Fourier base. Then, the inverse FFT produces the coefficients for individual scales.

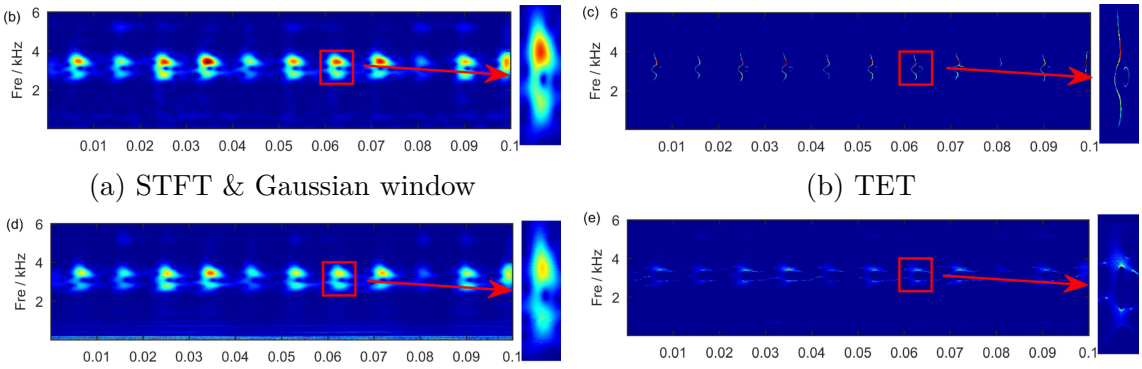
Synchrosqueezing Wavelet transform (SST) is a modification of CWT attempting to sharpen the representation of frequency components by coefficient reassignments from around the central frequencies towards the middle of the bands. The justification for these reallocations is rooted in the signal approximation as amplitude-modulated oscillating modes with additive noise $\eta(t)$ (Equation 2.10) [30].

$$s(t) = \sum_{k=1}^{K} A_k(t) \cos(\theta_k(t)) + \eta(t) \tag{2.10}$$

The components are defined by their instantaneous amplitudes $A_k(t)$ and instantaneous phases $\theta_k(t)$. The energy spread to adjacent bins can be effectively squeezed only in regions with constant phase and large enough component separation. Despite the promising properties of this transform, white noise causes severe interference in the resulting time-frequency map.

The spectrograms to illustrate the difference in the ability of the Fourier transform, the continuous Wavelet transform, and their modifications to pick up under-

lying patterns in bearing faults are shown in Fig. 2.7.



(a) STFT & Gaussian window

(b) TET

(c) CWT & Morlet wavelet

(d) SST

Figure 2.7: Comparison of time-frequency transform spectrograms [28]

The dyadic filter bank is another signal decomposition technique that generates subbands at multiple granularity levels. The practical realization of the multiscale description is **Discrete Wavelet Transform** (DWT). The DWT behaves as a quadrature mirror filter and splits waveforms using a wavelet filter to detail coefficients (D1) and approximation coefficients (A1) (Fig. 2.8a) [17]. The low-pass filter $h(k)$ creates approximation coefficients further decomposed at the successive levels. Detail coefficients represent the result of the high-pass filter $g(k) = (-1)^k h(1-k)$ after decimation by a factor of 2.

The maximum depth of the decomposition tree is $\log_2 n$ where n is the number of input samples. Energy, energy ratio, and entropy are prevalent features that succinctly encode the wavelet coefficients. Otherwise, the additional extracted levels raise the total number of data dimensions.

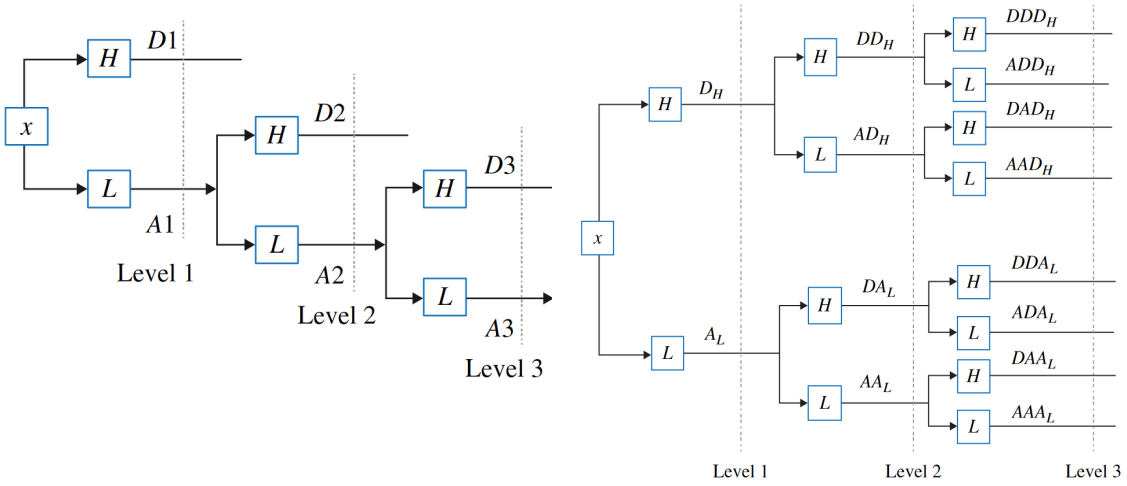
In washing machine status classification, the discrete wavelet transform with Daubechies wavelet (db4) and fifth-level decomposition provides features with a combination of approximation (cA5) and detail coefficients (cD1, ..., cD5). Washing machines belong to three categories: no fault, electric motor clamping screws problem, and a loose or broken counterweight. Extracted measures were sample mean and sample variances over autocorrelation functions of coefficients (AcDn) and smoothed coefficients cD1, cD2 by moving average filter [31].

Wavelet Packet Decomposition (WPD) applies filters to split detail coefficients identically as approximation ones (Fig. 2.8b) thus increasing the resolution in

the high-frequency bands and providing uniform spectrum partitioning.

$$w_{j,n,k} = \langle f, W^n_{j,k} \rangle = \langle f, 2^{j/2} W^n(2^j t - k) \rangle \tag{2.11}$$

Each wavelet packet coefficient $w_{j,n,k}$ captures subband frequency content around time instant $2^j k$ (Equation 2.11) [32]. This measure is an inner product of the source and scaled wavelet packet function. The aforementioned feature extraction established in DWT can be applied. For example calculation of the wavelet packet node energy.



(a) Discrete wavelet transform (b) Wavelet packet decomposition

Figure 2.8: Dyadic filter banks for discrete wavelet transform [17]

The wavelet packet energy ratio for weak feature extraction has been incorporated into the method of multiple frequency band demodulation (MFBD). The highest n energy coefficients are selected out of 8 narrow frequency bands to subsequently affect the principal components. Demodulation principle for the frequency bandwidth prescribes n to satisfy condition: $1 \,/\, 2^n > f_{\text{modulation}} \,/\, f_s$. The first few eigenvectors that explain together more than 80% of energy are retained to reconstruct the signal with a Fourier transform and retain the weak fault components [33].

Tool wear diagnosis based on acoustic emission (AE) signal considers *wavelet packet energy* in bands E_8, E_{10}, E_{12}, *energy ratios* P_8, P_{13}, and *energy entropy* as having high correlation ($|r| > 0.8$) with the band saw flank face width. The acoustic signal is decomposed into three layers using Daubechies db3 wavelet. The bottom layer contains bands numbered 7 through 14, each with a bandwidth of 62.5

kHz, because of the 1 MHz sampling frequency. The feature vector constructed in the article includes other statistical metrics out of power spectral density that have reached a notable correlation with evolving wear. The statistics are *skewness*, *kurtosis*, *shape factor*, and *centroid frequency* [34].

Discrete wavelet transform and wavelet packets partition the spectrum into predefined frequency bands that do not always adequately capture individual elementary oscillations. Adaptive spectral segmentation is needed to extract separate intrinsic mode functions (IMF).

Wavelet domain features and associated spectral segmentation appear to be a powerful tool for conserving data bandwidth. These variables turn out to be difficult to communicate and comprehend. Therefore, we further use more standardized metrics.

2.4 Feature transformations

Numerical features from the feature extraction phase have non-normal distributions and span wide ranges. Broad differences among features skew the spread on a particular axis. Inevitably, it can degrade the discernment of fault diagnostics models that map input onto a smooth function as regression does [35]. The feature scaling, power transform, and principal component analysis modify attribute values to gain more meaningful predictors, but one must be cautious in model interpretation.

2.4.1 Feature normalization

Vibrations are measured with an accelerometer in the form of 3D vectors where each axis has its own component. Models have to be resilient to any slight inclination or sensor orientation. To satisfy this prerequisite, feature f extracted from all three dimensions is first composed into a single vector and the Euclidean norm is computed [36].

$$\widetilde{f} = \sqrt{f_x^2 + f_y^2 + f_z^2} \tag{2.12}$$

Feature normalization most often takes two forms. **Min-max scaling** changes

the original range of values into an interval $[0, 1]$ (Equation 2.13). **Standardization** (Equation 2.14) constrains the mean of the variable to 0 with a variance of 1 [35].

$$\widetilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2.13}$$

$$\widetilde{x} = \frac{x - \bar{x}}{\sigma_x} \tag{2.14}$$

2.4.2 Principal components analysis

Highly correlated features generated from a relatively small original column space are redundant as they do not provide any additional information for diagnostics. **Principal Component Analysis** (PCA) solves this problem by projecting potentially linearly dependent features into a new feature space where the incoming information is preserved in a smaller number of features [35]. The threshold of how many principal components are picked depends on the amount of explained variance and desired quantity of data reduction.

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}} \tag{2.15}$$

PCA consists of taking the Singular Value Decomposition (SVD) (Equation 2.15) of the mean-centered input matrix. The disadvantage of this method is the loss of explainability in transformed space though it generally outperforms the model working with hand-crafted features [12]. The signal samples can be processed directly by PCA without going through an intermediate step of calculating statistical measures.

2.5 Feature selection

Features do not contribute to the predictive power of the model with an even share. A certain subset can achieve better results than others. Choosing the optimal feature subset is the NP-hard combinatorial problem.

2.5.1 Filtering methods

The features can be chosen intrinsically as a part of a model by *embedded methods* or by machine learning search algorithm at a serious computational expense in *wrapper methods*. However, we will focus on *filtering methods* which rank the predictors in order of their importance for the problem at hand and separate the best performing group [18]. The most common strategy is *K-Best Selection*.

The general steps in selecting the appropriate predictors are as follows [17]:

1. **Subset generation** - sets of features are generated in different search directions and with various strategies. Attributes are either appended to an empty set or pruned away from a universal set, sequentially or randomly.
2. **Subset evaluation** - comparison of subset quality is assessed with relevance measures, some of which are discussed below.
3. **Stopping criteria** - search is exhausted when the specified number of features has been found, subset metrics cannot be improved further, or satisfactory model performance is achieved. Subset generation and evaluation can be performed multiple times until the stopping criteria are met.
4. **Validation** - the resulting subset is tested for the specific model on synthetic and real-world datasets against well-known results.

Filter-based feature selection is preprocessing step independent of model choice with small computational requirements. Measures of information, correlation, similarity, and interdependence output the relevancy rating. Predictors are rated individually or in interacting congregations.

2.5.2 Feature importance ranking

Most of the scores are based on supervised learning, so they expect true class labels to apportion the measurements respectively. After the scores are assigned to the first n features, those below a threshold are removed.

The frequently used scores upon which the feature relevance is ordered are [17]:

- **Variance threshold** - removes low-variance features below the set threshold.

- **Correlation coefficient** - expresses the linear relationship between two variables. The codependent variables are of three sorts: quantitative, ordinal, and nominal. The choice of coefficient calculation is determined by the type of variables under consideration as shown in Table 2.6.

  Pearson correlation coefficient expresses similarity between two quantitative features: f_i and f_j. In the classification setting, the correlation of feature f makes sense only with dichotomous target class label c using *point biserial coefficient*. Mathematically, this is equivalent to the Pearson coefficient. Rank correspondence is quantified with either Spearman rho or Kendall's Tau [37].

Variable Y\X	**Quantitiative X**	**Ordinal X**	**Nominal X**
Quantitative Y	Pearson r	Biserial r_b	Point Biserial r_{pb}
Ordinal Y	Biserial r_b	Spearman ρ	Rank Biserial r_{rb}
Nominal Y	Point Biserial r_{pb}	Rank Bisereal r_{rb}	Phi, L, C, Lambda

Table 2.6: Correlation coefficients

  Attributes are ranked in descending order according to the absolute value of their correlation coefficient. We seek the highest correlation to the class label.

$$r(i) = \frac{\mathrm{cov}(f, c)}{\sqrt{\mathrm{var}(f) \cdot \mathrm{var}(c)}} \tag{2.16}$$

- **Fisher score** - measures the difference between the means of the classes. It is interchangeable with ANOVA F-value, but it is evaluated for each feature X^j separately. Ideally, the features in the subset have large distances between samples of various classes in C and distances within a class are the smallest possible. In the formula (2.17), n_j is the sample size of jth feature, μ^j is its sample mean, and μ is the overall mean.

$$\mathrm{FS}(X^j) = \frac{\sum_{i=1}^{C} n_i(\mu_i^j - \mu^i)^2}{\sum_{i=1}^{C}(n_i - 1) \cdot (\sigma_i^j)^2} \tag{2.17}$$

- **Mutual information** - quantifies the dependence between features, or between features and class labels. It is almost identical to Information Gain.

The probability distribution of proximity of variables derives from the relative entropy known as the Kullback-Leibler distance. Mutual information presumes variables are discrete. In the case of quantitative variables, mutual information is estimated by binning or nearest neighbours methods [38].

Probabilities $P(x)$, $P(y)$, $P(x, y)$ are estimated in the contingency table from event occurrence count to all sample population $|x|$ / N. Joint probability $P(x, y)$ represents samples of feature x simultaneously in class y.

$$\mathrm{MI}(X, Y) = \sum_{y \in Y} \sum_{x \in X} P(x, y) \cdot \log \left(\frac{P(x, y)}{P(x)P(y)} \right) \tag{2.18}$$

Multiple subsets of predictors produced by each evaluation metric can train several variants of a classification model. Sets of attributes can be combined into an ensemble by *electoral system*. One such example is **majority voting** which chooses the best feature out of the group. **Rank product** unifies several feature orderings by computing the geometric mean of feature rank in every experiment realization [39].

2.6 Diagnostics techniques

Fault identification in the rotating machinery is a one-class or multi-class classification problem acting in a semi-supervised manner because labels for degraded conditions are scarce in practice. The automation goals in monitoring can be broadly categorized as anomaly detection and recognizing the momentary fault type.

The guiding principles for algorithm selection are simplicity in terms of their straightforward visual explanation for the production managers, and the ability to progressively improve the model on the streaming data to address peculiarities in individual machine constructions.

2.6.1 Novelty detection

Anomaly, novelty, or outlier detection determines whether a health status deviates considerably from the baseline profile. The expert can then step in and diagnose the

machine after the notice. Anomaly is a rare observation different from the others, raising suspicion that it was created by unrelated behavior [40]. The observations get assigned anomaly scores, and those over the threshold are novelties.

The measurements coming in the steaming fashion have to be processed in a single pass. The detection model must deal with the minimal admissible assumptions about the nature of the input events. The outliers are derived based on non-parametric statistical models, nearest-neighbour clustering, and isolation-based approaches [41].

DenStream is a density-based algorithm adapted from DBSCAN to cluster streaming data of arbitrarily shaped groups. Samples it includes in the first step into coherent clusters are core data points in each other's neighbourhoods. Core points have at least *MinPts* (μ) points in their neighbourhood of radius *Eps* (ε) units. Then non-core points in the proximity area of the core point are attached to the cluster containing it [42].

Quality of clustering results is evaluated by *Silhouette score* in the range [-1; 1]. It demands points within clusters to have high cohesion and at the same time to have large separation from other clusters. The low score indicates a too-small or too-large number of clusters [43].

In the online maintenance phase, DenStream summarizes the nearby observations into core *micro-clusters* that can be potential micro-clusters or outlier *micro-clusters* (Fig. 2.9) [45]. The (outlier) *o-micro-clusters* can grow into (potential) *p-micro-clusters* when they encompass $\beta\mu$ points. The outliers are discounted after some time in accordance with the decay function: $f(t) = 2^{-\lambda t}$ or below the lower weight limit ξ. The on-demand offline stage runs DBSCAN over the approximate representation in micro-clusters to deliver final apportionment [46].

2.6.2 Classification

Accurate multi-class classification of machine faults according to the characteristics of known causes is a much more difficult task than novelty detection. Fault combinations have to be recorded and transformed into feature space. Interactions among fault root causes have to be considered. We are aware of rapid advances in

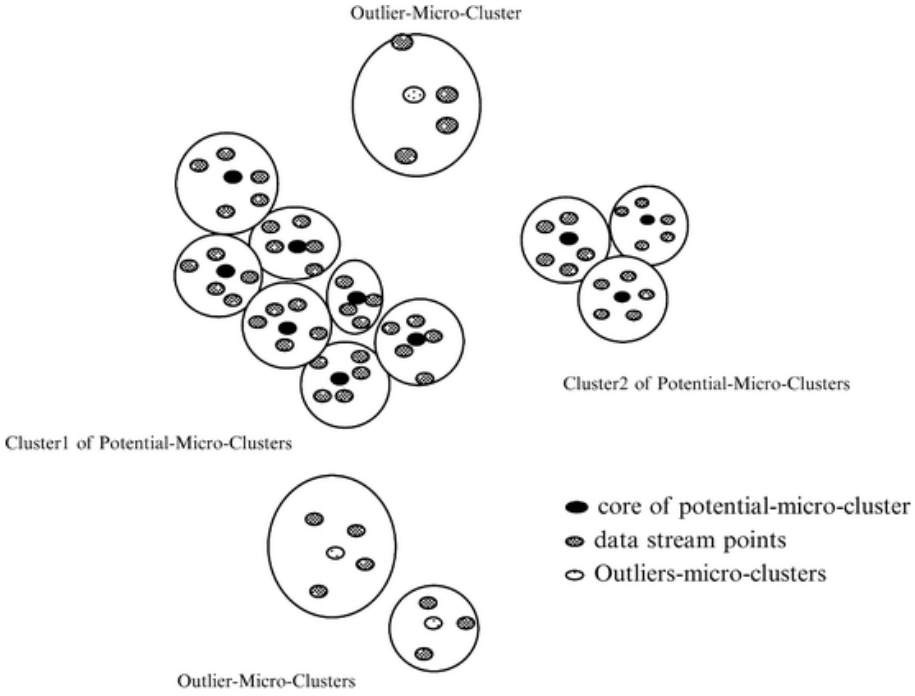


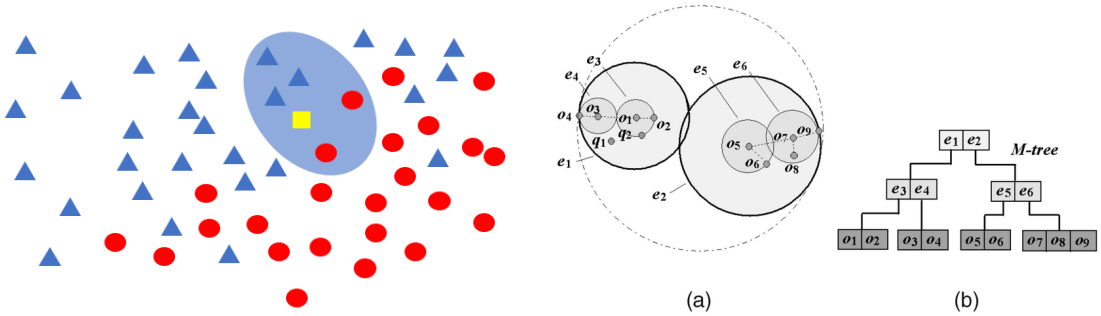
Figure 2.9: DenStream [44]

knowledge transfer for deep neural networks [47]. So far, solutions seem to be hard to implement. Therefore, we opt to use a simpler model.

The performance of classification is estimated by several metrics on the validation set obtained using hold-out or cross-validation techniques. Frequently used quantities for classifier model evaluation include accuracy, precision, recall, f1 score, area under the ROC curve, and counts of hits and misses in a confusion matrix.

K-nearest neighbours (k-NN) assigns the data point to the class where the majority of k closest instances belong (Fig. 2.10a). It means it can work in a semi-supervised environment because it can infer labels just from knowing a few annotations. The major drawback of k-NN is a preference for the majority class in imbalanced class-size datasets. The issue is mitigated with class weights or resampling classes by oversampling or undersampling. The model without class balancing drops in precision from 15% to over 40% depending on the imbalance ratio [48]. The k-NN algorithm requires features to be normalized to assign the same importance to each predictor.

The meaning of distance between feature vectors $\mathbf{x}$, $\mathbf{y}$ have to be defined, so several metrics are available like *Euclidean distance*, *Mahalanobis distance*, *Manhatann*



(a) k-NN with k = 5 (b) M-tree data structure

Figure 2.10: Nearest neighbours classification algorithm [49]

distance etc. (Tab. 2.7) [50, 51]. The optimal k parameter is set by supervised learning according to the breaking point in the elbow curve that plots choices of k against the error rate. The demanding neighbourhood queries are sped up using spatial indexes in databases that utilize search trees such as kd-tree, R-tree, or M-tree.

Distance	$d(\mathbf{x}, \mathbf{y})$
Manhattan distance	$\lvert x_i - y_i \rvert$
Euclidean distance	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Mahalanobis distance	$(\mathbf{x} - \mathbf{y})^T C^{-1} (\mathbf{x} - \mathbf{y})$

Table 2.7: Distance metrics for k-NN

The nearest-neighbour classifier has been successfully applied in machinery fault diagnostics. On the CWRU bearing dataset, the k-NN with the accuracy of 96.2% slightly outperformed SVM (95%) on the combination of time and frequency-domain features, on set of time-domain features (k-NN 91.2%, SVM 88.8%), and on frequency domain features (k-NN 98.8%, SVM 96.2)% [52].

Comparison of SVM, KNN, and KLDA for the classification of three kinds of bearing faults reaches more than 95% of data reduction decreasing the time complexity of the models. The test rig rotates at 800 rpm and is sampled at 40 kHz. Their approach uses feature sets of average, kurtosis, skewness, and standard deviation applied on vibration signal after the Fourier transform and its second derivatives over the moving window. The best accuracy was reached by 40 thousand *PSD* set

with 99.13% with KLDA compared to 95.64% with k-NN and Mahalanobis metric. The *Statistical* set of 8 × 228 features has accuracy 98.27% (KLDA) and 93.53% (k-NN) [53].

2.6.3 Incremental learning

Online or incremental machine learning operates on the streaming data, updating the model parameters with each new incoming event or in mini-batches. This approach finds its use in big data processing when the whole dataset is not available in advance or cannot be processed at once because of memory limitations.

There are some additional obstacles to watch out for with incremental learning in comparison to batch learning [54]:

1. **Concept drift** is defined as the change in data distribution function over time. The two types of concept drift are virtual and real. In virtual drift, changes occur only in the input distribution. Real drift means that the alteration comes to underlying functionality. Concept shift occurs with an abrupt change.
2. **Stability-plasticity tradeoff** concerns the speed with which the model adapts to new information. The model can react quickly, making it less stable, or retain patterns for longer but become irresponsive to sudden shifts.
3. **Model complexity** should be adjustable to ensure flexibility in unforeseen circumstances. Simpler models in the ensemble can also further increase prediction robustness. Resource limitation bound complexity from above.
4. **Memory model** can store aggregates from seen observations and its typical examples or finite window of latest samples with forgetting factor.

Model benchmarking in incremental learning is achieved by comparing models to their batch counterparts or using progressive validation [55, 56]. It has been shown that incremental clustering algorithms have overall worse accuracy than batch versions [54]. In the validation process, precautions should be taken to prevent data leakage from future events into the past.

Deployment of an online machine learning model alongside supporting services is different from established MLOps processes. Model store and Inference service

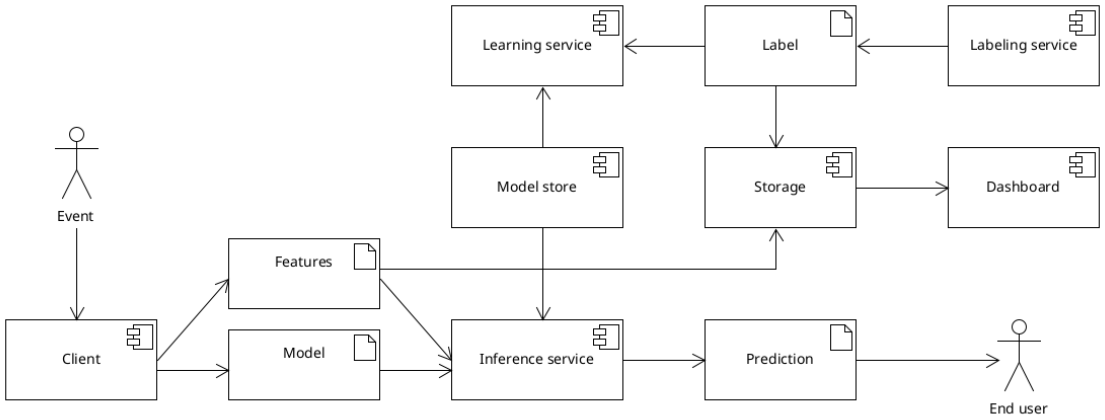


Figure 2.11: Incremental learning deployment architecture [57]

components are supplemented with Labelling and Learning service [57]. Their goal is to tune model parameters gradually as additional ground truth labels are provided. Labels can be provided later in a scheme called "log and wait", but data features are stored until such time.

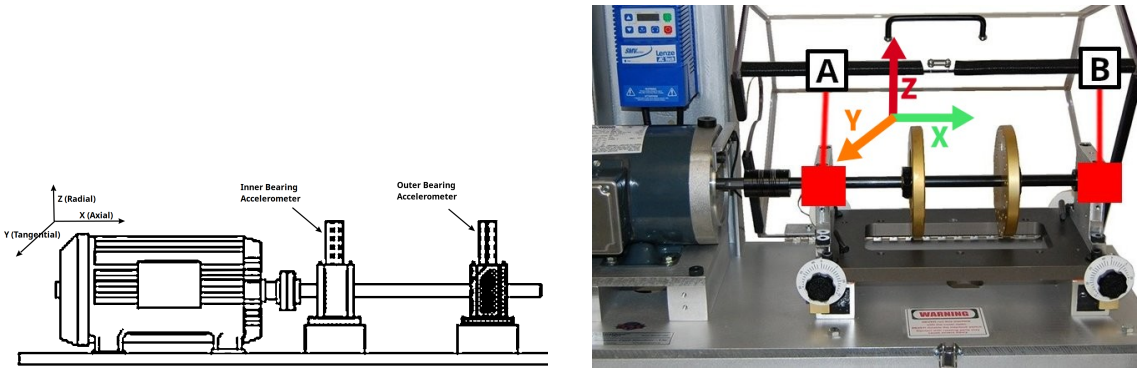
2.7 Datasets of machinery faults

The experimentally designed features' relevancy is first proven in comparison to comprehensive benchmark datasets. There are a few standardized datasets used in the related work, e.g. [58].

MaFaulDa dataset combines vibration and acoustic measurements of the shaft in deviating positions and bearing abnormalities. *CWRU dataset* focuses solely on faults in ball bearings. Another less known dataset concerns shaft unbalance, but compared to the previous two, it demonstrates behavior during speed up.

2.7.1 Machinery Fault Database

MaFaulDa [59] is a collection of 1951 multivariate time series for 4 different operational conditions on rotor kit Alignment Balance Vibration Trainer (ABVT) (Fig. 2.12). Each series has 5 seconds in duration and is captured at 50 kHz. Vibration signals were obtained with piezoelectric accelerometers with a linear response up to 10 kHz, amplitude range to ± 490 m/s^2, and resolution step of 10.2 mV per m/s^2.



(a) Schematic diagram [60] (b) Mechanical construction [61]

Figure 2.12: Machinery fault simulator for MaFaulDa

Observations were conducted on three cardinal axes simultaneously with 2 sets of accelerometers, each one associated with one bearing (inner and outer bearings) (Fig. 2.12). Additionally, a magnetic speedometer produced a pulse on shaft turn. The cardioid condenser microphone recorded sound emissions at a frequency range 20 Hz - 20 kHz. Sensors were fed into a four-channel dynamic signal acquisition module.

Columns in the dataset are organized as depicted in table 2.8. Machine rotational speeds were kept constant during a particular measurement, but covered a range from 737 to 3686 rpm with steps of approximately 60 rpm (equiv. 10 Hz - 60 Hz) [60]. The maximal rotational frequency achieved with a high unbalance load is 3300 rpm.

Columns	**Description**
1.	Pulse with modulation of speedometer signal to estimate rotation frequency (in TTL levels)
2., 3., 4.	Underhang bearing accelerometer (inner - between the rotor and motor) - axial, radial, tangential direction
5., 6., 7.	Overhang bearing accelerometer (outer - outside most position after the rotor) - axial, radial, tangential direction
8.	Microphone

Table 2.8: MaFaulDa description of columns

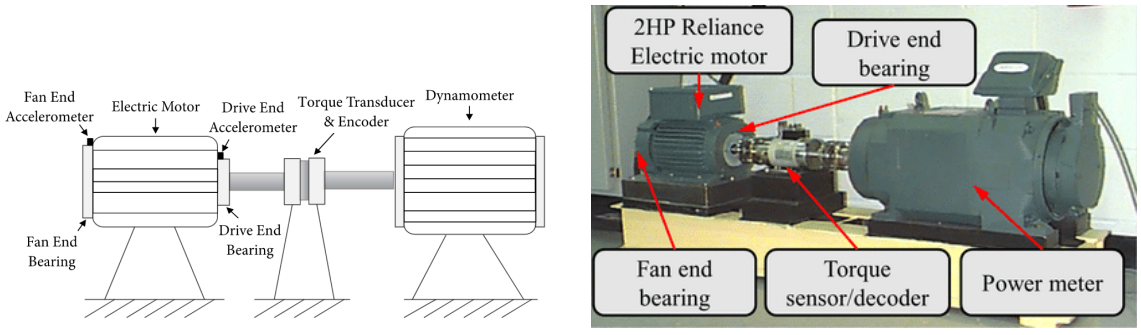
This database contains normal operating conditions, faults out of unbalance, horizontal and vertical shaft misalignment, and three types of faulty bearings in

inner and outer positions: outer track, inner track, rolling elements [60].

- **Normal** conditions are baseline without the adverse effect of fault at 49 different rotation speeds.
- **Unbalance** shaft time series uses 8 unbalancing weights from 6 to 35 grams and varying 45 - 49 speeds for each weight, adding to 333 mass unbalance loads.
- **Vertical misalignment** set comprises 50 signals each (or 51 in one instance) obtained under displacements: 0.51, 0.63, 1.40, 1.90, 1.27, 1.78 mm.
- **Horizontal misalignment** signals were recorded under displacements: 0.50, 1.00, 1.50, 2.00 mm, each with 49 different speeds (or 50 in one instance) [60].
- **Bearing faults** are unnoticeable without unbalance. Therefore, weights of 6, 20, and 35 grams were attached to induce a detectable effect. Each mass was combined with cage, outer race, and ball faults at multiple rotation speeds, usually at 50 different speeds.

2.7.2 CWRU bearings dataset

In Case Western Reserve University (CWRU) bearing dataset [62] recordings were made of a fan end and drive end bearings under motor loads of 0, 1, 2, and 3 Horsepower (equivalently 0, 0.75, 1.49, 2.24 kW). Shaft speed was unaltered in all experiments, but it fluctuated between 1720 and 1797 rpm (approx. 29 Hz).



(a) Schematic diagram [63] (b) Mechanical construction [64]

Figure 2.13: CWRU machine apparatus

Single point defects were created with diameters of 0.007, 0.014, 0.021, 0.028, and 0.040 inches (equivalently 0.18, 0.36, 0.72, 1.02 mm). Fault locations on bearings

are in the inner raceway, in the outer raceway directly and orthogonally relative to the load zone, and on rolling ball elements (Fig. 2.13) [52].

Columns	Description
1. DE	Drive end accelerometer samples
2. FE	Fan end accelerometer samples
3. BA	Base accelerometer samples (optional)
4. RPM	Rotation speed of the motor in rpm

Table 2.9: CWRU dataset description of columns

The sampling frequency during baseline set, drive end, and fan end bearing capture is 12 kHz. For drive end bearings, samples were taken at 48 kHz. The duration of the time series varies from 5 to 40 seconds. Drive end and fan end bearing signals are measured in each experiment. Accelerometer was sometimes mounted on the supporting baseplate.

2.7.3 Unbalance of the rotating shaft

Unbalance Detection of a Rotating Shaft [65] is a Kaggle dataset that simulates 4 different unbalance strengths.

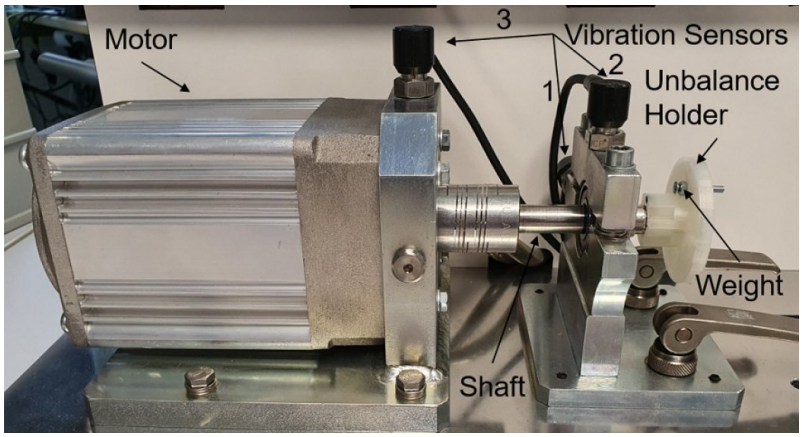


Figure 2.14: Motor driving shaft in unbalance measurement [66]

The setup is shown in Fig. 2.14. A mass of 3.28 grams (or 6.61 grams during the severe unbalance test) is attached to the unbalance holder in 5 sets (numbered 0 - 4) on the radii 0, 14, 18.5, 23, 23 mm. The rotation speed of the motor is perpetually rising between 630 and 2330 rpm in development datasets (marked with

suffix D) and speeds from 1060 to 1900 rpm in the evaluation datasets (suffix E). The vibrations were recorded at a sampling rate of 4 kHz [66].

Columns	**Description**
1. V_in	Input voltage to the motor controller (V)
2. Measured_RPM	Rotation speed of the motor (rpm)
3. Vibration_1	1. Vibration sensor (samples)
4. Vibration_2	2. Vibration sensor (samples)
5. Vibration_3	3. Vibration sensor (samples)

Table 2.10: "Unbalance on the rotating shaft" dataset description of columns

The accelerometers used are piezoelectric and have a frequency range of up to 10 kHz, dynamic range of $\pm 490\ m/s^2$, and resolution step of 10.2 mV per m/s^2. These sensor parameters are the same as in the case of MaFaulDa. In total, three different uniaxial accelerometers are mounted on the motor housing.

3 Design

During the design stage, we set out research questions and explain the machine learning pipeline for machinery defects diagnostics in the MaFaulDa dataset. The explanatory analysis of MaFaulDa hints at the underlying dependencies within feature sets that can degrade the machine learning model of choice. We apply the methodology for vibration measurements found in technical standards to requirements for the data logger and data acquisition procedures for the industrial equipment.

3.1 Research questions

The thesis aims to provide answers to five research questions. The main focus is on making data flow more efficient in an industrial sensor network that could monitor rotating machines. The **research questions** are:

RQ1. Which time-domain and frequency-domain features are extractable from vibrations to provide the most accurate record of machinery faults?

RQ2. What is the reduction in transmission goodput with signal features?

RQ3. What accuracies of fault classification are achievable with tiny feature subsets?

RQ4. How do the vibration signals collected on the industrial machinery behave?

RQ5. How can machinery faults be identified incrementally from collected events?

In accomplishing the objectives of our research, we propose several **goals**:

- Statistically and visually describe vibration signals from the Machinery fault database (MauFaulDa).
- Establish a list of conditions that will be investigated in the experiments.
- Prepare the dataset to be used in conjunction with machine learning models namely by identifying labels and balancing classes.

- Find the best subsets of features in the time and frequency domain using previously analyzed feature extraction and selection methods.
- Evaluate the performance of models described in the diagnostics section with a significant focus on the k-nearest neighbours algorithm.
- Acquire measurements of vibrations from machines in the real environment to form a novel dataset of machinery behavior.
- Develop hardware for a sensor unit and implement its firmware to obtain vibration measurements with the quality demanded by vibrodiagnostics standards.

3.2 Machine learning pipeline

We adopt a similar data processing framework found also elsewhere in the literature. It generally consists of signal acquisition, feature extraction, dimensionality reduction, and pattern recognition or fault detection [67]. However, the realization of the steps must be decomposed closer to suit the dataset structure, desired processing goals, and investigated parameters. The design of the data flow is in Figure 3.1.

The signals are in columns of CSV files within the MaFaulDa ZIP archive. **In preprocessing step**, they are first associated with metadata according to their file's path in the directory structure. The hierarchically topmost folder describes the simulated defect. There are directories for *normal*, *imbalance*, *horizontal-misalignment*, *vertical-misalignment*, *overhang* (outer bearing), *underhang* (inner bearing). Each bearing has an additional folder layer for dividing up bearing faults: *ball_fault*, *cage_fault*, *outer_race*. Fault categories, except the normal class, contain various fault severities marked with in grams or millimeters. Fault severities, in turn, consist of individual files with different rpm speeds.

The labels for shaft misalignment in vertical and horizontal directions are merged into the same group because we neglect the **accelerometer spatial orientation**. This merging leaves **six types of labels**: baseline, two shaft faults are imbalance and misalignment, and three bearing faults are cage fault, ball fault, and outer race fault.

Depending on the **chosen bearing position**, only time series associated with

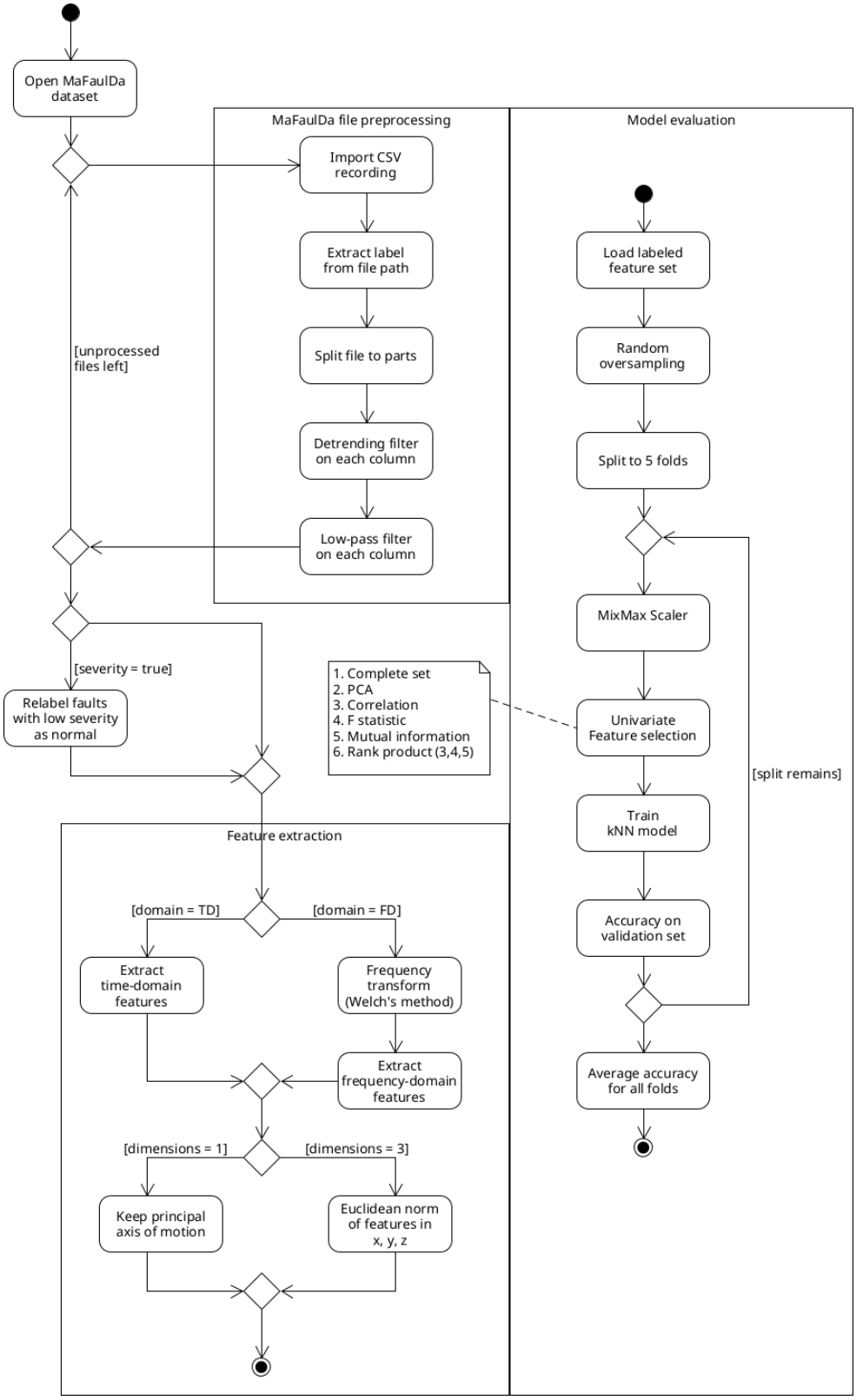


Figure 3.1: Machine learning pipeline for MaFaulDa dataset

that bearing are retained because the standards demand that bearings are assessed separately. Fault classification is concerned with bearing in direct contact and the shaft mechanically passing through it.

The strength of the vibrations in response to the defect is dependent on the shaft's **rotational speed**. Speed in units of rpm is calculated from pulsed speedometer output. It is an average distance between two successive rising edges:

$$\text{rpm} = 60 \,/\, \overline{\Delta t} \tag{3.1}$$

File description is the triplet of **fault**, **severity**, and **rpm**. After labeling, we insert a step to split the time series into parts when the duration is long enough to introduce more rows artificially. This assumption for splitting is not met for MaFaulDa because, for the desired frequency resolution for features of close to 1 Hz, we need a window size of 32768 samples. Variation reduction in spectral estimation necessitates averaging at least twelve overlapping windows. The full length of five seconds provides fifteen complete windows.

The **DC component** in the three-dimensional vibration signal is removed by subtracting the global mean. Digital IIR Butterworth low pass filter of 5^{th} order immediately follows with cutoff frequency 10 kHz at -3 dB. Before its use, the peak appeared at 20 kHz with a sideband as an unwanted artifact. It could not have been reliably recorded due to the linear frequency response of the sensor up to 10 kHz. At the same time, such a frequency is outside the range of any feasible MEMS accelerometer.

Initial labels can be exchanged for a label of fault-free condition when the defect severity is low. The severity levels are not identical in every group. The numerical amounts for each level are sorted in ascending order within each fault group, and then these levels are normalized using a min-max scaler.

The preprocessed signal from each axis of accelerometer is packed up to **two base feature sets** in time domain (TD) and frequency domain (FD). The feature extraction formulas are identical with formulas presented in the section 2.3 of analysis chapter. The frequency transform is Welch's method on FFT vectors from 32768 samples after Hann windowing and 50% overlap. The feature sets are:

- **TD has 10 time-domain features** - peak-to-peak amplitude (*pp*), zero-crossing rate (*zerocross*), root mean square (*rms*), skewness, kurtosis, shape factor, crest factor, impulse factor, clearance factor, average amplitude change (aac),
- **FD has 11 frequency-domain features** - spectral centroid, standard deviation (*std*), skewness, kurtosis, roll-on frequency, roll-off frequency, spectral flux, noisiness, spectral negentropy, energy, entropy.

The four conditions are applied in each instance of the pipeline to filter observations with calculated features to create **24 scenarios**:

- **Source bearings** - samples and fault labels are left in either just for the inner bearing (A), for only the outer bearing (B), or for both bearings (A+B).
- **Feature domain** - the set of extracted features as an input to downstream models is changed either to TD or FD set,
- **Accelerometer axis** - either one principal direction of motion or all three dimensions are aggregated under the same feature name (1 or 3),
- **High severity faults** - to more precisely simulate fault frequencies found in the real environment, the low severity defects can be considered as normal machine operation (Yes or No).

The **fault classification** performance metrics are evaluated after balancing classes, feature normalization, and 5-fold cross-validation in the k-nearest neighbour classifier with Euclidean distance metric. The population of classes is evened out with random oversampling of all but the majority class. In the multi-class classification, the micro-averaged performance metrics of accuracy, precision, and recall give the same number. Therefore, accuracy is only one on the output.

The hyperparameter of k-neighbours and a feature subset size are tweaked. The results are compared under different preprocessing conditions and in multiple experiments:

- **Complete feature sets**: compare an accuracy of the k-NN model with k-value as an odd number in the range from 1 to 37 on the dataset under different conditions. The TD and FD sets are treated separately.

- **Feature combinations** - every combination of feature subsets is evaluated with 2, 3, and 4 members to obtain the statistical distribution of all possible k-NN models. The most informative attributes for defect identification are found by brute force in the distribution. The k-value is chosen from options of 3, 5, and 11 neighbours sequentially. The single k-value generates $\sum_{j=2}^{4} \binom{n}{j}$ combinations where n is cardinality of complete feature set. In total 375 models for TD and 550 models for FD are tested under identical conditions. The observed effect is a decrease in accuracy while shrinking the number of features to the bare minimum.

- **Feature selection methods** - the choice of feature subset is performed in linear time complexity as opposed to exponential in combinatorial case. The dimensionality reduction with **principal components analysis** (PCA) does further extraction and retains only the components that the best explain the variance. The disadvantage is that the resulting linear combination cannot be interpreted easily.

  The best feature subset from complete sets is picked after ranking the features according to their importance. The bivariate scores for ranking are the mean of point biserial **correlation** to every class label as a dichotomous variable, **Fisher score** (ANOVA F statistic), and **Mutual information**. The rank product combines the three latter scores into the ensemble. The accuracy comparison of the feature selection strategies is made in relation to the entire model accuracy distribution.

- **Incremental learning** - the order of samples simulates the gradually worsening state of the machine and the delayed annotation of defects. Observations are sorted based on increasing relative severity. The faults within severity levels are shuffled randomly.

  The **tumbling window** of lengths 1, 10, 100 measurements imitates regular expert visits annotating observations recorded until that moment. Labels for the whole previous window are supplied at once. Another common problem with online learning is **missing annotations** due to the size of the dataset. The equal-length gaps of 0, 2, 10, and 50 labels are skipped before another ob-

servation is annotated. This approach of skipping samples without considering their representativeness can harm the predictions.

The experimental design involves the filtering conditions that create 24 forms of the original MaFaulDa. The four main experiments test hyperparameters on each dataset variation. Those are k-neighbors, a number and kind of features in batch learning tumbling windows, and label skips in incremental learning.

3.3 Exploratory data analysis of MaFaulDa

A better rationale for future models behaviour is given by examining the number of observations per fault category, spatial separability of the categories, scales of the attributes and their interdependencies.

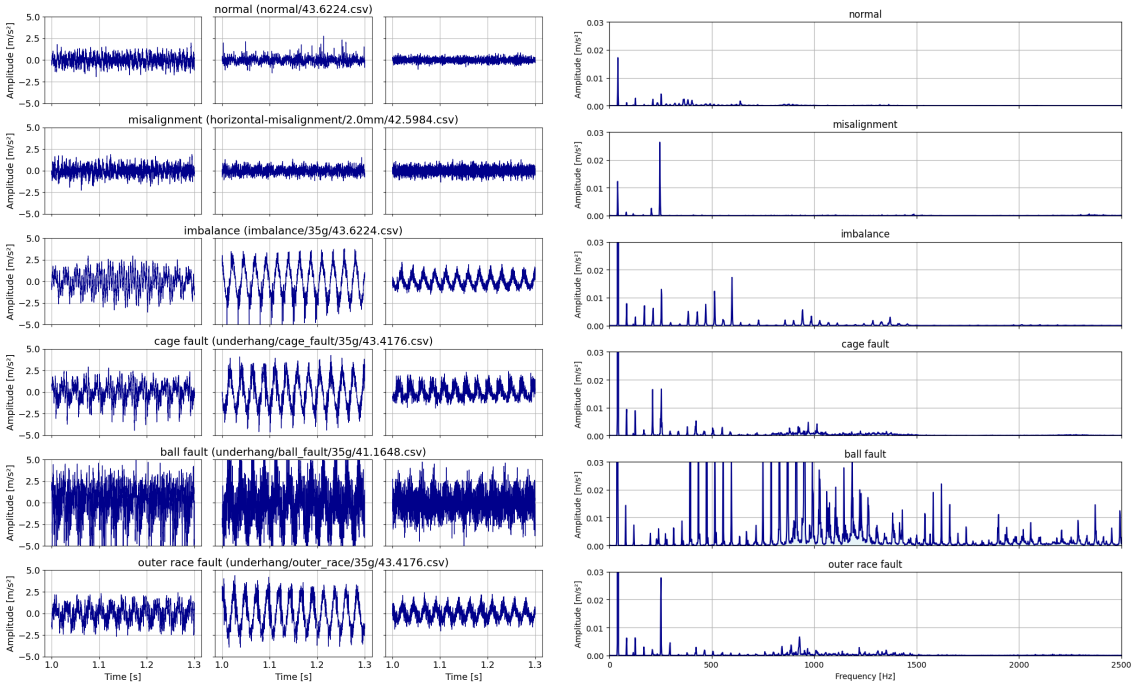
Out of the 1951 time series, the inner bearing (A) has 1438, and the outer bearing (B) has 1393 (Tab. 3.1). Source bearing A+B just concatanes observations from both postions so the column sums their counts together. The observations for shaft defects are shared regardless of the bearing.

Source bearing		A	A+B	B	A	A+B	B
High severity faults		No	No	No	Yes	Yes	Yes
Fault	**misalignment**	498	996	498	248	496	248
	imbalance	333	666	333	188	376	188
	cage fault	188	376	188	91	181	90
	ball fault	186	323	137	87	132	45
	outer race fault	184	372	188	86	176	90
	normal	49	98	49	738	1470	732
Σ		**1438**	**2831**	**1393**	**1438**	**2831**	**1393**

Table 3.1: Number of observation in MaFaulDa split by class label according to source bearing

The fusion of vertical and horizontal misalignment gives rise to the label that occurs the most at 34.63%. The baseline class is the least populous at 3.4% for the inner bearing. After relacing low severity faults, the normal status became the largest group with 51.32%. The imbalance ratio as a proportion of the largest to the smallest classes is 10.16 for original labels and 8.58 after relabeling.

Exemplary waveforms from each fault class illustrate their superficial differences in oscillatory behavior. The files with the highest severity levels and around rotational speed of 2500 rpm (42 Hz) are displayed to discern patterns.



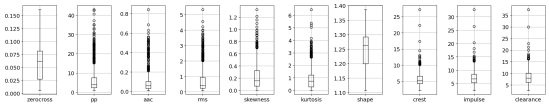
(a) Time domain for all axis (b) Frequency domain in radial direction

Figure 3.2: Vibrations from inner bearing (A) for every fault class with the highest fault severity at 2500 rpm

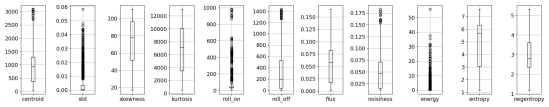
Time-domain waveforms of the 300 ms signal slice are shown in the graphs in Figure 3.2a. Subplots for radial, tangential, and axial directions are in columns from left to right. Amplitudes vary with limits from ± 3 m/s^2 for normal condition and misalignment, up to ± 11 m/s^2 in case of severe bearing faults.

The frequency spectrum in Figure 3.2b is obtained by FFT and Hann window of 16384 samples. The signal chunk represents an uncertainty box with a duration of approximately 328 ms and a spectral resolution of little over 3 Hz. The graph is cropped to make the most important peaks visible.

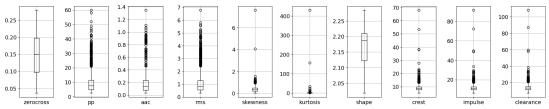
The numeric scales of distinct attributes calculated from inner bearing signals differ widely (Fig. 3.3). The magnitude of three spatial dimensions stretches the scale towards larger values or introduces more outliers. Many of the indicators used for fault detection are dimensionless numbers, but for others, the information about units of measurement is still attainable.



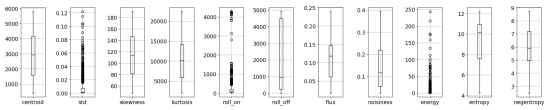
(a) Time-domain features for 1 axis



(b) Frequency-domain features for 1 axes



(c) Time-domain features for 3 axes

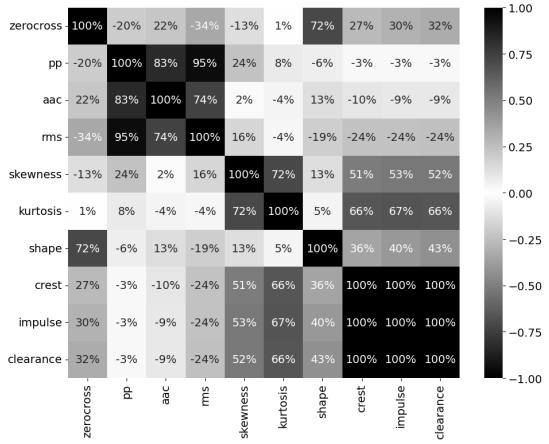


(d) Frequency-domain features for 3 axes

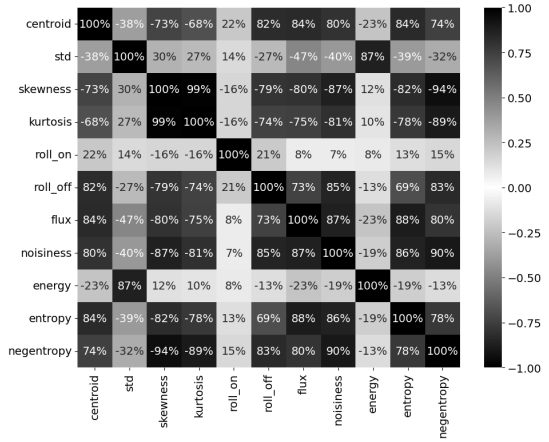
Figure 3.3: The value ranges of attributes for bearing A, depending on the number of directions that feature is aggregated out of

Inside the interquartile range of three axis features, the peak-to-peak is between 4.89 - 11.39 m/s^2, root-mean-square is 0.56 - 1.31 m/s^2, average amplitude change represents 0.08 - 0.23 m/s^2 of level jump in 20 μs. In the frequency domain, the spectral centroid is in the range of 1575 - 4175 Hz, roll-off frequencies arise above 246 and below 4476 Hz, roll-on frequencies are from 50 to 207 Hz, and entropy occurs from 7.57 to 10.94 nats.

Despite the differences in scales, in some instances by several orders of magnitude, there are still dependencies among variables in the complete feature sets. The TD set (Fig. 3.4a) appears to have less strongly correlated pairs than the FD set (Fig. 3.4b) in terms of Pearson correlation coefficient.



(a) Time-domain features



(b) Frequency-domain features

Figure 3.4: Pearson correlations of variables from inner bearing (A) in feature sets aggregated out of 3 axes

Crest, impulse, and clearance have correlations of 100% in TD. Other very strongly positively correlated variables are peak-to-peak to root-mean-square at 95%

and average amplitude change at 83%. Kurtosis is strongly correlated with skewness at 72% and with crest, impulse, and clearance at 66%. Shape and zero-crossing rate are also correlated at 72%. In the FD set, the roll-on frequency, energy, and standard deviation are the most uncorrelated to all other variables. However, the energy and standard deviation are highly correlated to each other. Skewness and kurtosis of the frequency spectrum have a strong association of 99%. The rest of the attributes have an absolute value of correlation greater than 68% in pairs amongst themselves.

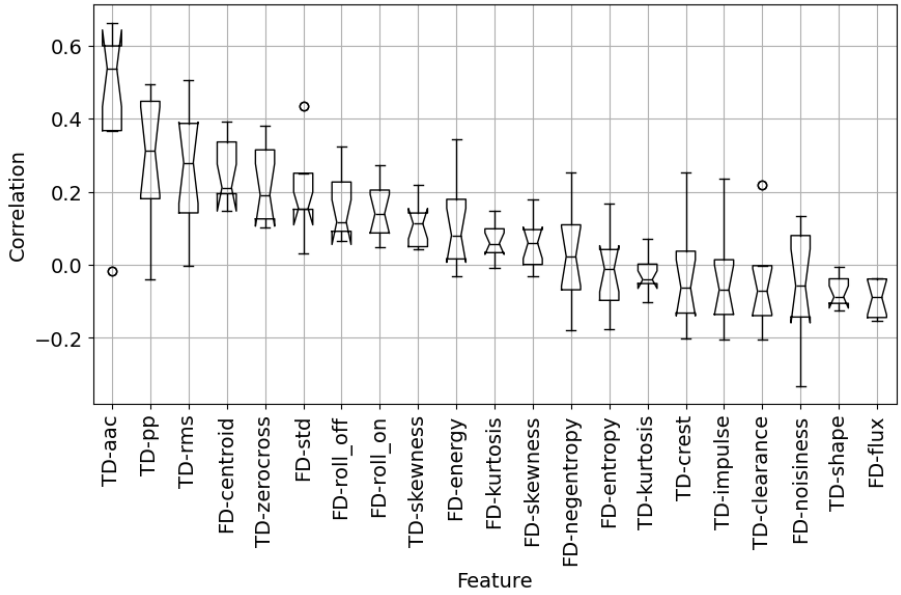
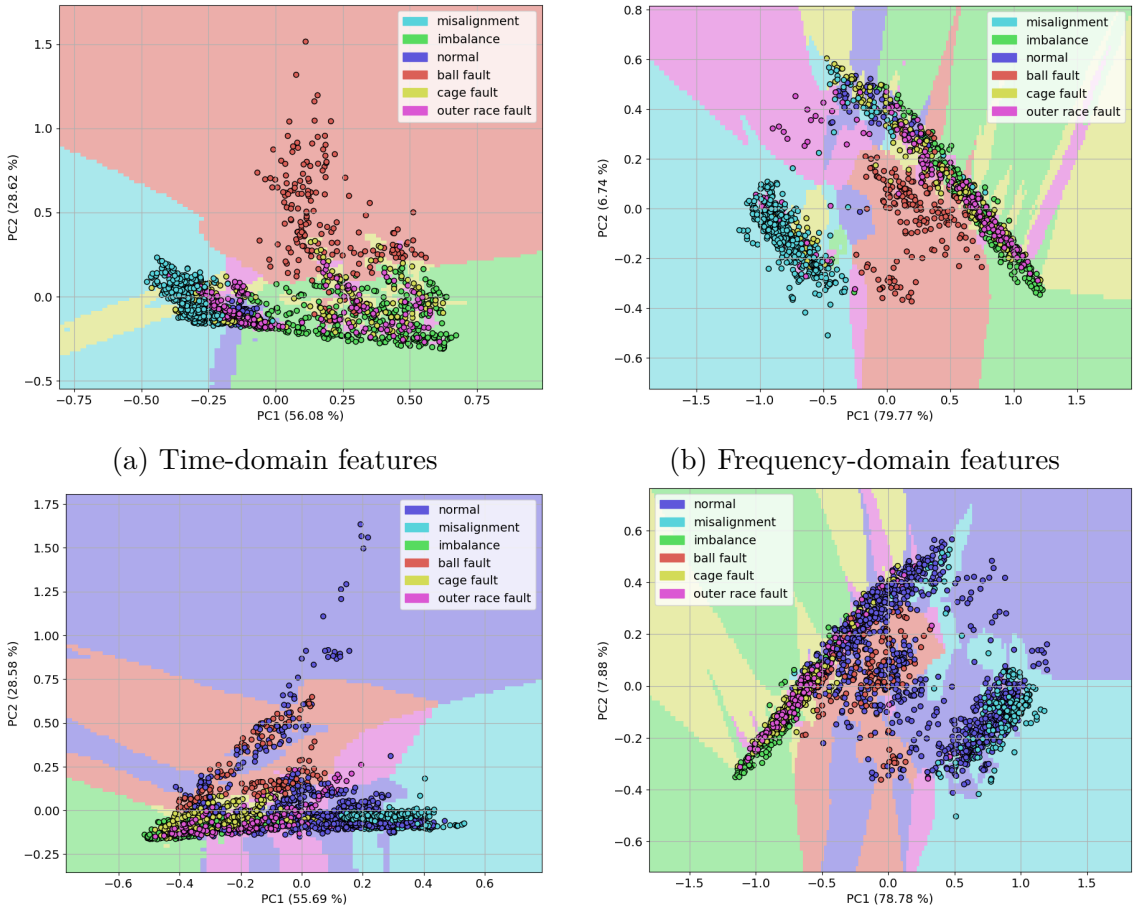


Figure 3.5: Correlation of features to rpm over all experiments

The faults in the MaFaulDa are observed at several rotational speeds. This increases the robustness of the machine-learning model. In practice, if the model is trained only at one speed and then the speed alters, the features tied to rpm could inhibit the prediction accuracy. The statistical distribution for correlations between predictors and rpm under all experimental conditions is in Figure 3.5.

An overwhelming majority of predictors are weakly correlated to rotational speed. The moderate correlation appears for average amplitude change with a median of 54% and peak-to-peak with a median of 31%. These two features are less suited for identifying fault types in variable speed situations.

Fault group separability into clusters is qualitatively analyzed by projecting normalized complete feature sets onto the plain utilizing two principal components.



(a) Time-domain features

(b) Frequency-domain features

(c) Time-domain features (severity)

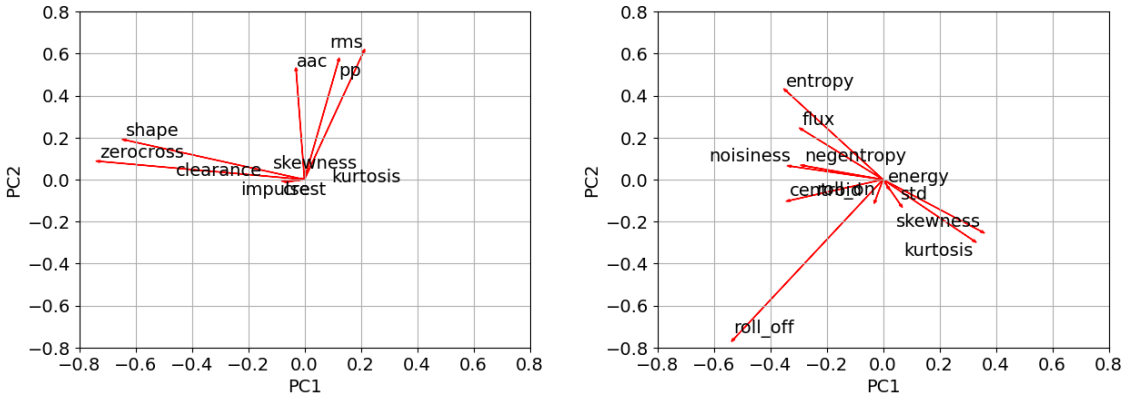
(d) Frequency-domain features (severity)

Figure 3.6: PCA of complete feature set into two principal components (bearing A and three axes

The transformed TD and FD sets are in Figure 3.6 with original labels and after low severities are assigned the normal class. The decision boundaries that color the feature space around the data points are determined in 0.02 *times*0.02 square areas by k-NN with five neighbours. We observe the inability to separate faults by linear boundaries in given feature spaces and the noncompactness of clusters.

Shaft misalignment and ball bearing fault look lumped together, whereas imbalance is interweaved with cage fault. The silhouette score quantifies the cluster overlap by a weak score of around 0.08. The explained variance with two principal components is 84.7% for TD and 86.5% with original labels.

The PCA loading plots (Fig. 3.7) illustrate correlations of features to two principal components. The first PC in the time domain mainly describes the impulsiveness of the waveform: *shape, impulse, crest, clearance, zero-crossing rate*. The second



(a) Time-domain features

(b) Frequency-domain features

Figure 3.7: PCA loading plots for bearing A

PC focuses more on the amplitude range: *rms, peak-to-peak, aac.*

However, the groups are not as clear-cut for frequency-domain features. Overall chaos in the frequency spectrum can be attributed to PC1: *flux, entropy, negentropy, noisiness*, and the shape of frequency distribution to PC2: *roll-on, roll-off.* PCA efficiently expresses attributes in less dimensional space, but the resulting linear combination is hard to comprehend for explaining decisions.

3.4 Accelerometer data logger

Data acquisition of vibrations from industrial machinery necessitates the assembly of the standalone data logger. The required hardware modules of the device include a triaxial MEMS accelerometer as a sensor, a fast persistent memory unit, and a microcontroller to control communication with both peripherals. The recording starts after a button press, and the indicator LED provides feedback that the device is recording.

Detailed specification according to standards makes demands especially on the accelerometer, in terms of bandwidth for linear response to at least 1 kHz, ideally more. The sampling rate should be greater than 2 kHz for each of the three directions (6 kSps). Out of the options available on the market, we opted for the accelerometers with digital communication bus, more than 12-bit sample resolution, high sensitivity, and low noise density.

The hardware configuration we arrived at for the data logger is in Table 3.2. The

Accelerometer	**IIS3DWB**
Vendor	STMicroelectronics
Bus	SPI (to 10 MHz)
Axis	1 or 3
Range	$\pm$ 2 - 16 g (19 - 157 m/s^2)
Bandwidth	5 - 6.3 kHz
Sensitivity	0.06 - 0.49 mg/LSB
Noise density	75 μg/$\sqrt{\text{Hz}}$ rms
Output data rate	26.8 kHz
Sample resolution	16 bit
FIFO	3 kB (512 samples)
Microcontroller	**ESP32-PoE-ISO**
CPU SoC	ESP32-WROOM-32
CPU clock rate	80 MHz

Table 3.2: Hardware parameters of accelerometer data logger

power bank powers the device via a MicroUSB connector. The IIS3DWB accelerometer is a cheap enough MEMS accelerometer that approaches industrial standards for vibration monitoring. The evaluation board STEVAL-MKI208V1K with the SMD accelerometer is connected via a ribbon cable to the microcontroller. This enables proper attachment of the sensor to designated places.

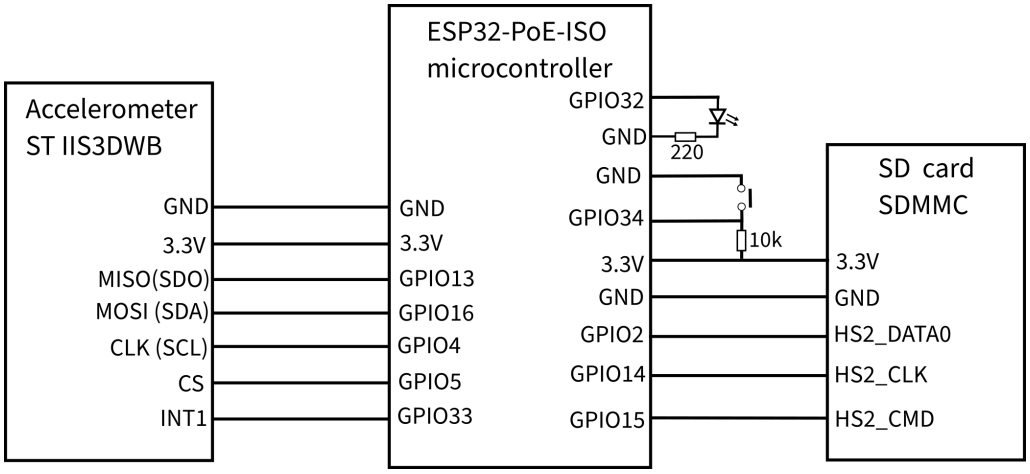


Figure 3.8: Data logger hardware block diagram

ESP32-PoE-ISO is chosen as a microcontroller development kit because it has an SD card slot connected via an SD/MMC bus. The accelerometer uses an SPI bus with a maximum speed of 10 MHz that can be connected to any physical GPIO

pin. The block diagram of the designed hardware device is in Figure 3.8.

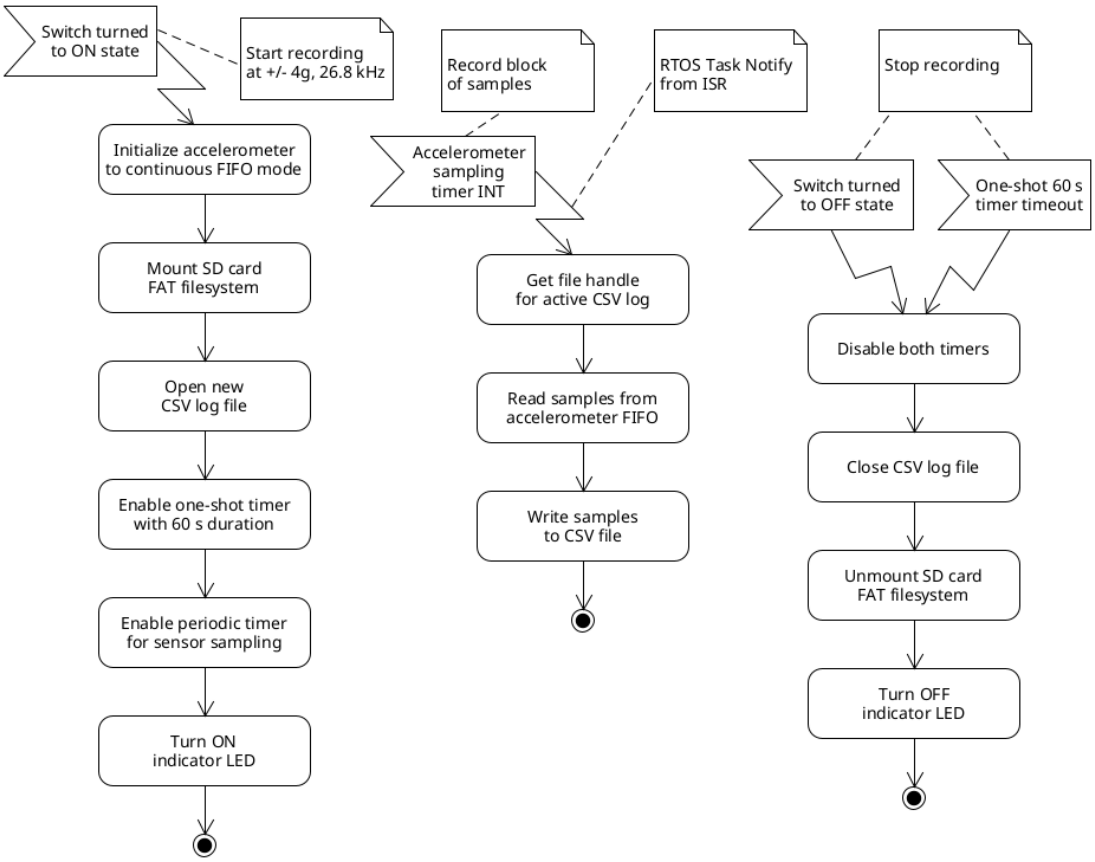


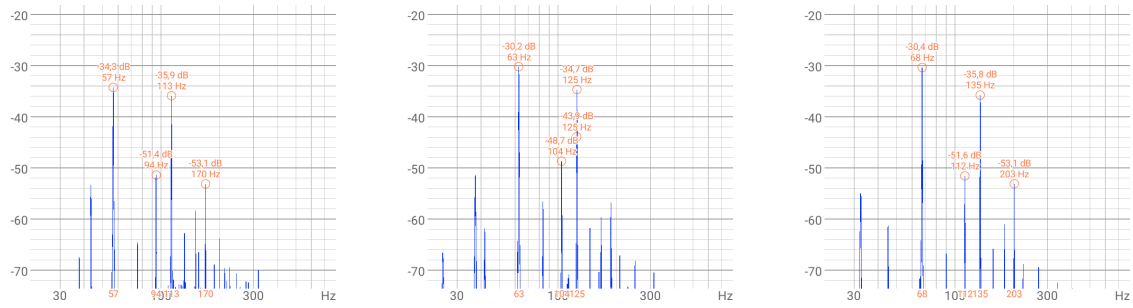
Figure 3.9: Activity diagram of data logger firmware functions

The firmware acts as a coordinator of hardware adapters and aids in straightforward vibration acquisition for later analysis. The provided functionality is consequently minimalistic. Two user-driven actions, one periodic event, and one timer-generated event satisfy these basic requirements. Activity diagrams outline the functionality of the firmware actions (Fig. 3.9).

3.5 Industrial equipment

Three kinds of rotating machines are at our disposal for vibration measurements. Those are a standing fan, scroll compressor, and water pump.

Standing fan is model *Kalorik TKG VT1037* (Fig. 3.11a) and one unit is available to us. It serves as a test bed during the data logger development. The accelerometer is glued to the plastic casing at the back of the drive motor. The fan



(a) Fan speed 1 (b) Fan speed 2 (c) Fan speed 3

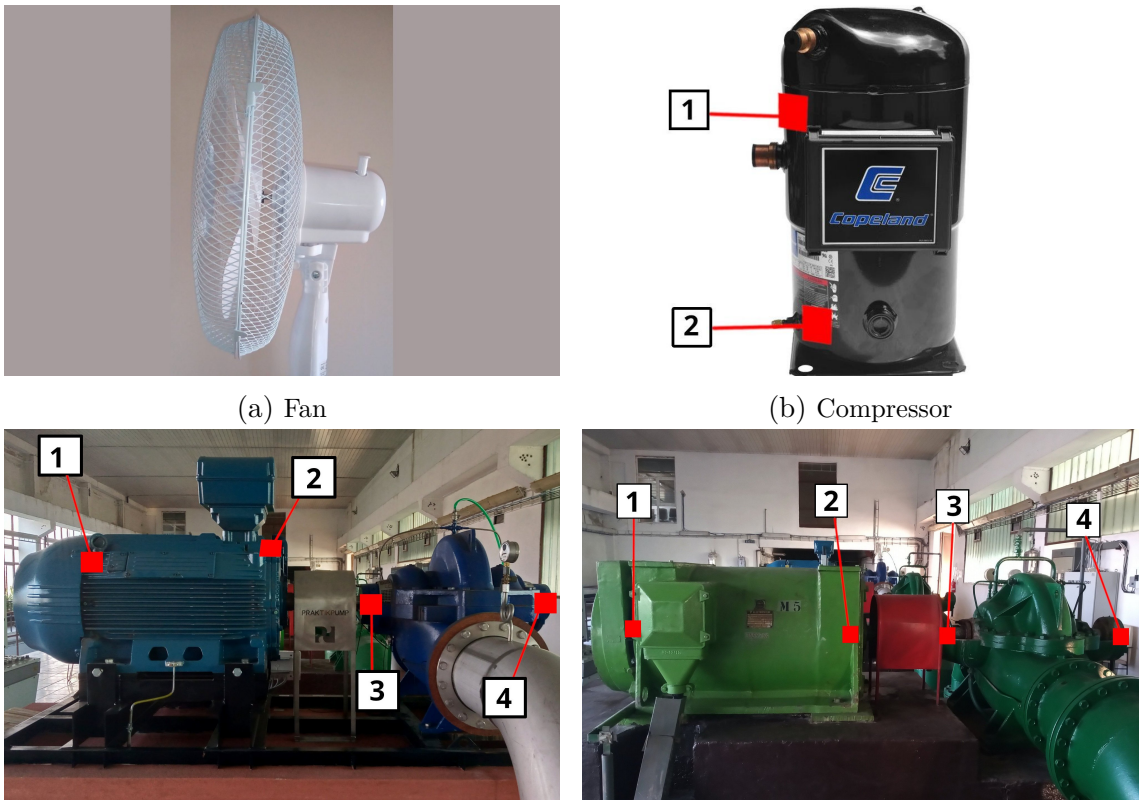
Figure 3.10: Frequency spectrum of audio recorded in close proximity to the side of the standing fan

blades have a 45 cm diameter with 3 propellers and a power of 45 W (class I). It has a switch for three rotational speeds of approximately 19 Hz, 21 Hz, and 22.7 Hz ($\Delta f = 0.18$ Hz). Speed was estimated by spectral analysis of audio (Fig. 3.10) in *Spectroid* app and confirmed by 240 fps high-speed camera.

Scroll compressor is model *Copeland ZR16* (Fig. 3.11b). In the data center, two units are part of independent air conditioning units. The compressor has 9.7 kW of power (class I) and rotates at 2900 rpm (48.3 Hz). Possible measurement placements are located on the sides of the compressor atop the bearings, just above the base and below the scroll. Steel casing is not in direct contact with the bearing, which causes significant alteration to the signal.

Water pumps are available as three units in municipal drinking water pumping station. The apparatus consists of a single-stage axially split volute casing pump and an attached electric induction motor. The pump and motor have each two bearings, hence there are four measurement positions.

The newer primary pumps, commissioned in 2018, with bundled wireless cloud monitoring are two units of *KSB Omega 300-560* (Fig. 3.11c). The pumps and motors rotate at 1493 rpm (24.9 Hz). The motor *WEG W50* provides 400 kW of power (class III). The bearing designations at numbered places for bearing defect frequency calculations are 6319-C3 (1), 6324-C3 (2), and 6317-2Z (3 & 4). The secondary pump is one unit named *Sigma 300-OVD-600* (Fig. 3.11d) installed in 1986. The Sigma pump rotates at 1485 rpm (24.75 Hz), and its electric motor has a power of 450 kW (class III).



(a) Fan (b) Compressor

(c) Water pump KSB (d) Water pump Sigma

Figure 3.11: Machines dedicated for vibration measurements

3.6 Data collection methodology

Vibration measurement for one placement on the machine involves three trials (two for Sigma pump). After each trial, the sensor is detached and attached again to introduce "noise". The sensor is mounted to the machine on a flat surface with adhesive of thin double-sided carpet tape.

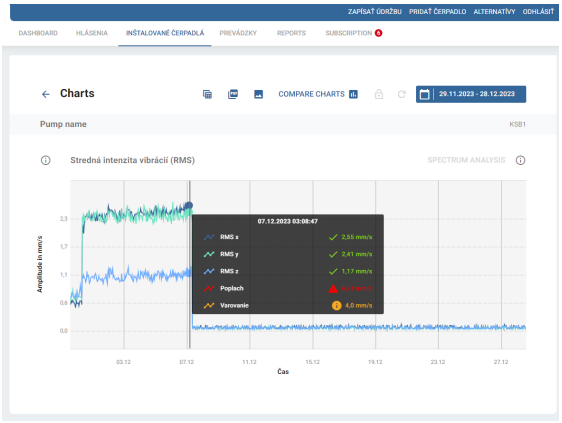
Sensor placements are marked by single digit numbers for simplicity (Fig. 3.11) but the precise description follows the notation of MIMOSA convention from ISO 13373-1. The accelerometer is mounted on compressor's positions: *SFTA001AT000TN* (1), *SFTA002AT000TN* (2), on WEG motor's positions: *MTRA001AT000TN* (1), *MTRA002AT045TN* (2), and on KSB pump positions: *PMPA003AT000TN* (3), *PMPA004AT000TN* (4).

The triaxial recording has a duration 60 s (at $f_s = 26.8$ kHz) and a 16-bit resolution. One file has 24.6 MiB as a binary stream and 48 MiB in TSV format. Measurements were executed and organized into folders according to schedule in

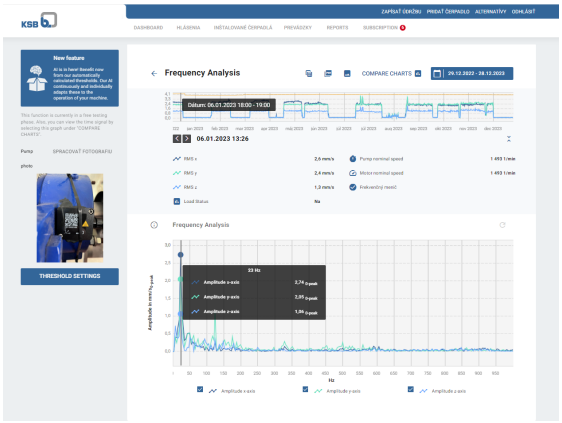
Machine		**Scroll compressor**		**Water pump**			**Pump's motor**		
Facility code		K3	K5	KSB-1	KSB-7	Sigma-5	KSB-1	KSB-7	Sigma-5
Label		C1	C2	P1	P2	P3	M1	M2	M3
Date	**20/02/2024**	×	×						
	27/02/2024			×	×		×	×	
	05/03/2024	×	×						
	19/03/2024	×	×						
	26/03/2024			×	×	×	×	×	×

Table 3.3: Data collection schedule

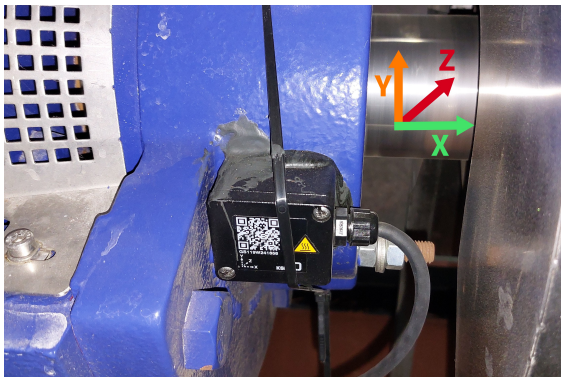
Table 3.3. It should be noted that because of operational regulations, the second KSB pump was measured the day after the first. The estimated total space requirements for all 92 uncompressed raw recordings is 2.21 GiB and approximately 4.31 GiB as TSV files (Tab-separated values). The dataset shall be called *Pump dataset*.



(a) Vibration rms velocity dashboard



(b) Vibration frequency analysis dashboard



(c) KSB Guard Sensor Unit

Figure 3.12: KSB Guard cloud monitoring for pumps

Logs from **KSB Guard** cloud monitoring tool contain vibration rms velocities and frequency spectra for two monitored KSB water pumps (Figure 3.12). We

were able to export the historical measurements of vibration velocity for last year at hourly intervals and sample frequency spectra. The sensor unit (Fig. 3.12c) is permanently attached to KSB pumps on bearing in position no. 3.

3.7 Data volume savings

The apparent advantage of feature discovery is reducing the amount of data downstream. Data compression must occur on edge devices to enable the utilization of wireless low-power wide area networks (LPWAN). The protocol stack may differ, so goodput in this section is compared without node configuration metadata and keepalive messages.

The machinery monitoring system relies on determining a few parameters:

- **Number of source channels** (S) - comprises the number of monitored machines, measurement locations for sensors, and active sensor axes.
- **Sampling frequency** (f_s) - is set based on the linear response of the accelerometer, the types of faults intended for detection, and how soon they should be noticed after they arise. The higher required sensitivity means a higher sampling rate derived according to the Nyquist theorem. At a minimum, it should be 15 kHz to 20 kHz.
- **Interval between successive measurements** (T) - specifies the minimal response time to sudden failure. The more critical the machine is, the shorter the interval should be. The bigger the machine parts, the slower the defect evolves.
- **Duration of valid recording** (D) - is the captured snapshot of machine unaltered behavior associated with a timestamp. Duration should cover at least twelve windows for spectral estimation.
- **Number of extracted features** (F) - are ideally key trend indicators pointing to symptoms of common malfunctions. We aim for a total of six features.

Equation 3.2 expresses the lossy compression ratio ($\mathcal{C}$) formula if trend indicators are stored instead of full recording. The number of raw channels (S_{in}) can differ from

those extracted in features (S_{out}). Parameter $D = 0.5$ when we use frequency bins with 1 Hz resolution.

$$\mathcal{C} = \frac{D \cdot f_s \cdot S_{\text{in}}}{F \cdot S_{\text{out}}} \tag{3.2}$$

Compression ratio for MaFaulDa dataset compared to all 21 extracted features in 3 dimensions is 2381:1. With just six features, the compression is 25000:1, which is a saving of the original data by 99.996%.

As an example to approximate required network goodput and storage in practice, we consider continuous vibration **monitoring for municipal water pumping station**. The station has three pumps and three electric motors.

A pump and motor pair have four bearings together for drive end and non-drive end positions. Each position has a sensor mounted in three directions that makes a total of *36 source channels*. The sampling frequency at each position is *20 kHz*. The recordings have *five seconds* and are triggered regularly every 1 hour (*8760 times per year*).

In a year, the system gathers 31.54 Gs (gigasamples), which is 58.74 GiB with a 16-bit ADC resolution. Reasonably precise spectral estimation with 10 thousand bins needs 3.15 Gs per year. On the other hand, six features out of each channel keep only 1.89 Ms per year for a lossy compression ratio of 16667:1. Low data volumes potentially enable feature selection and models to be offloaded directly to edge devices. The entire history of machines' health can be saved into a small flash memory module.

4 Implementation

The tools we implement include the exploration of datasets using statistical overviews and visualizations, feature selection experiments that involve a machine learning pipeline, and firmware for the data logger's hardware.

4.1 Data analysis

The data processing and data mining on the MaFaulDa and Pump dataset takes place in several *JupyterLab* notebooks written in *Python* language. The purpose of individual notebooks is described in detail in Appendix A. Utility functions are located in separate package *vibrodiagnostics* so that the experiments can be realized under multiple conditions.

The tabular data is handled using *Pandas* dataframes that are training batch models from *scikit-learn* library and *imbalanced-learn*, or online models from *RiverML*. The wide variety of graphs and other visualizations are stylized with *Matplotlib* and *Seaborn*. Feature calculation is crafted according to mathematical formulas atop libraries *Numpy*, *SciPy*, and *Time Series Feature Extraction Library* (TSFEL).

```
METRICS = {
    "corr": selection.corr_classif,
    "f_stat": sklearn.feature_selection.f_classif,
    "mi": sklearn.feature_selection.mutual_info_classif }
r = pandas.DataFrame()
for name, metric in METRICS.items():
    # Order the features to leaderboard
    r[name] = leaderboard
ranks = r.rank(axis="rows", method="first", ascending=False)
return ranks.apply(scipy.stats.gmean, axis=1).sort_values()
```

Source code 4.1: Rank product of feature matrix X to label column Y

```
pandas.DataFrame(zip(X.columns, metric(X, Y)),
                 columns=["feature", "score"])
      .set_index("feature")
      .sort_values(by="score", ascending=False))
```

Source code 4.2: Leaderboard of feature importance metric scores

Since the central focus of this work targets feature selection, the source code listing 4.1 and 4.1 demonstrates the ranking scores calculated for features in data frame X. Leaderboards are united by geometric mean to produce rank product ordering.

4.2 Firmware

The drivers for low-level interfaces of ESP32 microcontroller and FAT32 filesystem are already available in Espressif ESP-IDF SDK. The accelerometer driver is made available by the vendor[1]. The entry routine of the firmware mounts the SD card, sets up GPIO pins for LED and button, and executes three *FreeRTOS* tasks. Procedures in firmware are described in Appendix A. The hardware (Fig. 4.1) was built by the thesis consultant based on the detailed specification supplied by the author.

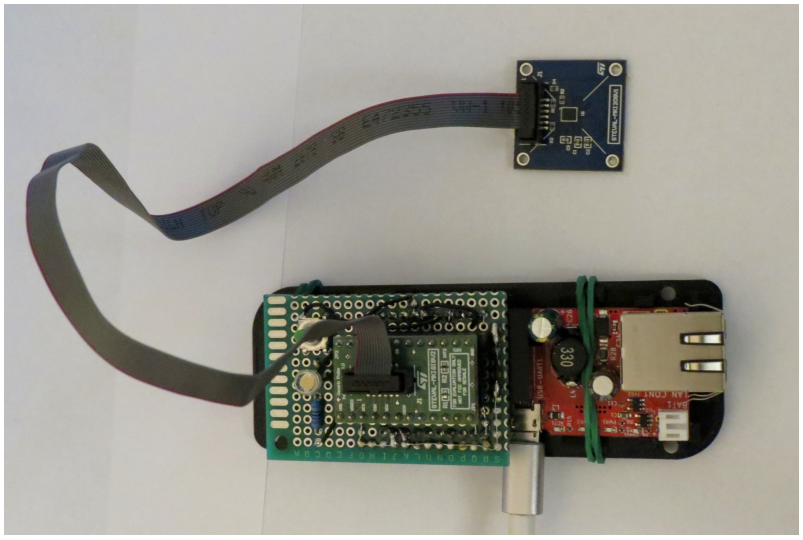


Figure 4.1: Accelerometer Data Logger

The tasks run on an event-driven basis through notifications and a queue. They have the following purposes:

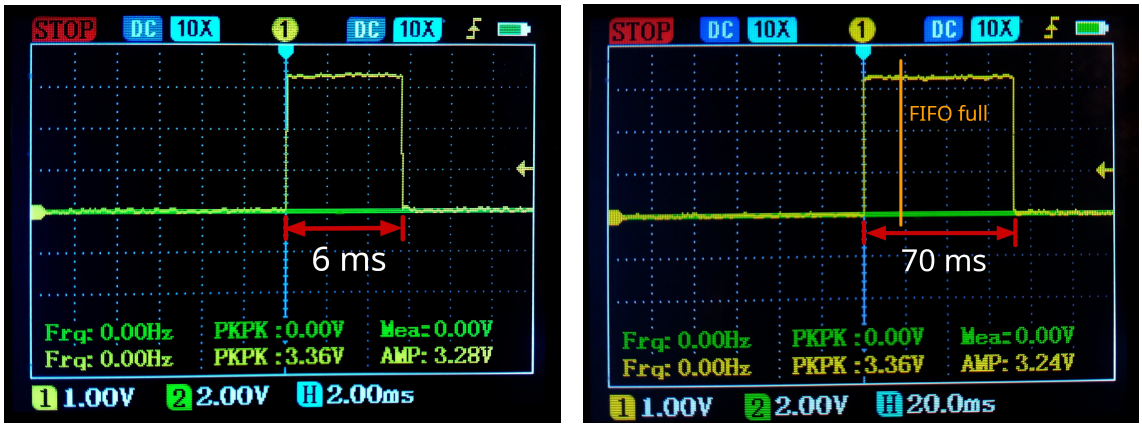
- **Trigger task** - reacts to notification from button's interrupt handler. Depending on whether the recording is in progress, it either starts peripherals and creates a file or stops them and closes the file. Button debouncing has the

[1]IIS3DWB driver: `https://github.com/STMicroelectronics/iis3dwb-pid`

form of a two-second delay when interrupts are ignored. The task is pinned to core 1 with priority 2 (higher numbers have more priority).

- **Read task** - waits for notification from 9 ms periodic timer to read around half of the accelerometer's FIFO via half-duplex SPI bus at 8 MHz. It sends read-out samples to the queue. In case of any buffer overrun, it turns off the LED prematurely. The task is pinned to core 0 with priority 1.
- **Write task** - reads the samples from the queue, and after locking the mutex for the opened file, it writes them to the card. The file is also manipulated within the trigger task, so the lock prevents race conditions. The task is pinned to core 1 with priority 1.

The testing process of the firmware revealed two issues that were resolved subsequently. The STEVAL evaluation board is plagued with broken hardware interrupt lines. The FIFO watermark interrupt on the INT1 pin of the accelerometer stopped firing after a few seconds and stayed at a high logic level. We noted the problem first on the digital multimeter, then confirmed it on an oscilloscope and found the same issue in the vendor's forum thread[2].



(a) Binary format (b) TSV format

Figure 4.2: Timing issue of writing samples to SD card captured on oscilloscope

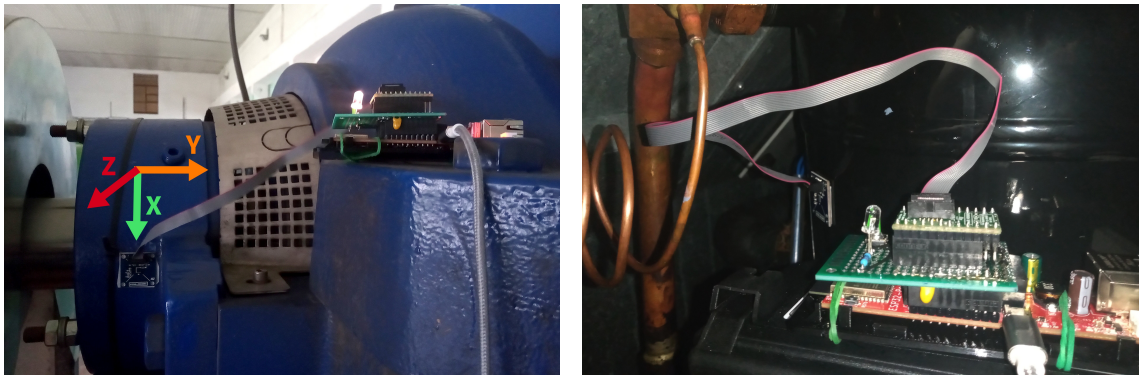
Slow string formatting with *printf* family of functions during the accelerometer sampling presented another impediment. The readings cannot be written to TSV

[2]IIS3DWB Interrupt stops triggering while sampling for long periods: `https://community.st.com/t5/mems-sensors/iis3dwb-interrupt-stops-triggering-while-sampling-for-long/td-p/203630`

directly because the real-time (RT) deadline is not met due to the high sampling rate. This is true even if the queue is utilized and the processing takes place in another task. The binary files from the SD card are converted in bulk to TSV with custom Python script `bin2tsv.py`.

During debugging, the pin was set to change the output level immediately before the one buffer started file write and after its completion. Waveforms in Figure 4.2 show the time it took to write in binary format compared to the TSV format. The relation of already processed buffers x to those left to process y is $y = 3.7x$ for TSV format. It is a monotonically increasing function, so the system would eventually overrun queue of any size.

The hardware FIFO of 512 sample vectors fills up in 19.05 ms. The processing time t has to satisfy inequality for deadline: $t \leqslant 2 \cdot \frac{1}{f_s} \cdot N_{FIFO}$. The RT deadline is 38.1 ms, beyond which the samples get dropped. In an embedded device, the short recoding burst (80.4 kSps) would need to be written temporarily in binary format to external memory and processed in two passes.



(a) Sensor on water pump (b) Sensor on scroll compressor

Figure 4.3: Data logger on the machines in measurement positions

Measurement conditions in the industry and axis orientation of the accelerometer are depicted in Figure 4.3. The sensor is attached by tape to machines with a connector facing up and is pressed gently toward the surface.

5 Evaluation

The verification of proposed solutions for machinery fault diagnostics is focused on two activities, these are vibration measurement and defect identification. The accuracy of supervised learning using the k-nearest neighbour classifier is determined by the MaFaulDa dataset under various experiments. Data logger recording is compared to the known reference, and the collected Pump dataset is analyzed.

5.1 Fault classifiction in MaFaulDa

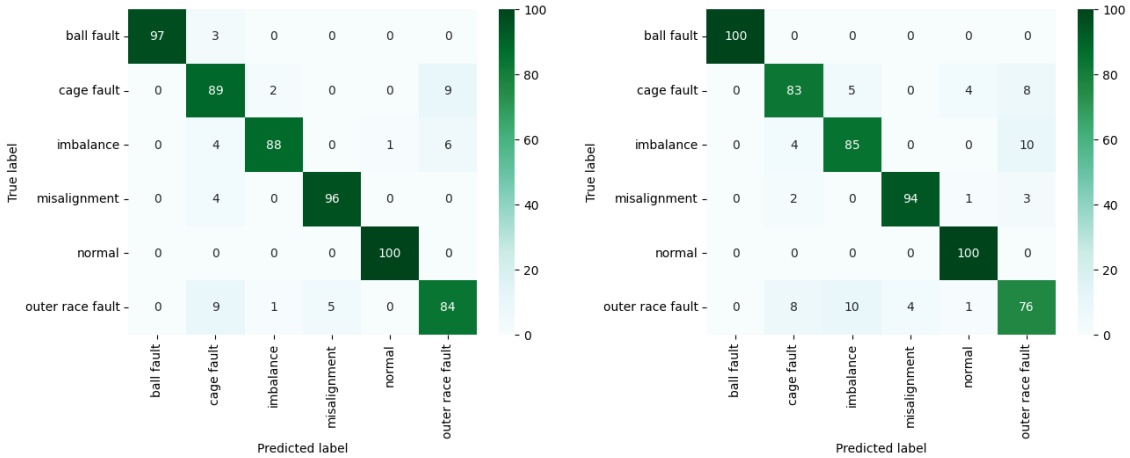
The four outlined experiments on MaFaulDa involve testing sets of features with very few members. The effect in predictions is observed when reducing the information content about machine's status to the minimum. First, the attributes are left in full after extraction. Then, all of their combinations are enumerated, and the resulting accuracy distribution is matched against accuracy after feature selection with similarity metrics.

5.1.1 Complete feature sets

The two full sets of features include ten extracted from time-domain and eleven from frequency-domain of the vibrations. The feature spaces can incorrectly interchange different fault labels and better separate out some groups than the others.

The inner bearing observations are selected to train the k-NN with five neighbours and Euclidean distance metric. The attributes are normalized beforehand, rows are oversampled to a majority label, and data is split into training and testing sets with an 80:20 proportion. The 598 observation of validation data determines the confusion matrix (Fig. 5.1).

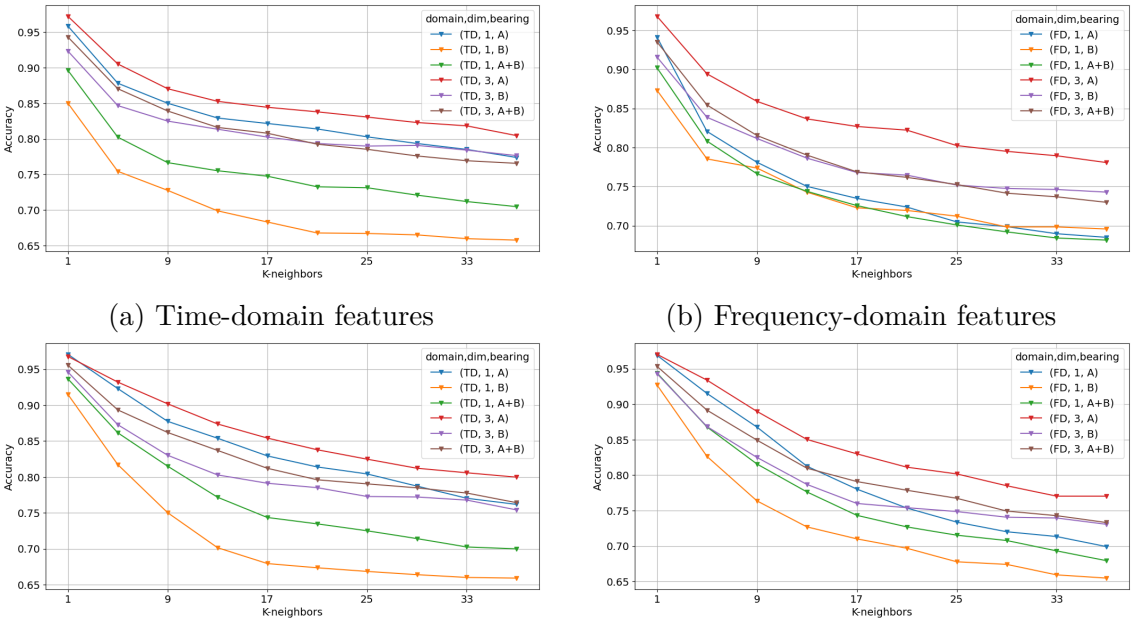
The label "normal" is not falsely attributed to other classes in either feature set, but other classes can get assigned to be "normal". Most mistakes happen while



(a) Time-domain features (b) Frequency-domain features

Figure 5.1: Confusion matrix for complete sets of features

predicting outer race fault, which gets confused with cage fault, imbalance, and less often with misalignment. The shaft imbalance is applied to simulate bearing faults, which is a natural reason for this high error rate.



(a) Time-domain features (b) Frequency-domain features

(c) Time-domain features (severity) (d) Frequency-domain features (severity)

Figure 5.2: Accuracy on complete feature sets depending on the k-value

An increase in the number of neighbours used for classification in five-fold cross-validated k-NN shows a substantial decrease in accuracy on validation sets (Fig. 5.2). The most prominent drop in performance of around 10% occurs at the beginning until the k-value of nine, and then the accuracy curve slowly plateaus.

Under every circumstance, the magnitude of the triaxial feature vector reaches better accuracy than those from only the axis of motion for the same source domain and bearing. The model for inner bearing A is more accurate than outer bearing B. The TD set is generally better in predictions than the FD set for the equivalent k-value. The dataset relabeled for high severity has a steeper decrease in accuracy for the same number of neighbours.

5.1.2 Feature subset combinations

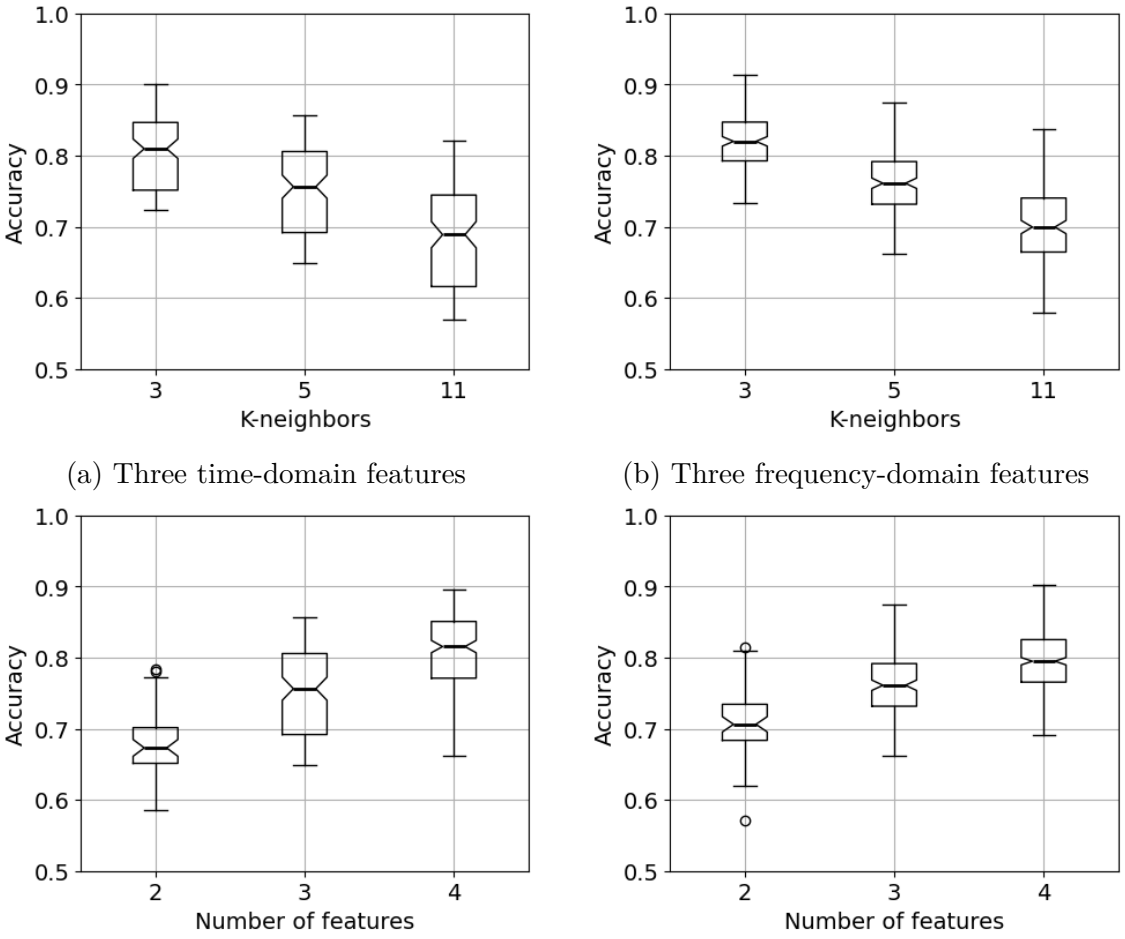
The complete sets of predictors are even greatly shrunken to representation that could be presented in 3D plot or in perpendicular cross sections. These trend variables could be used in distinguishing faults the same way as rms amplitude indicates their presence. Each possible combinations of pairs, triplets, and quadruplets construct a separate k-NN model on which prediction accuracy is evaluated.

The distribution of model accuracies is documented on features in both domains coined from three dimensions on bearing A with original and high-severity defect labels. The boxplots display the relation of the k-value to accuracy with three features, and the relation of number of features to accuracy with five neighbours (Fig. 5.4 and 5.4).

The decrease in accuracy with additional neighbours is apparent and similar to the trend in complete sets of features. We are interested in comparing maximum accuracies of the accuracy distribution because the optimal feature selection method tries to approach them. The reduction in maximal accuracy is more noticeable between three and five features of 3% to 5%, and almost the same amount between five and eleven neighbours.

The complete sets reach better accuracy than subsets when the number of features is at most three and simultaneously the number of neighbours is five or less. For more triaxial features and more neighbours, the complete set is only about 2% percent worse at the most than subsets because the curse of dimensionality is not so substantial for ten dimensions.

The spread in model accuracies in the interquartile range is from 5 to 10%, and measured between whiskers is 25% at a maximum. The standard deviation is about



(a) Three time-domain features

(b) Three frequency-domain features

(c) Five neighbours in time domain

(d) Five neighbours in frequency domain

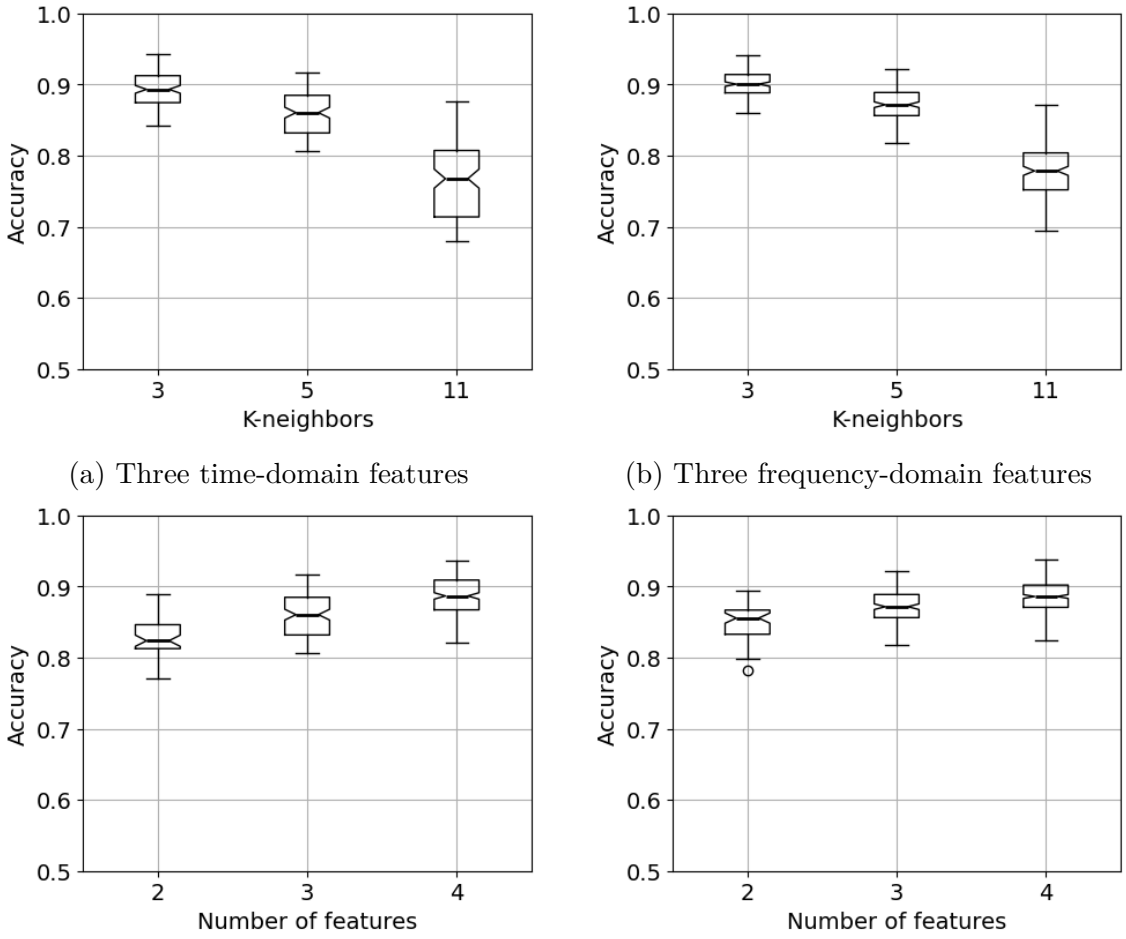
Figure 5.3: Model accuracy distribution for bearing A and three axis features

7%. Overall, the time-domain features are better than frequency-domain features for these specific feature spaces.

The number of features has a direct proportionality effect on the optimal model accuracy. An increase from two to three features has more weight than allowing a fourth attribute. The contributions of adding features are around 6% and 3%, and the relabeled dataset has an increase of 3% and 2%. The absolute accuracies are consecutively for 2, 3, and 4 features when k is equal to 5: 78.31%, 85.67%, and 89.55% for the time domain and 81.52%, 87.52%, 90.26% for the frequency domain.

5.1.3 Feature selection techniques

The predictors chosen with supervised selection strategies are compared to the accuracy distribution of enumerated combinations of same-sized attribute sets and the



(a) Three time-domain features

(b) Three frequency-domain features

(c) Five neighbours in time domain

(d) Five neighbours in frequency domain

Figure 5.4: Model accuracy distribution from bearing A and three axis features after relabeling for high severity faults

performance of its source superset. The metrics for bivariate feature selection are correlation, F statistic, mutual information, and their ensemble by the rank product. The PCA of the complete feature set that retains the same number of features as a selection method is a benchmark to tell whether the linear combination or meaningful attributes get better performance on the MaFaulDa.

Table 5.1 compares the concrete case of choosing the three features from each domain on bearing A and k-NN with five neighbours. Table 5.2 uses labels for high-severity faults. There is a significant difference in train and test accuracy, which means the model is likely overfitting. The percentile within the distribution is measured to its respective data, the train or validation set. However, the percentile of best features in the test data is calculated against distribution of the training data.

Feature set	Accuracy		Percentile		Domain	Best features
	Train	Test	Train	Test		
All features	96.03	92.80	100.00	100.00	TD	
PCA PC	91.20	84.67	95.00	93.33	TD	
Best features	91.93	85.47	100.00	99.17	TD	zerocross, pp, skewness
Rank product	91.21	85.04	95.83	97.50	TD	zerocross, shape, rms
Correlation	91.21	85.04	95.83	97.50	TD	shape, zerocross, rms
F statistic	90.59	84.07	91.67	90.00	TD	rms, pp, zerocross
Mutual information	88.24	80.62	75.83	76.67	TD	zerocross, shape, crest
All features	93.67	88.45	100.00	100.00	FD	
PCA PC	86.76	78.51	64.85	70.91	FD	
Best features	92.86	87.52	100.00	100.00	FD	centroid, roll_off, entropy
Rank product	85.79	77.18	51.52	57.58	FD	roll_off, flux, skewness
Correlation	85.79	77.18	51.52	57.58	FD	roll_off, skewness, flux
F statistic	85.79	77.18	51.52	57.58	FD	roll_off, flux, skewness
Mutual information	90.73	83.60	94.55	94.55	FD	roll_off, entropy, noisiness

Table 5.1: Feature selection method accuracy and percentile within accuracy distribution of all three member subsets. (bearing = A, dimension = 3, k=5)

Combining the rankings from several metrics is necessary to get consistent results. This is evidenced by the variability in the success of selected feature sets in final prediction performance under multiple conditions. The PCA with three components for the complete feature sets is comparable in accuracy to selection methods with original attributes.

The triplet of variables with the best results is in TD set: zero-crossing rate, peak-to-peak, skewness, and in FD set: centroid, roll-off, and entropy. With high severity labels, the best attributes for FD stay the same, but average amplitude change and shape factor are preferred along the zero-crossing rate. The rank product picked up roll-off, flux, and skewness for the FD set, which is suboptimal. The two of the three methods in the ensemble arrive at the same set overruling the superior set produced by mutual information. In the TD set, the zero-crossing rate, shape, and rms, are chosen by rank product.

The entire accuracy distribution for the training and testing set with original labels is drawn as a histogram (Fig. 5.5). The results of the chosen predictors are mapped out onto the distribution as vertical lines that stack up near the maximum for the time domain or disperse slightly above the median for the frequency domain. The accuracy of all features is unreachable for three feature subsets. Another noticeable difference between distributions is their shift down and greater spread in

Feature set	Accuracy		Percentile		Domain	Best features
	Train	Test	Train	Test		
All features	96.42	94.76	100.00	100.00	TD	
PCA PC	94.63	92.07	100.00	100.00	TD	
Best features	94.58	91.71	100.00	99.17	TD	zerocross, aac, shape
Rank product	94.31	91.40	98.33	95.83	TD	zerocross, shape, rms
Correlation	94.31	91.40	98.33	95.83	TD	zerocross, shape, rms
F statistic	94.31	91.40	98.33	95.83	TD	shape, zerocross, rms
Mutual information	91.90	88.51	72.50	76.67	TD	zerocross, shape, clearance
All features	95.20	93.20	100.00	100.00	FD	
PCA PC	92.14	88.84	69.09	73.94	FD	
Best features	94.64	91.94	100.00	99.39	FD	centroid, roll_off, entropy
Rank product	93.89	91.24	95.15	97.58	FD	entropy, noisiness, centroid
Correlation	94.50	92.19	99.39	100.00	FD	entropy, centroid, flux
F statistic	94.50	92.19	99.39	100.00	FD	entropy, flux, centroid
Mutual information	93.32	90.67	90.91	91.52	FD	noisiness, roll_off, entropy

Table 5.2: Feature selection method accuracy and percentile within accuracy distribution of all three member subsets. (severity, bearing = A, dimension = 3, k=5)

Method	Median percentile [%]	Median accuracy [%]
Rank product	88.97	79.82
Mutual information	91.81	80.87
F statistic	86.90	79.63
Correlation	84.49	79.04

Table 5.3: Feature selection methods evaluated in summary over all experimental conditions

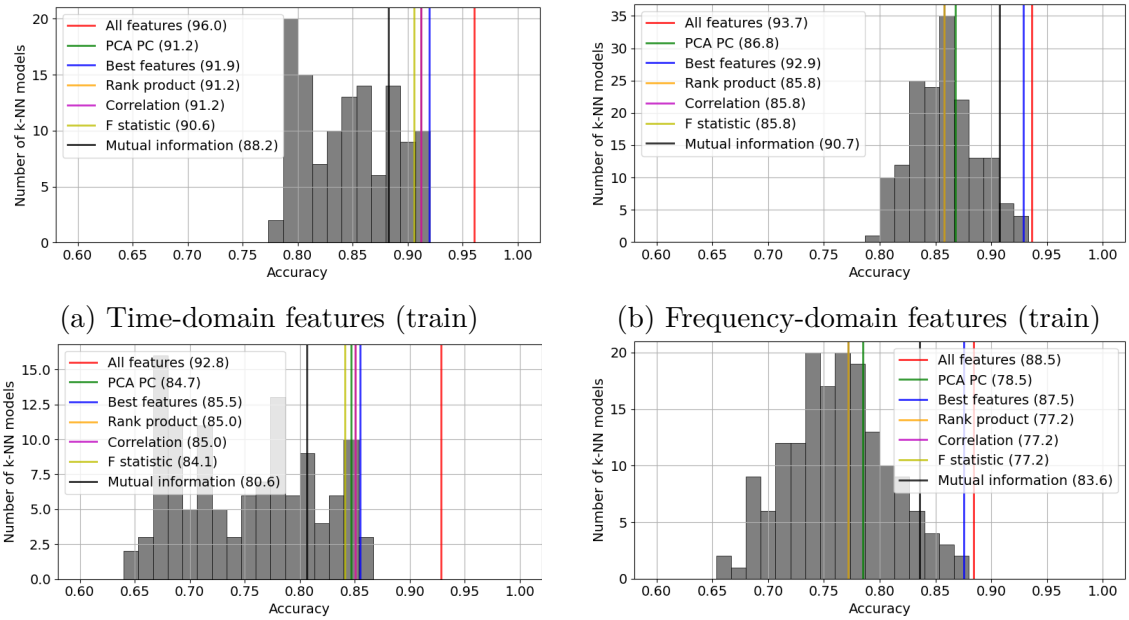
testing sets compared to training sets.

In 216 scenarios, the mutual information has better median accuracy (80.87%) and distribution percentile (91.81%) followed by rank product with an accuracy of 79.82% of and percentile of 88.97% (Tab. 5.3). The scenarios are composed of 24 base dataset modifications and options for hyperparameters k-value and number of features.

The rank product is the best method in the majority of 43.52% scenarios. Mutual

	Best in scenarios	Scenarios [%]	Mean percentile [%]
Rank product	94	43.52	92.38
Mutual information	87	40.28	91.79
Correlation	26	12.04	97.54
F statistic	9	4.16	96.10
Σ	216	100	-

Table 5.4: The experimental scenarios in which the method is found to be the best



(a) Time-domain features (train)

(b) Frequency-domain features (train)

(c) Time-domain features (test)

(d) Frequency-domain features (test)

Figure 5.5: Model accuracy statistical distributions with feature selection methods for three predictors (k = 5)

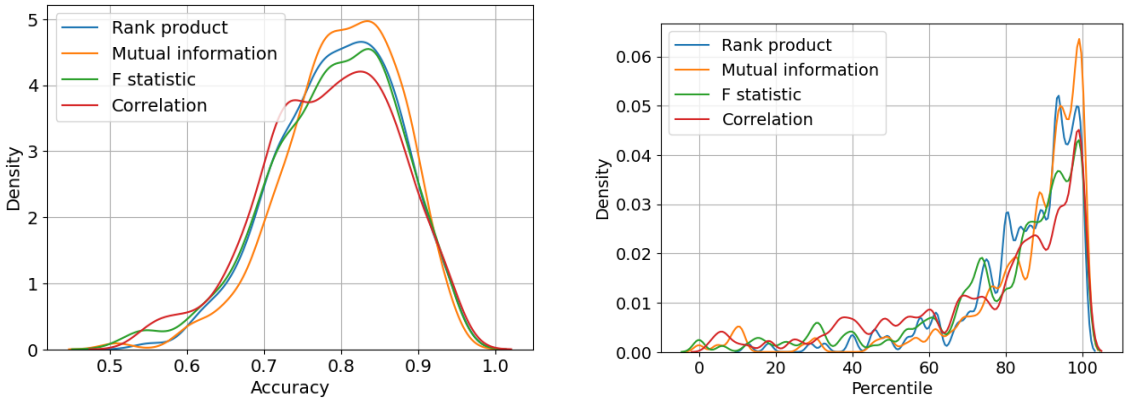
information comes second in 40.28% cases, where it is deemed the best strategy (Tab. 5.4). The median accuracy, in cases where they are the best method, is also better for rank product 92.38% compared to 91.79% mutual information.

Kernel density estimate plot (Fig. 5.6) shows the distribution of accuracies for the selection methods and percentiles for predictor subsets they choose. The feature selection usually picks variables so that they stay in the upper quartile of the distribution above the 75% percentile. The median accuracy of all methods is the vicinity of 80%.

The features chosen by rank product are visualized as a three-dimensional scatter plot. The colors of data points represent correct labels with ex marks for misses. The scales on graph axes are inverse transformed of the min-max scaler. The visualization of predicted groups suggests another transform should be applied to even out the distances and handle the outliers.

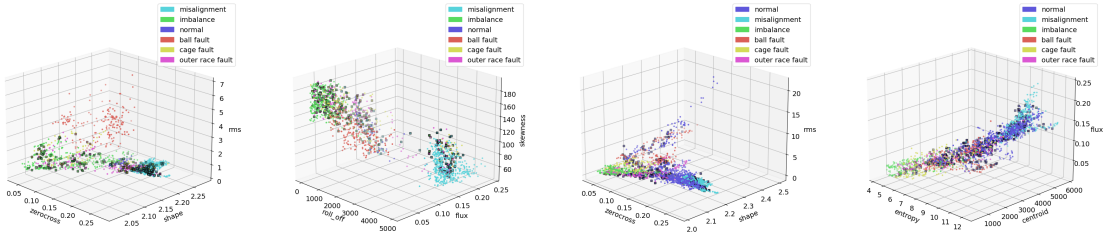
5.1.4 Incremental learning

Online learning imitates hardened conditions for machinery diagnostics that appear in deployment. Delayed provision or omission of actual labels undoubtedly degrades



(a) Accuracy (b) Percentile in distribution of accuracies

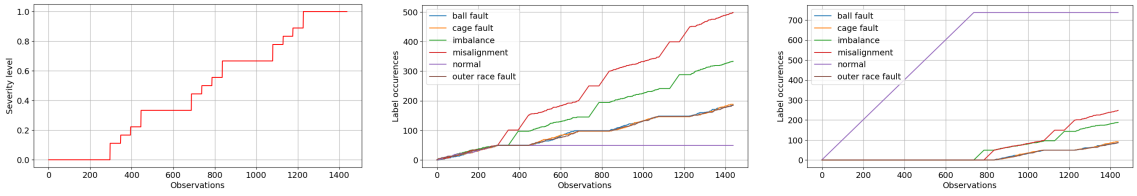
Figure 5.6: Quality of choice for feature selection methods



(a) TD [89.05%] (b) FD [77.18%] (c) TD (s) [91.40%] (d) FD (s) [91.24%]

Figure 5.7: Three features in both domains chosen by rank product with prediction accuracy in brackets. Relabeled dataset is marked with (s)

the reliability of the classification. The question is how quickly the accuracy approaches the optimal one from the nearest neighbors trained in batch and what is the effect on the classifier from routine difficulties associated with the continuous labeling process.



(a) Relative severity levels (b) Original labels (c) High severity labels

Figure 5.8: Ordering of faults in dataset according to relative severity levels

The k-NN models in incremental learning experiments train on the same base training dataset as those in batch learning for bearing A. Online learning metrics are evaluated by progressive valuation on a still unbalanced dataset. In this manner,

we can compare the training accuracies for the last sample of online models to their batch counterparts.

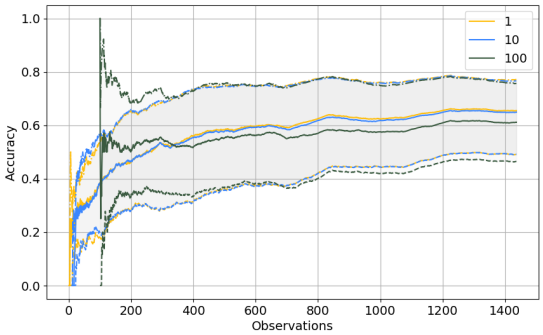
The **stream of events** is ordered by rising severity levels (Fig. 5.8a), which ensures steady increments in label counts throughout the simulation (Fig. 5.8b). The same sorting approach is applied to the dataset where low-severity faults are annotated as baseline. During most of its lifespan, the machine simulator looks to be in a fine operating state. Near the end of the simulation, faults start to develop (Fig. 5.8c). These artificially streamed event sequences are a bit unrealistic because all types of faults never begin to appear simultaneously with equal strengths. It is meant to approximate the gradual overall degradation of the machine.

A significant change in the data stream occurs after 294 out of 1438 observations (or after 737 for high severity faults) when all 49 (or 737) normal conditions are consumed in the training process. Counters of other faults show that predictions are skewed towards more represented classes of imbalance and misalignment. The uneven evolution of categories in a stream impacts the evolution of accuracy in the remaining experiments. The test accuracies of comparable batch models for three best features are 85.47% (TD), 87.52% (FD), 91.71% (TD severity), and 91.94% (FD severity).

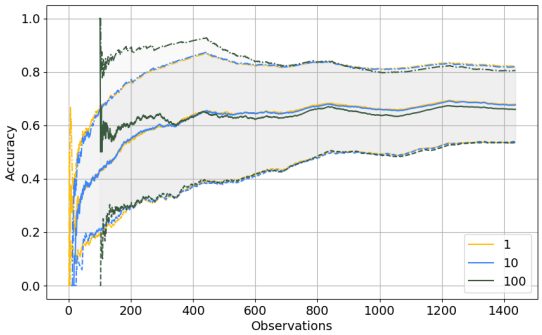
During gradual learning, the correct labels are supplied in bulk at a fixed period. Labeling delay in **tumbling window** decreases towards the window's end. Figure 5.9 plots the evolution of the k-NN model's distribution of accuracy for windows of size 1, 10, and 100 observations. The models consist of three predictors and use five neighbours for prediction. Each window length is described with three curves of identical colour. The median is drawn with a solid line, the maximum with a dash-dotted line, and the minimum with a dashed line.

Initial zero accuracy is caused by a warming-up period in data collection during the span of the first few windows. The true labels are unknown at that point. After just a handful of windows in the beginning, accuracy jumps above 60% for the best triplet of attributes and stabilizes after 400 observations. In an alternative labeling scheme, the accuracy for one class is 100% and only after encountering other fault types, it decreases to a level above 75% for the best set.

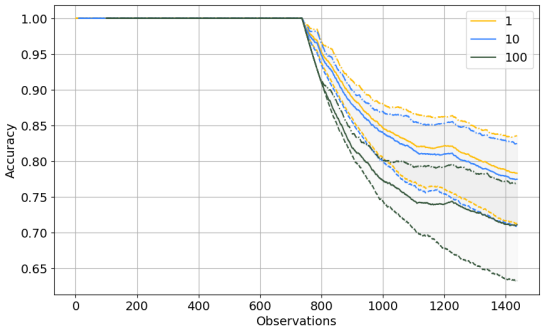
The top accuracies after sequentially seeing samples in the longest tumbling win-



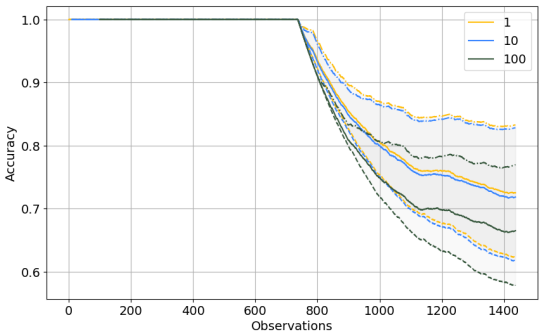
(a) Time-domain features



(b) Frequency-domain features

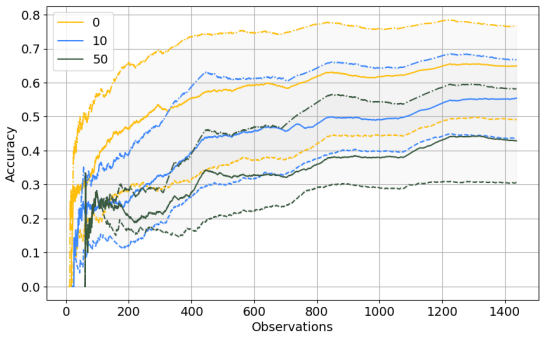


(c) Time-domain features (severity)

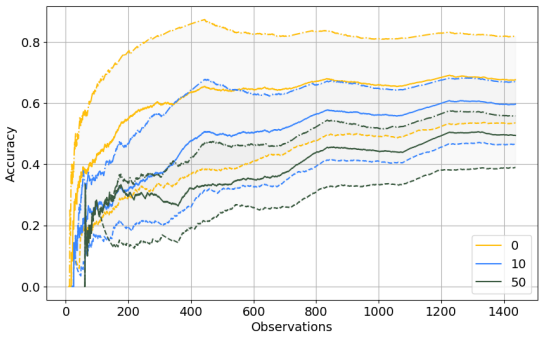


(d) Frequency-domain features (severity)

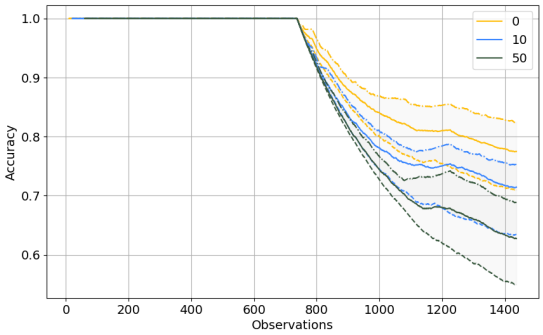
Figure 5.9: Tumbling window of lengths 1, 10, 100 during incremental learning



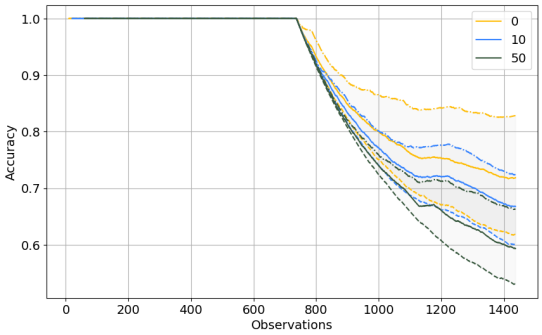
(a) Time-domain features



(b) Frequency-domain features



(c) Time-domain features (severity)



(d) Frequency-domain features (severity)

Figure 5.10: Omission of labels during incremental learning with tumbling window of length 10 and gaps of size 0, 10, and 50 samples

dows of 100 observations are 75.71% (TD), 76.98% (FD), 76.98% (TD severity), and 80.56% (FD severity). The degradations in accuracy when compared to the batch model are by 9.76% (TD), 10.54% (FD), 14.73% (TD severity), and 11.38% (FD severity). The effect of severe performance hit could be attributed to unbalanced data based on the analysis chapter. The online multiclass oversampling method would need to be inserted into the pipeline to prove this hypothesis.

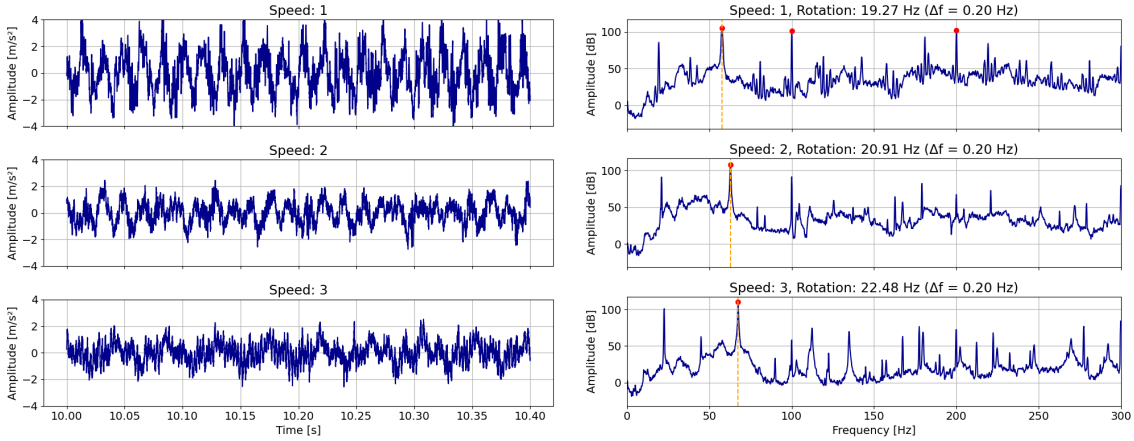
Labeling every 10^{th} sample (10% of the dataset) with a tumbling window of 10 samples reduces maximum accuracy for the three-predictor model compared to data points without gaps in labels. In time-domain features, it is decreased by 9.9% to 66.78%, and by 14.93% in frequency-domain features to 67.00% (Fig. 5.10). More annotations can be scrapped in the case of a relabeled dataset where keeping just 0.02% of the class labels produces a decrease by 13.7% to 68.79% for time domain attributes, and by 16.59% to 66.25% for frequency domain attributes.

5.2 Industrial dataset analysis

The latter part of the solution evaluation focuses on signal properties of vibrations from air compressors, water pumps, and electric induction motors. Our custom dataset is compiled from measurements collected during regular operation of machines. The behaviour of different machines is compared using static frequency estimations and time-frequency spectrograms. The domain expert methodology is carried out to diagnose the current status of water pumps. The presence of fault is confronted with sensor logs from the pump's vendor.

5.2.1 Data logger verification

Data logger capabilities were verified by capturing vibrations on the back of the plastic casing for the electric motor of the standing fan. The recorder is able to run continuously for more than one minute without dropping any samples. The amplitude range after conversion is constrained in the expected range (Fig. 5.11a). The same amplitudes were obtained previously on a proof-of-concept device using analog accelerometer ADXL335 on BeagleBone Black, albeit with lower sensitivity.

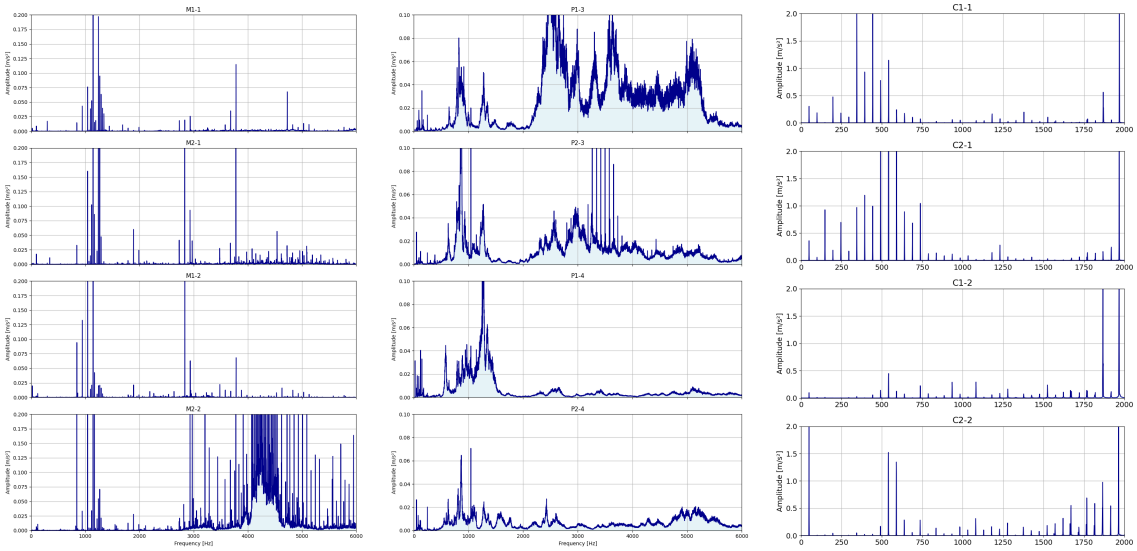


(a) Time waveform (b) Estimation of rotation speed

Figure 5.11: Vibrations from the back of a standing fan in the radial direction

Time waveform captures mostly regular patterns of oscillatory motion (Fig. 5.11a). At slower speeds, the vibrations are higher because the fan is less stable and wobbles around the support. Timestamps from the sensor calibrate the sampling frequency from the theoretical 26667 Hz to the actual 26866 Hz. The estimate of fan rotational speed (Fig. 5.11b) is accurate within the margin of error.

5.2.2 Signal waveforms



(a) Motors (b) Pumps (c) Compressors

Figure 5.12: Wideband frequency waveform of machinery vibartions in Pump dataset

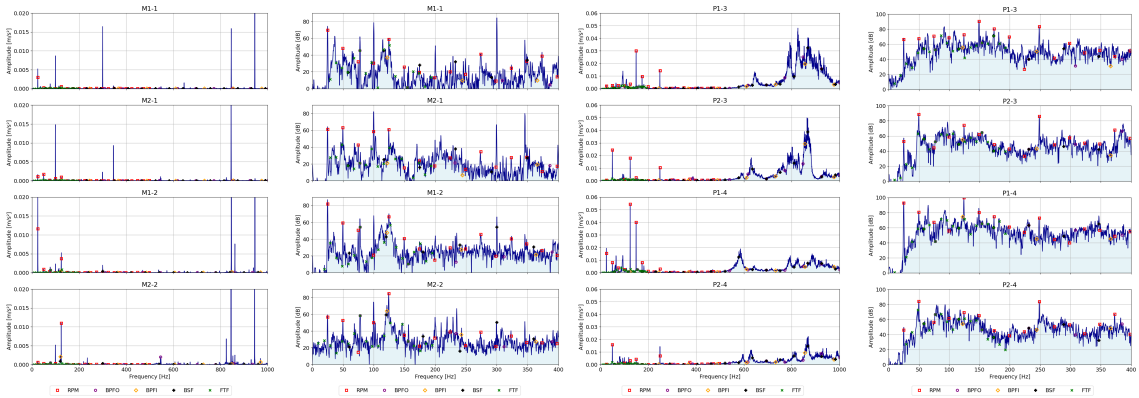
The overview of wideband vibration frequency spectra from various sensor place-

ments shows clear distinctions among the machines. The recordings in Figure 5.12 present consistent shapes of waveforms found during repeated trials. The spectra are Welch's average of 32768 long windows with 50% overlap over minute of recording.

The electric motor signals are polluted by noise from the wind blowing out of a large cooling fan at the back of the motor. The M2 in place two compared to M1 has elevated amplitudes above 4 kHz. The pumps have richer signal content than motors split into several frequency bands, likely due to the flow of water. The outer bearing (4) has a more attenuated amplitude above 1.5 kHz than the inner bearing (3). The P2 exhibits less vibration in comparable bands in general. The compressor casing produces a series of harmonics of rotational frequency, which are stronger near the scroll and suction valve than near the base.

Placement	**M1**	**M2**	**P3, P4**
Bearing	6319-C3	6324-C3	6317-2Z
Rolling elements n	8	8	8
Rotational speed f_s **[rpm]**	1493	1493	1493
Inner diameter d **[mm]**	33.12	41.28	30.00
Outer diameter$D$ **[mm]**	147.5	190.0	132.5
Contact angle β	0	0	0
RPM [Hz]	24.88	24.88	24.88
BPFO [Hz]	77.18	77.91	77.00
BPFI [Hz]	121.88	121.16	122.07
BSF [Hz]	58.20	59.97	57.77
FTF [Hz]	9.65	9.74	9.63

Table 5.5: Bearing characteristic harmonic frequencies of pumps and their motors

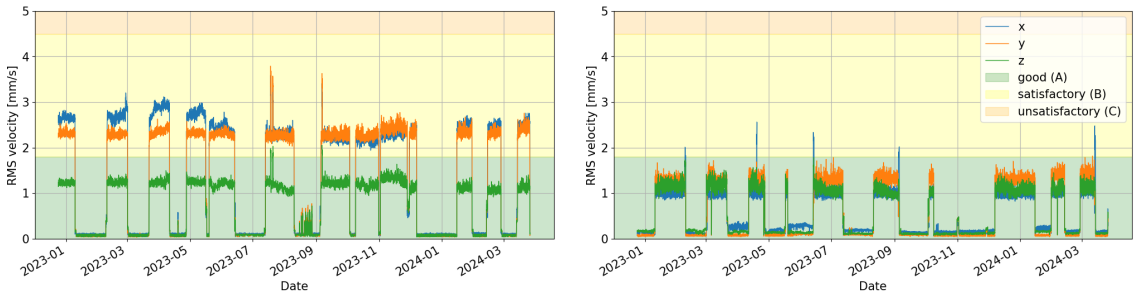


(a) Motors (b) Motors (dB) (c) Pumps (d) Pumps (dB)

Figure 5.13: Characteric bearing frequencies of pumps

Domain experts recommended the procedure for fault identification by calcula-

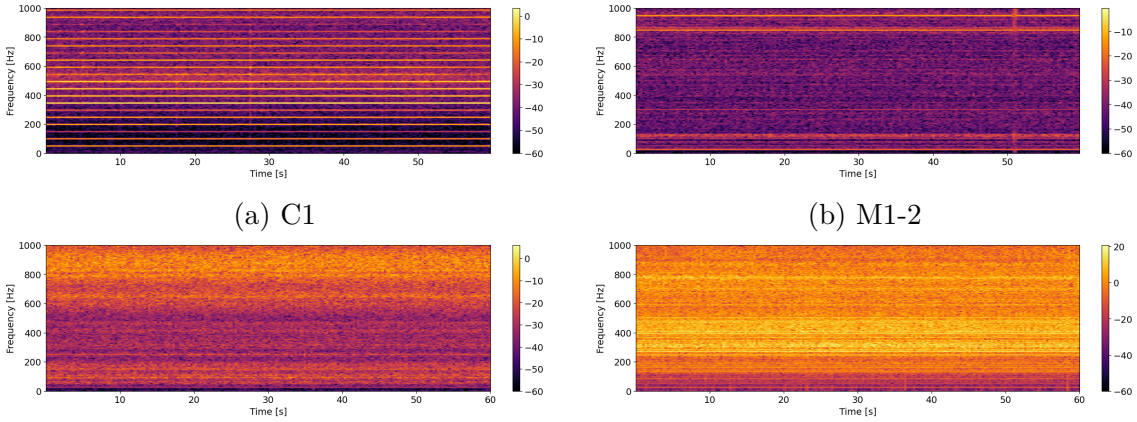
tion of bearing characteristic frequencies (Tab. 5.5), and they approved the following results. Figure 5.13 identifies harmonics of rotational speed and BPFO frequency in every machine and BPFI for M2-2. It can be assumed that those frequencies will be the reason for damage in the future. The absolute acceleration is minuscule in current frequency bands and year-long sensor logs of rms velocity (Fig 5.14).



(a) Pump P1 (b) Pump P2

Figure 5.14: Vibration rms levels for a period over a year from KSB Guard

Therefore, the current results suggest that the bearings are in flawless condition. During the over five years the pumps have been in service, there is not one instance of bearing fault due to rated lifespan and yearly prophylactic maintenance. This underscores the scarcity of recording faulty states in industrial environments. The monitoring is better suited in situations without preventive maintenance or with smaller machinery like the plunger pump BC 21/20S we considered during air conditioning inspections.



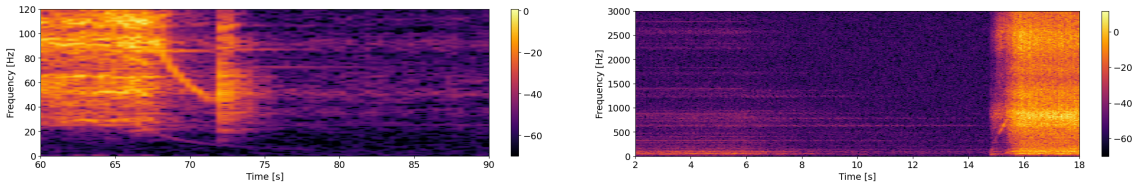
(a) C1 (b) M1-2

(c) P1-3 (d) P3-3

Figure 5.15: Machinery vibration spectrograms ($f_s = 26.8$ kHz, $w = 8$ kS)

Spectrograms point out that under constant machine load, the significant fre-

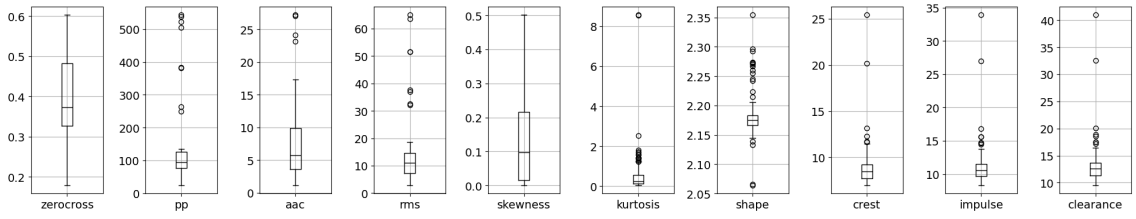
quencies stay more or less unchanged over time (Fig 5.15). The resolution of the uncertainty box is 306.95 ms and 3.26 Hz. Its magnitude is colored for decibel value.



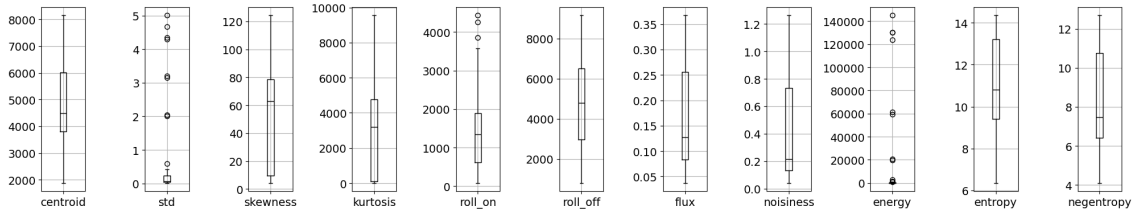
(a) P1 turns off ($w = 4$ kS) (b) P2 turns on ($w = 32$ kS)

Figure 5.16: Vibrations of water pumps during switch over ($f_s = 26.8$ kHz)

The water pump switch on and off creates a fast change in shaft speed. Resonance bands cannot be separated with enough time resolution by an automatic system at a chosen sampling rate. Due to inertia, slowing down takes about 10 seconds, whereas the speed up takes just half a second (Fig. 5.16. The interference from excessive vibrations of the adjacent Sigma pump is picked up by the accelerometer on the KSB pump just before their switchover occurs (Fig 5.16b).



(a) Time-domain features



(b) Frequency-domain features

Figure 5.17: The attribute value ranges of Pump dataset

The range of feature values for compressors and pumps are higher than in MaFaulDa despite most of them describing baseline status (Fig 5.16b). The increase can be seen for example in the interquartile range of rms (7.33 - 14.62 m/s^2), centroid (3806 - 6003 Hz) or entropy (9.40 - 13.22 nats).

6 Conclusion

The focus of this master's thesis is on the optimal choice of trend indicators from vibration signals in an inexpensive industrial condition monitoring solution. The goal is to enable timely fault detection of machinery parts with as little input data as possible. We answer **five research questions** for those purposes.

The first question is about finding numerical features that can represent machinery behaviour for accurate diagnosis. The technical standards in vibrodiagnostics and the descriptive statistics for time-varying phenomena are valuable in that regard. Vibrations have to be recorded at a high sampling rate. The researched formulas narrow down this massive data stream to groups of a few features. We define ten features that summarize the waveform in the time domain (TD) and eleven features that describe its spectral density estimation (FD). The features are either just taken from the axis of the jitter, or aggregated from all three spatial directions.

The second question answers the effort to achieve even more pronounced savings in transmission goodput. Groups of three predictors are chosen from each base feature set that together have the most similarity to the target variable. Feature selection scores of the correlation coefficient, F statistic, mutual information, and their rank product are computed in supervised learning scenarios. Lossy compression ratios attained are 2381:1 for all features and 25000:1 for six features from the MaFaulDa dataset. We managed to discard more than 99.995% of the irrelevant data.

The third question asks about the accuracy of the machine learning model for machinery fault diagnostics. The major constraint put on the model is to provide inference from reduced sets of predictors. The k-nearest neighbours classifier is utilized because it works on a simple to-explain assumption that proximate data points mean the same type of fault. Additionally, the k-NN retains individual historical observations and can be applied in an incremental learning approach.

We implement the **data processing pipeline** that transforms the MaFaulDa

dataset for multiclass classification with the k-NN algorithm and a tiny number of features. The fault labels with severity are composed of the file's path from the dataset structure. The features are extracted from vibration time series after digital filtering. Then, the features are arranged into multiple configurations: complete feature sets, enumerated feature subset combinations, and feature subsets chosen by filtering feature selection methods. The accuracy results are validated by five-fold cross-validation.

We found that for this particular dataset, the chosen time-domain features have reached higher accuracies than the chosen frequency-domain set because of fewer intercorrelations. An increase in k-neighbours lessens the model accuracy because class boundaries overlap and are relatively noisy. Predictors aggregated from tri-axis recording performed better in classification than those from just one axis.

The feature selection methods can pick a group of predictors with accuracy above the upper quartile of the statistical distribution of accuracy values for the same-sized sets. The rank product method found the best-performing attributes in the majority of 44% of scenarios, closely followed by mutual information. The rank product for the predictor's triplets reaches the 97.50% percentile from the accuracy distribution.

The number of features has a direct proportionality effect on accuracy. The absolute best predictor triplets for MaFaulDa are in TD (accuracy of 85.47%): zero-crossing rate, peak-to-peak, skewness, and in FD (accuracy of 87.52%): spectral centroid, roll-off frequency, and entropy.

The fourth question investigates machinery behavior in an industrial environment using signal processing techniques. The vibrations were collected a month apart by an accelerometer data logger according to the specification. The designated machines were air conditioning compressors and water pumps with motors in municipal pumping station. We implemented a data logger's firmware that reads samples after a button press from the accelerometer at 26.8 kHz and stores them onto a memory card.

Analysis of the collected dataset confirms the stationarity of the signal spectrum under constant load and assures that five-second bursts are indeed satisfactory for feature extraction. The dataset provides a practical exercise in labeling potential future faults from bearing defect frequency amplitudes. It also demonstrates the

immense difficulty of obtaining observations of faulty machinery in the field without purposefully damaging it.

The fifth question is concerned with incremental machine learning because the considerable obstacle in the deployment of an autonomous fault detection system is the availability of annotations that can be assigned belatedly or even never. Incremental k-NN algorithm on an unbalanced dataset for three predictors achieves at best 77% accuracy with labels in tumbling windows of hundred samples. The accuracy of 67% is reached for the online model when just 10% of observations stay associated with the label. The comparable model trained in batch reaches an accuracy of 88%.

We conclude that an inexpensive industrial condition monitoring solution with low data rates is indeed possible. It uses the available accelerometer sensors and feature discovery techniques. According to the analysis, the system could diagnose the machine's health with high accuracy in a local model on the edge device.

6.1 Future work

Future work would focus on the deployment of our machine learning model on a large scale to sensor units and setting up the necessary infrastructure. The components for such an endeavour are described in the analysis for incremental learning. The way to annotate observations could be introduced, either automatically depending on the overall vibration level or manually during maintenance or after an unexpected failure.

The natural research extension of the thesis is to confirm that the results hold in other publicly available datasets of machinery faults, e.g. in CWRU. The other alteration is to use more and different starting base feature sets. Harmonic frequencies or wavelet coefficients can generate predictors that could be reduced further to the smallest possible subsets.

Ensemble strategies inspired by classical election systems may be utilized in the task of feature selection. Different types of classifiers for online learning could be compared to k-nearest neighbours, e.g. Hoeffding tree. A combination of local models from multiple sensors is desirable, e.g. in a federative learning approach.

7 Resumé

7.1 Úvod

Vzostup priemyslu 4.0 so sebou prináša väčšiu mieru automatizácie s cieľom dosiahnuť optimálne využitie dostupných zdrojov. Na základe nepretržitého sledovania opotrebenia zariadení v reálnom čase sa majú zabezpečiť nápravné opatrenia na opravu alebo výmenu súčiastok včas, v reakcii na trendové ukazovatele.

Cieľom je zachovať požadovanú bezpečnosť a efektivitu výroby a zároveň predĺžiť životnosť rotujúcich komponentov. Proaktívna diagnostika porúch je nevyhnutná na začatie opráv bez nadbytočných nákladov. Vibrácie predstavujú nerušivý spôsob, ako zistiť a zaznamenať prípadne fatálne zlyhania hneď v zárodku. Hlavným problémom pri monitorovaní veľa strojov s vibráciami, je to, že vzniká množstvo záznamov, ktoré nie sú priamo užitočné pre operátora výrobnej linky. Väčšina signálov sa zobrazí maximálne raz, preto je zbytočné ich ukladať alebo prenášať vcelku.

Zároveň na dosiahnutie maximálnej presnosti detekcie musí byť model strojového učenia trénovaný pre cieľové prostredie. Poruchy sú navyše pomerne zriedkavé udalosti, ktoré sa zvyčajne vyskytujú s odstupom niekoľkých mesiacov. Za týchto okolností je ťažké rýchlo získať dostatočne veľkú vzorku poruchových udalostí.

7.2 Sledovanie prevádzkového stavu

Existujú tri rôzne prístupy k údržbe strojov: reaktívny, preventívny a prediktívny. Pri reaktívnej údržbe beží stroj až do úplného zlyhania a je prijateľná vtedy, keď je možná úplná a rýchla výmena pokazeného stroja za záložný. Preventívna údržba prebieha v pravidelných intervaloch odvodených od vopred určeného rozvrhu v alebo strednej doby medzi poruchami. Prediktívna údržba zlepšuje predvídateľnosť oproti reaktívnej údržbe a eliminuje plytvanie voči príliš obozretnej prevencii. Odstávka stroja je naplánovaná po zistení kritických hodnôt a po odhalení problematických

komponentov.

Mechanické problémy počas prevádzky strojov spôsobujú v mnohých prípadoch vibrácie. Vibroakustická diagnostika sa preto považuje za jednu z najdôležitejších metód pri včasnej identifikácii porúch komponentov. Najbežnejšie sa vyskytujúcimi poruchami sú nevyváženosť, nesúosovosť, vôľa, excentricita, deformácia, trhlina a nadmerné trenie.

Symptómy porúch rotačných strojov sa prejavujú rôznymi frekvenčnými pásmami, ale väčšina je závislá od rotačnej rýchlosti súčiastky. Nevyváženosť, nesúosovosť a vôľa sa bežne objavujú v frekvenciách do 300 Hz. Poruchy ložísk a prevodovky v neskorých štádiách vývoja sa prejavujú v rozsahu medzi 300 Hz a 1 kHz. Vyššie frekvencie do 10 kHz pomáhajú odhaliť poruchy ložísk v skorších štádiách rozvoja.

Postupy monitorovania stavu založené na vibráciách musia byť v súlade s normatívnymi smernicami ISO 20816 a ISO 13373. Normy sa týkajú umiestnenia meracích zariadení, zberu údajov, konvencií nastavenia úrovní závažnosti porúch.

7.3 Extrakcia a výber atribútov

Prediktívna údržba má ideálne predpoklady na využitie extrakcie atribútov, pretože signál je zvyčajne stacionárny a trendové premenné v časovej a frekvenčnej oblasti vychádzajú z expertných znalostí v oblasti mechaniky. Výhody dodatočného úsilia v porovnaní so spracovaním pôvodných vzoriek spočívajú v dosahovaní lepšej presnosti klasifikácie, znížení výpočtovej záťaže a znížení potreby úložnej kapacity. Výber atribútov nie je samostatným krokom v procese strojového učenia, ale mal by sa vykonávať iteratívne na zlepšenie výsledného modelu.

Najrozšírenejšími používanými atribútmi sú štatistické miery centrálneho momentu: priemer, rozptyl, štandardná odchýlka, šikmosť a špicatosť. Charakteristiky amplitúdy zahŕňajú kvadratický priemer (rms), vzdialenosť špička-špička, maximum, absolútnu zmenu amplitúdy (aac), a početnosť prechodov nulou. Ostatné významné atribúty časovej oblasti sú odvodené ako pomery a sú nim: faktor výkyvu, faktor rozpätia, faktor impulzu a faktor tvaru.

Vo frekvenčnej oblasti sú získané obvyklé štatistické vlastnosti distribúcie, ktorými sú spektrálne ťažisko, šikmosť a špicatosť. Okrem toho sa extrahujú frekvencie roll-

on a roll-off, fundamentálna frekvencia, entropia, negentropia, vzájomná korelácia spektier, pomer signálu k šum, a energia vo frekvenčných pásmach.

Atribúty neprispievajú k prediktívnej sile modelu s rovnakým podielom. Výber ich optimálnej podmnožiny je NP-ťažký kombinatorický problém. Kroky všeobecného postupu pri výbere atribútov metódou filtrovania sú generovanie podmnožín, vyhodnotenie podmnožín, ukončovacie kritérium hľadania, a validácia.

Hodnotenie relevancie atribútov je založené na skórovaní podobnosti s predikovanou premennou. Často používané spôsoby zoraďovania dôležitosti atribútov sú prah rozptylu, koeficienty korelácie, ANOVA F štatistika, a vzájomná informácia. Viaceré podmnožiny prediktorov produkovaných každou z výberových metrík môžu slúžiť na trénovanie viacerých variantov klasifikačného modelu. Množiny atribútov je možné kombinovať do súboru volebným systémom ako sú väčšinové hlasovanie alebo súčin poradí.

7.4 Diagnostické prístupy

Identifikácia porúch rotujúcich strojoch je binárny alebo viac-triedny klasifikačný problém, ktorý pracuje na princípe učenia čiastočne s učiteľom, pretože označenia pre degradované stavy stroja sú v praxi zriedkavé. Ciele automatizácie monitorovania možno rozdeliť na detekciu anomálií a rozpoznanie typu poruchy.

Detekcia anomálií, novostí alebo odľahlých hodnôt určuje, či sa prevádzkový stav stroja výrazne odchyľuje od normálu. Po upozornení môže zasiahnuť odborník a stroj diagnostikovať. Odľahlé hodnoty sú odvodzované na základe neparametrických štatistických modelov, zhlukovania podľa najbližších susedov a prístupov založených na izolácii anomálnych vzoriek. DenStream je algoritmus zhlukovania založený na hustote prispôsobený z DBSCAN na zhlukovanie prúdových dát do ľubovoľne tvarovaných skupín.

Presná viac-triedna klasifikácia príčin porúch stroja podľa vopred známych charakteristík je oveľa náročnejšia úloha ako objavenie anomálií. Algoritmus k-najbližších susedov (k-NN) priradí pozorovanie triede, do ktorej patrí väčšina k bodov v blízkom okolí podľa použitej miery vzdialenosti. Nachádza uplatnenie aj v učení čiastočne s učiteľom, pretože dokáže odvodiť označenia len zo znalosti niekoľkých anotácií.

Ďalším prístupom je online alebo postupné učenie, ktoré aktualizuje parametre modelu s každou novou prichádzajúcou udalosťou. Tento prístup je užitočný pri spracovaní veľkých dát, kedy celý súbor údajov nie je k dispozícii vopred alebo ho nemožno spracovať naraz z dôvodu pamäťových obmedzení.

7.5 Výskumné otázky

Cieľom tejto práce je poskytnúť odpovede na päť výskumných otázok:

1. Aké atribúty dokážeme extrahovať z vibračných signálov?
2. Akú úsporu dát dosiahneme výberom atribútov?
3. Aké budú presnosti diagnostiky porúch s rôznymi sadami atribútov?
4. Ako sa správajú vibračné signály zozbierané na priemyselných strojoch?
5. Ako môžeme priebežne označovať poruchové stavy?

7.6 Návrh spracovania pre MaFaulDa

Vo všeobecnosti pozostáva spracovanie signálov vibrácií zo získavania signálu, extrakcie atribútov, redukcie rozmerov a rozpoznávania vzorov alebo detekcie porúch. Realizácia uvedeného postupu je dekomponovaná podľa štruktúry dát a parametrov jednotlivých fáz.

Predspracovanie najskôr priradí metadáta časovým radom v CSV súboroch z MaFaulDa podľa ich cesty v adresárovej štruktúre. Na základe pôvodného označenia porúch a polohy ložiska je ponechaných šesť tried: bezporuchový stav, nevyváženosť, nesúosovosť, porucha klietky, guľôčok a vonkajšej dráhy ložiska. Každé ložisko sa posudzuje zvlášť. Každú nahrávku popisuje trojica typu poruchy, závažnosti a rýchlosti otáčok.

Nasleduje rozdelenie časových radov na menšie časti, ak je ich trvanie dlhšie ako dvanásť okien a majú rozlíšenie okolo 1 Hz, čo však neplatí pre MaFaulDa. Jednosmerná zložka a frekvencie nad 10 kHz sú odfiltrované z pôvodných signálov vibrácií. Pôvodne priradené triedy porúch sú pri niektorých experimentoch vymenené za bezporuchový stav, keď je závažnosť poruchy nízka.

Z predspracovaného signálu v každej osi akcelerometra sú vypočítané dve základné sady atribútov v časovej (TD) a frekvenčnej doméne (FD). Na filtrovanie pozorovaní sa používajú štyri podmienky: umiestnenie ložiska (A alebo B), doména základnej sady atribútov (TD alebo FD), osi akcelerometra použité na výpočet atribútu (jedna alebo tri), ponechanie označení iba pre poruchy vyššej závažnosti (áno alebo nie). Uvedené podmienky vytvárajú dohromady 24 scenárov.

Úspešnosť klasifikácie porúch strojov sa vyhodnocuje po normalizácii atribútov, vyvážení mohutnosti tried a päťnásobnej krížovej validácii na klasifikátore k-najbližších susedov s euklidovskou metrikou vzdialenosti. Presnosti sa porovnávajú pre rôzne hodnoty hyperparametrov k-susedov a veľkostí podmnožiny atribútov podľa uvedených scenárov.

Návrh experimentov zahŕňa popísané filtračné podmienky, ktoré upravia MaFaulDa do 24 podôb. Navyše v postupnom učení sa porovnávajú dĺžka oneskorovania anotácii a ich vynechávanie. Štyri hlavné experimenty pre porovnanie výberu podmnožiny atribútov zahŕňajú klasifikáciu porúch na základných sád atribútov, na všetkých kombinácií podmnožín prediktorov s určitou veľkosťou, na najlepších atribútoch vybraných cez metriky podobnosti s cieľovou premennou a pomocou modeloch postupného učenia s sťaženým prístupom k pravdivým označeniam tried.

7.7 Zber vibrácií v priemysle

Metodiku uplatnenú na súbore údajov MaFaulDa z laboratórneho prostredia aplikujeme na vibrácie z priemyslu. Pri monitorovaní je zužitkovaný postup z technických noriem. Ten zahŕňa výber strojov určených na monitorovanie, identifikáciu pozícií na meranie, predbežné merania a vývoj senzorovej jednotky.

Na zber údajov boli vyčlenené dva špirálové kompresory ako súčasť klimatizačných jednotiek pre dátové centrum a tri čerpadlá s troma elektromotormi v prečerpávacej stanici na pitnú vodu. Merania sú uskutočnené s mesačným rozostupom vlastným data loggerom na báze vývojovej dosky ESP32-PoE-ISO so slotom na SD kartu. Ako senzor vibrácii je použitý MEMS akcelerometer ST IIS3DWB. Vyznačuje sa vysokou šírkou pásma až 6.3 kHz, nízkym šumom, a vysokou výstupným dátovým tokom 26.7 kHz cez SPI zbernicu.

Meranie vibrácií na jednom ložisku stroja zahŕňa tri pokusy s nahrávkou o dĺžke 60 sekúnd. Po každom zázname je akcelerometer znova pripevnený na meraciu pozíciu. Snímač je pripevnený k stroju na rovnom povrchu obojstrannou kobercovou páskou.

7.8 Vyhodnotenie presnosti diagnostiky

Overovanie navrhovaných riešení diagnostiky porúch strojov sa zameriava na dve činnosti, ktorými sú meranie vibrácií a identifikácia porúch.

Väčšinu zámen pri odhaľovaní porúch spraví model pri poruche vonkajšieho krúžku ložiska, ktorú označuje za poruchu klietky ložiska alebo nerovnováhu hriadeľa a menej často za nesúosovosť. Nevyváženosť hriadeľa zabezpečuje v laboratóriu simuláciu porúch ložísk, čiže tam nastáva prirodzene k značnej zámene týchto porúch.

Zvyšujúci sa počet susedov použitých na klasifikáciu s k-NN ukazuje podstatné zníženie presnosti na testovacích dátach. Najvýraznejší pokles o približne 10% nastáva po deväť susedov. Atribúty vytvorené z trojosového vektora dosahujú lepšiu presnosť ako tie výhradne z osi pohybu pre rovnakú zdrojovú doménu a ložisko. Model pre vnútorné ložisko A je presnejší ako vonkajšie ložisko B. Sada TD je lepšia v identifikácií porúch ako sada FD pre ekvivalentnú hodnotu k. Dátová sada s anotáciami pre poruchy vysokej závažnosti má prudšie zníženie presnosti pre rovnaký počet susedov.

Základné sady prediktorov sú ešte značne zmenšené na reprezentáciu, ktorá by mohla byť prezentovaná v 3D grafe alebo v rovinných rezoch. Každá možná kombinácia párov, trojíc a štvoríc natrénuje samostatný k-NN model, na ktorom sa hodnotí presnosť klasifikácie. Zníženie presnosti so zvzšujúcim sa počtom susedov je zrejmé a podobné trendu v základných sadách atribútov. Zníženie maximálnej presnosti je výraznejšie medzi tromi a piatimi susedmi a takmer o rovnaké množstvo sa degraduje model medzi piatimi a jedenástimi susedmi.

Ak je počet atribútov je najviac tri a súčasne počet susedov je päť alebo menej, základné množiny atribútov dosahujú lepšiu presnosť ako ich podmnožiny. Počet prediktorov má priamy úmerný vplyv na presnosť optimálneho modelu. Presun z

dvoch na tri atribúty má väčšiu váhu ako pridanie štvrtého prediktora.

Prediktory vybrané pomocou výberových metód sú porovnané s presnosťou klasifikácie kombinácií skupín atribútov rovnakej veľkosti a s presnosťou ich zdrojovej nadmnožiny. Na získanie konzistentnej presnosti je potrebné kombinovať hodnotenia z niekoľkých metrík výberu atribútov. PCA o troch komponentoch transformovaných zo základných sád atribútov je presnosťou porovnateľná s metódami výberu s pôvodnými atribútmi. Trojica prediktorov s najlepšími výsledkami zo sady TD sú: početnosť prechodov nulou, vzdialenosť špička-špička, šikmosť a zo sady FD: spektrálne ťažisko, roll-off frekvencia a entropia.

Metrika vzájomnej informácie dosahuje lepšiu strednú presnosť (80,87%) a percentil v distribúcii presností (91,81%). Nasledovaná je súčinom poradí s presnosťou 79,82% a percentilom 88,97% Súčin poradí je považovaný za najlepšiu stratégiu v 43,52% prípadov. Vzájomná informácia je na druhom mieste s 40,28% prípadov. Stredná presnosť vo sitáciach, kde je zvolená metóda výberu najlepšia, je tiež lepšia pre súčin poradí s 92,38% v porovnaní s 91,79% pri vzájomnej informácii. Výber atribútov zvyčajne vyberá premenné tak, že ich presnosť objaví v hornom kvartile distribúcie.

Postupné učenie napodobňuje sprísnené podmienky pre diagnostiku strojov, ktoré sa objavujú pri nasadení v praxi. Oneskorené poskytnutie alebo vynechanie skutočných označení nepochybne znižuje spoľahlivosť klasifikácie. Modely k-NN v experimentoch s postupným učením sa učia na rovnakom základnom súbore trénovacích údajov ako pri dávkovom učení pre ložisko A. Metriky postupného učenia sa vyhodnocujú progresívnym vyhodnotením na nevyváženom súbore údajov. Presnosti klasifikácie medzi postupným a dávkovým učením sú porovnané na finálnej presnosti online modelu.

Udalosti sú usporiadané podľa stúpajúcej úrovne relatívnej závažnosti, čím simulujeme postupnú celkovú degradáciu stroja. Presnosť testov porovnateľných dávkových modelov z troch najlepších vlastností je 85,47% (TD), 87,52% (FD), 91,71% (TD, závažnosť) a 91,94% (FD, závažnosť). Najvyššie presnosti po postupnom zhliadnutí vzoriek v najdlhších posuvných oknách s dĺžkou 100 pozorovaní sú znížené v porovnaní s dávkovým modelom o 9,76% (TD), 10,54% (FD), 14,73% (TD, závažnosť) a 11,38% (FD, závažnosť). Ponechanie 10% anotácií ich rovnomerným vynechá-

vaním a trénovanie modelu s oknom 10 vzoriek zníži maximálnu presnosť modelu s tromi prediktormi v časovej doméne o 9,9% na 66,78% a o 14,93% na 67,00% vo frekvenčnej doméne.

7.9 Rozbor dátovej sady z priemyslu

Správanie sa rôznych strojov je porovnané vo frekvenčnej doméne a cez časovo-frekvenčné spektrogramy. Aktuálny stav čerpadiel nameraný vlastným data loggerom je obohatený o záznamy zo senzora vibrácií od výrobcu. Potenciálne poruchy sú diagnostikované podľa postupu od doménových expertov.

Motor M2 na pozícii dva má zvýšené amplitúdy nad 4 kHz v porovnaní s motorom M1. Čerpadlá majú bohatší frekvenčný signál ako motory, pravdepodobne v dôsledku toku vody. Vonkajšie ložisko čerpadla (4) vykazuje nižšiu amplitúdu vibrácií nad 1,5 kHz ako vnútorné ložisko (3). Čerpadlo P2 vo všeobecnosti vykazuje menšie vibrácie v porovnateľných pásmach. Skriňa kompresora vytvára sériu harmonických frekvencií, ktoré sú silnejšie v blízkosti sacieho ventilu ako v blízkosti základne.

Doménoví experti odporučili postup identifikácie porúch výpočtom charakteristických frekvencií ložísk. Vo frekvenciách sú viditeľné harmonické frekvencie rýchlosti rotácie frekvencie BPFO pri každom stroji a BPFI pre M2-2. Dá sa predpokladať, že práve tieto frekvencie budú v budúcnosti dôvodom poškodenia týchto strojov.

Aktuálne výsledky naznačujú, že ložiská sú v bezchybnom stave. Počas viac ako piatich rokov prevádzky čerpadla sa nevyskytol ani jeden prípad poruchy ložiska v dôsledku ich dlhej životnosti a každoročnej profylaktickej údržby. To podčiarkuje náročnosť záznamu poruchových stavov v priemyselnom prostredí.

7.10 Záver

V diplomovej práci sme sa zamerali na výber trendových ukazovateľov pre monitorovanie prevádzkového stavu rotačných strojov a odhaľovanie porúch z vibračných signálov. Cieľom je umožniť včasnú detekciu porúch strojných častí s čo najmenším množstvom vstupných údajov, pričom odpovedáme na päť výskumných otázok.

Prvá otázka sa týka hľadania numerických atribútov, ktoré môžu reprezentovať správanie strojov pre ich presnú diagnostiku. V tomto ohľade sú dôležitým zdrojom technické normy v oblasti vibrodiagnostiky a popisné štatistiky. Definujeme desať atribútov v časovej doméne a jedenásť vo frekvenčnej oblasti.

Druhá otázka odpovedá na snahu dosiahnuť ešte výraznejšie zníženie nárokov na objem prenesených dát. Výber najdôležitejších atribútov sa hodnotí cez korelačný koeficient, F štatistiku, vzájomnú informáciu a súčin poradí jednotlivých metód. Dosiahnuté stratové kompresné pomery pre MaFaulDa sú 2381:1 pre všetky atribúty a 25000:1 pre šesť atribútov. Podarilo sa nám dosiahnuť úsporu dát o viac ako 99,995%.

Tretia otázka sa pýta na presnosť modelu strojového učenia na diagnostiku porúch stroja. Implementujeme postup spracovania údajov z dátovej sady MaFaulDa, ktorá umožní klasifikáciu algoritmom k-najbližších susedov pri malom počte atribútov. Zistili sme, že pre nami použitá dátová sada dosiahla pre zvolené atribúty časovej domény vyššiu presnosť ako zvolená množina z frekvenčnej domény. Zvýšenie počtu susedov pre k-NN vedie k menšej presnosti modelu. Prediktory agregované z trojosového záznamu dosiahli lepšie výsledky v klasifikácii ako prediktory len z jednej osi. Metódy výberu atribútov dokážu vybrať skupinu prediktorov s presnosťou nad horným kvartilom štatistického rozdelenia hodnôt presnosti. Metóda súčinu poradí našla najlepšie výkonné atribúty vo väčšine scenárov a pre trojicu prediktorov dosahuje percentil 97.5% z rozdelenia presnosti.

Štvrtá otázka skúma správanie strojov v priemyselnom prostredí pomocou techník spracovania signálov. Vibrácie boli zbierané s odstupom jedného mesiaca pomocou vlastného data loggera. Ich analýza potvrdzuje stacionárnosť spektra signálu pri konštantnej záťaži a overuje, že päťsekundové zhluky sú skutočne uspokojivé na extrakciu atribútov.

Piata otázka sa týka postupného učenia, kde k-NN dosahuje presnosť prinajlepšom 77% s anotáciami, ktoré prichádzajú v posuvných oknách s dĺžkou sto pozorovaní. Presnosť 67% dosahuje online k-NN model, keď len 10% pozorovaní zostane označených. Porovnateľný model trénovaný v dávkach dosahuje presnosť 88%.

Výber malého počtu trendových ukazovateľov sa ukázal ako dostatočne spoľahlivý

na určovanie porúch rotačných strojov. Natrénovaný model umožňuje nasadenie na senzorovú jednotku IoT zaradenia pri výraznej úspore posielaných dát.

Bibliography

1. MOHANTY, Amiya Ranjan. *Machinery Condition Monitoring: Principles and Practices*. CRC Press, 2015. ISBN 978-1-4665-9305-3.

2. EL-THALJI, Idriss. Predictive Maintenance (PdM) Analysis Matrix: A Tool to Determine Technical Specifications for PdM Ready-Equipment. *IOP Conference Series: Materials Science and Engineering*. 2019, vol. 700, p. 012033. Available from DOI: 10.1088/1757-899X/700/1/012033.

3. SCHEFFER, C.; GIRDHAR, P. *Practical Machinery Vibration Analysis and Predictive Maintenance*. IDC Technologies, Elsevier, 2004. ISBN 0-7506-6275-1.

4. ÇINAR, Zeki Murat; ABDUSSALAM NUHU, Abubakar; ZEESHAN, Qasim; KORHAN, Orhan; ASMAEL, Mohammed; SAFAEI, Babak. Machine Learning in Predictive Maintenance towards Sustainable Smart Manufacturing in Industry 4.0. *Sustainability*. 2020, vol. 12, no. 19, p. 8211. ISSN 2071-1050. Available from DOI: 10.3390/su12198211.

5. ŽIARAN, Stanislav. *Technická diagnostika*. 1. vyd. Bratislava: Vydavateľstvo STU, 2013. ISBN 978-80-227-4051-7.

6. DAVIES, A. *Handbook of Condition Monitoring: Techniques and Methodology*. Dordrecht: Springer Netherlands, 2012. ISBN 978-94-011-4924-2.

7. JENNIONS, Ian K. *Integrated Vehicle Health Management: Perspectives on an Emerging Field*. SAE International, 2011. ISBN 978-0-7680-6432-2. Available from Google Books: PIi7pwAACAAJ.

8. BOUSDEKIS, Alexandros; MENTZAS, Gregoris. Enterprise Integration and Interoperability for Big Data-Driven Processes in the Frame of Industry 4.0. *Frontiers in Big Data*. 2021, vol. 4. Available from DOI: 10.3389/fdata.2021.644651.

9. OKOH, C.; ROY, R.; MEHNEN, J.; REDDING, L. Overview of Remaining Useful Life Prediction Techniques in Through-life Engineering Services. *Procedia CIRP*. 2014, vol. 16, pp. 158–163. ISSN 2212-8271. Available from DOI: 10.1016/j.procir.2014.02.006.

10. TORRES, Pedro; RAMALHO, Armando; CORREIA, Luis. Automatic Anomaly Detection in Vibration Analysis Based on Machine Learning Algorithms. In: *Innovations in Mechatronics Engineering II*. Cham: Springer International Publishing, 2022, pp. 13–23. Lecture Notes in Mechanical Engineering. ISBN 978-3-031-09385-2. Available from DOI: 10.1007/978-3-031-09385-2_2.

11. *ISO 13373-1:2002 - Condition Monitoring and Diagnostics of Machines - Vibration Condition Monitoring - Part 1: General Procedures*. International Organization for Standardization, 2002.

12. BRITO, Lucas Costa; SUSTO, Gian Antonio; BRITO, Jorge Nei; DUARTE, Marcus Antonio Viana. Fault Detection of Bearing: An Unsupervised Machine Learning Approach Exploiting Feature Extraction and Dimensionality Reduction. *Informatics*. 2021, vol. 8, no. 4, p. 85. ISSN 2227-9709. Available from DOI: 10.3390/informatics8040085.

13. LU, Goodenough. *What Is A Ball Bearing?* 2023.

14. *ISO 20816-1:2016 - Mechanical Vibration - Measurement and Evaluation of Machine Vibration - Part 1: General Guidelines*. International Organization for Standardization, 2016.

15. TITTELBACH-HELMRICH, Klaus. Digital DC Blocker Filters. *Frequenz*. 2021, vol. 75, no. 9-10, pp. 331–339. ISSN 2191-6349. Available from DOI: 10.1515/freq-2020-0177.

16. LYONS, Richard G. *Understanding Digital Signal Processing*. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2011. ISBN 978-0-13-702741-5.

17. NANDI, Asoke Kumar; AHMED, Hosameldin. *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines*. Hoboken, NJ, USA: Wiley-IEEE Press, 2019. ISBN 978-1-119-54462-3.

18. JOHNSON, Max Kuhn and Kjell. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. 2019.

19. MOSTAFAVI, Alireza; SADIGHI, Ali. A Novel Online Machine Learning Approach for Real-Time Condition Monitoring of Rotating Machines. In: *2021 9th RSI International Conference on Robotics and Mechatronics (ICRoM)*. 2021, pp. 267–273. ISSN 2572-6889. Available from DOI: 10.1109/ICRoM54204.2021.9663495.

20. MOCTAR, Sidi Mohamed Sid'El; RIDA, Imad; BOUDAOUD, Sofiane. Time-domain features for sEMG signal classification: A brief survey. 2023.

21. *ISO 13373-2:2016 - Condition Monitoring and Diagnostics of Machines - Vibration Condition Monitoring - Part 2: Processing, analysis and presentation of vibration data*. International Organization for Standardization, 2016.

22. PEETERS, Geoffroy. A Large Set of Audio Features for Sound Description. 2004.

23. AVOCI, Moise. Spectral Negentropy and Kurtogram Performance Comparison for Bearing Fault Diagnosis. In: Dubrovnik, Croatia: International Measurement Confederation (IMEKO), 2020.

24. ADIKARAM, K.K. Lasantha Britto; HUSSEIN, Mohamed; EFFENBERGER, Mathias; BECKER, T. Non-Parametric Local Maxima and Minima Finder with Filtering Techniques for Bioprocess. *Journal of Signal and Information Processing*. 2016, vol. 07, pp. 192–213. Available from DOI: 10.4236/jsip.2016.74018.

25. GERBER, Timothée; MARTIN, Nadine; MAILHES, Corinne. Identification of Harmonics and Sidebands in a Finite Set of Spectral Components. 2013, vol. 1.

26. HÁJEK, Miroslav; BALÁŽ, Marcel. *Spracovanie dát generovaných senzorovou IoT sietou*. Bratislava, 2022. Bachelor's thesis. Faculty of Informatics and Information Technologies, Slovak University of Technology.

27. SHI, Xiangfu; ZHANG, Zhen; XIA, Zhiling; LI, Binhua; GU, Xin; SHI, Tingna. Application of Teager–Kaiser Energy Operator in the Early Fault Diagnosis of Rolling Bearings. *Sensors*. 2022, vol. 22, no. 17, p. 6673. ISSN 1424-8220. Available from DOI: 10.3390/s22176673.

28. YU, Gang. A Concentrated Time–Frequency Analysis Tool for Bearing Fault Diagnosis. *IEEE Transactions on Instrumentation and Measurement*. 2020, vol. 69, no. 2, pp. 371–381. ISSN 1557-9662. Available from DOI: 10.1109/TIM.2019.2901514.

29. ARTS, Lukas P. A.; van den BROEK, Egon L. The Fast Continuous Wavelet Transformation (fCWT) for Real-Time, High-Quality, Noise-Resistant Time–Frequency Analysis. *Nature Computational Science*. 2022, vol. 2, no. 1, pp. 47–58. ISSN 2662-8457. Available from DOI: 10.1038/s43588-021-00183-z.

30. HERRERA, Roberto; BAAN, Mirko; HAN, Jiajun. Applications of the Synchrosqueezing Transform in Seismic Time-Frequency Analysis. *Geophysics*. 2014, vol. 79, pp. V55–V64. Available from DOI: 10.1190/geo2013-0204.1.

31. GOUMAS, Stefanos; ZERVAKIS, Michalis; STAVRAKAKIS, G. Classification of Washing Machines Vibration Signals Using Discrete Wavelet Analysis for Feature Extraction. *IEEE Transactions on Instrumentation and Measurement*. 2002, vol. 51, pp. 497–508. Available from DOI: 10.1109/TIM.2002.1017721.

32. YEN, Gary; LIN, K.C. Wavelet Packet Feature Extraction for Vibration Monitoring. *IEEE Transactions on Industrial Electronics*. 2000, vol. 47, pp. 650–667. Available from DOI: 10.1109/41.847906.

33. SONG, Yongxing; LIU, Jingting; WU, Dazhuan; ZHANG, Linhua. The MFBD: A Novel Weak Features Extraction Method for Rotating Machinery. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*. 2021, vol. 43, no. 12, p. 547. ISSN 1806-3691. Available from DOI: 10.1007/s40430-021-03259-z.

34. ZHUO, Rongjin; DENG, Zhaohui; CHEN, Bing; LIU, Tao; GE, Jimin; LIU, Guoyue; BI, Shenghao. Research on Online Intelligent Monitoring System of Band Saw Blade Wear Status Based on Multi-Feature Fusion of Acoustic Emission Signals. *The International Journal of Advanced Manufacturing Technology*. 2022, vol. 121, no. 7, pp. 4533–4548. ISSN 1433-3015. Available from DOI: 10.1007/s00170-022-09515-3.

35. ZHENG, Alice; CASARI, Amanda. *Feature Engineering for Machine Learning*. O'Reilly Media, 2018. ISBN 978-1-4919-5324-2.

36. KAMMINGA, Jacob W.; LE, Duc V.; MEIJERS, Jan Pieter; BISBY, Helena; MERATNIA, Nirvana; HAVINGA, Paul J.M. Robust Sensor-Orientation-Independent Feature Selection for Animal Activity Recognition on Collar Tags. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 2018, vol. 2, no. 1, 15:1–15:27. Available from DOI: 10.1145/3191747.

37. CALKINS, Keith G. *More Correlation Coeficients (Lesson 13)*. 2005.

38. ROSS, Brian C. Mutual Information between Discrete and Continuous Data Sets. *PLOS ONE*. 2014, vol. 9, no. 2, e87357. ISSN 1932-6203. Available from DOI: 10.1371/journal.pone.0087357. Publisher: Public Library of Science.

39. BREITLING, Rainer; ARMENGAUD, Patrick; AMTMANN, Anna; HERZYK, Pawel. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*. 2004, vol. 573, no. 1-3,

pp. 83–92. ISSN 0014-5793. Available from DOI: 10.1016/j.febslet.2004.07.055.

40. AGGARWAL, Charu C. *Outlier Analysis*. 2nd ed. Springer Publishing Company, Inc., 2016. ISBN 978-3-319-47578-3.

41. TOGBE, Maurras Ulbricht; BARRY, Mariam; BOLY, Aliou; CHABCHOUB, Yousra; CHIKY, Raja; MONTIEL, Jacob; TRAN, Vinh-Thuy. Anomaly Detection for Data Streams Based on Isolation Forest Using Scikit-Multiflow. In: GERVASI, Osvaldo; MURGANTE, Beniamino; MISRA, Sanjay; GARAU, Chiara; BLEČIĆ, Ivan; TANIAR, David; APDUHAN, Bernady O.; ROCHA, Ana Maria A. C.; TARANTINO, Eufemia; TORRE, Carmelo Maria; KARACA, Yeliz (eds.). *Computational Science and Its Applications – ICCSA 2020*. Cham: Springer International Publishing, 2020, vol. 12252, pp. 15–30. ISBN 978-3-030-58810-6. Available from DOI: 10.1007/978-3-030-58811-3_2.

42. AGGARWAL, Charu C.; REDDY, Chandan K.. *Data Clustering - Algorithms and Applications*. CRC Press, 2014. ISBN 978-1-4665-5822-9.

43. ROUSSEEUW, Peter. Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. *Journal of Computational and Applied Mathematics*. 1987, vol. 20, pp. 53–65. Available from DOI: 10.1016/0377-0427(87)90125-7.

44. AMINI, Amineh; WAH, Teh Ying. A Comparative Study of Density-based Clustering Algorithms on Data Streams: Micro-clustering Approaches. In: *Intelligent Control and Innovative Computing*. Ed. by AO, Sio Iong; CASTILLO, Oscar; HUANG, Xu. New York, NY: Springer New York, 2012, pp. 275–287. ISBN 978-1-4614-1695-1. Available from DOI: 10.1007/978-1-4614-1695-1_21.

45. GHESMOUNE, Mohammed; LEBBAH, Mustapha; AZZAG, Hanene. State-of-the-Art on Clustering Data Streams. *Big Data Analytics*. 2016, vol. 1, no. 1, p. 13. ISSN 2058-6345. Available from DOI: 10.1186/s41044-016-0011-3.

46. CAO, Feng; ESTERT, Martin; QIAN, Weining; ZHOU, Aoying. Density-Based Clustering over an Evolving Data Stream with Noise. In: *Proceedings of the 2006 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2006, pp. 328–339. ISBN 978-0-89871-611-5. Available from DOI: 10.1137/1.9781611972764.29.

47. MAURYA, Seetaram; SINGH, Vikas; VERMA, Nishchal K.; MECHEFSKE, Chris K. Condition-Based Monitoring in Variable Machine Running Conditions Using Low-Level Knowledge Transfer With DNN. *IEEE Transactions on Automation Science and Engineering*. 2021, vol. 18, no. 4, pp. 1983–1997. ISSN 1558-3783. Available from DOI: 10.1109/TASE.2020.3028151.

48. SHI, Zhan. Improving k-Nearest Neighbors Algorithm for Imbalanced Data Classification. *IOP Conference Series: Materials Science and Engineering*. 2020, vol. 719, no. 1, p. 012072. ISSN 1757-8981, ISSN 1757-899X. Available from DOI: 10.1088/1757-899X/719/1/012072.

49. CHEN, Lei; LIAN, Xiang. Efficient Processing of Metric Skyline Queries. *IEEE Transactions on Knowledge and Data Engineering*. 2009, vol. 21, pp. 351–365. Available from DOI: 10.1109/TKDE.2008.146.

50. SHENG, Hao; CHEN, Zhongsheng; XIA, Yemei; HE, Jing. Review of Artificial Intelligence-based Bearing Vibration Monitoring. In: *2020 11th International Conference on Prognostics and System Health Management (PHM-2020 Jinan)*. 2020, pp. 58–67. ISSN 2166-5656. Available from DOI: 10.1109/PHM-Jinan48558.2020.00018.

51. ABU ALFEILAT, Haneen Arafat; HASSANAT, Ahmad B.A.; LASASSMEH, Omar; TARAWNEH, Ahmad S.; ALHASANAT, Mahmoud Bashir; EYAL SALMAN, Hamzeh S.; PRASATH, V.B. Surya. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*. 2019, vol. 7, no. 4, pp. 221–248. ISSN 2167-6461, ISSN 2167-647X. Available from DOI: 10.1089/big.2018.0175.

52. JAMIL, Mohd Atif; KHAN, Md Asif Ali; KHANAM, Sidra. Feature-Based Performance of SVM and KNN Classifiers for Diagnosis of Rolling Element Bearing Faults. *Vibroengineering PROCEDIA*. 2021, vol. 39, pp. 36–42. ISSN 2345-0533. Available from DOI: 10.21595/vp.2021.22307.

53. ALTAF, Muhammad; AKRAM, Tallha; KHAN, Muhammad Attique; IQBAL, Muhammad; CH, M. Munawwar Iqbal; HSU, Ching-Hsien. A New Statistical Features Based Approach for Bearing Fault Diagnosis Using Vibration Signals. *Sensors*. 2022, vol. 22, no. 5, p. 2012. ISSN 1424-8220. Available from DOI: 10.3390/s22052012.

54. GEPPERTH, Alexander; HAMMER, Barbara. Incremental learning algorithms and applications. *European Symposium on Artificial Neural Networks (ESANN)*. 2016.

55. BLUM, Avrim; KALAI, Adam; LANGFORD, John. Beating the hold-out: bounds for K-fold and progressive cross-validation. In: *Proceedings of the twelfth annual conference on Computational learning theory*. Santa Cruz California USA: ACM, 1999, pp. 203–208. ISBN 978-1-58113-167-3. Available from DOI: 10.1145/307400.307439.

56. HALFORD, Max. *The correct way to evaluate online machine learning models*. 2020. Section: blog.

57. GAIA. *Online Machine Learning With RiverML by Max Halford*. 2022.

58. RIBEIRO, Felipe; MARINS, Matheus; NETTO, Sergio; SILVA, Eduardo. Rotating Machinery Fault Diagnosis Using Similarity-Based Models. In: *Anais de XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. Sociedade Brasileira de Telecomunicações, 2017. Available from DOI: 10.14209/sbrt.2017.133.

59. *MaFaulDa - Machinery Fault Database*. [N.d.].

60. PESTANA-VIANA, Denys; ZAMBRANO-LOPEZ, Rafael; de LIMA, Amaro A.; DE M. PREGO, Thiago; NETTO, Sergio L.; da SILVA, Eduardo A.B. The Influence of Feature Vector on the Classification of Mechanical Faults Using Neural Networks. In: *2016 IEEE 7th Latin American Symposium on Circuits & Systems (LASCAS)*. Florianopolis: IEEE, 2016, pp. 115–118. ISBN 978-1-4673-7835-2. Available from DOI: 10.1109/LASCAS.2016.7451023.

61. *SpectraQuest Inc.,: Machinery Fault Simulator - Lite*. Available also from: https://spectraquest.com/machinery-fault-simulator/details/mfs-lt/.

62. LOPARO, K.A. *Bearings vibration data set, Case Western Reserve University,* [n.d.].

63. SONG, Renwang; BAI, Xiaolu; ZHANG, Rui; JIA, You; PAN, Lihu; DONG, Zengshou. Bearing Fault Diagnosis Method Based on Multidomain Heterogeneous Information Entropy Fusion and Model Self-Optimisation. *Shock and Vibration*. 2022, vol. 2022, e7214822. ISSN 1070-9622. Available from DOI: 10.1155/2022/7214822.

64. YUHONG, Jin; HOU, Lei; CHEN, Yushu. A New Rotating Machinery Fault Diagnosis Method Based on the Time Series Transformer. 2021.

65. MEY, Oliver; NEUDECK, Willi; SCHNEIDER, André; ENGE-ROSENBLATT, Olaf. Machine Learning-Based Unbalance Detection of a Rotating Shaft Using Vibration Data. In: 2020, pp. 1610–1617. Available from DOI: 10.1109/ETFA46521.2020.9212000.

66. MEY, Oliver; NEUDECK, Willi; SCHNEIDER, André; ENGE-ROSENBLATT, Olaf. *Machine Learning-Based Unbalance Detection of a Rotating Shaft Using Vibration Data*. 2020-07-31. Available from DOI: 10.48550/arXiv.2005.12742.

67. WANG, Xiang; ZHENG, Yuan; ZHAO, Zhenzhou; WANG, Jinping. Bearing Fault Diagnosis Based on Statistical Locally Linear Embedding. 2015, vol. 15, pp. 16225–47. Available from DOI: 10.3390/s150716225.