



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA MECÂNICA

MARIA VITORIA SIKORA

**COMPARISON OF MACHINE LEARNING BINARY
CLASSIFIERS FOR DETECTION OF GEAR DEFECTS**

FLORIANÓPOLIS

2024

Maria Vitoria Sikora

**COMPARISON OF MACHINE LEARNING BINARY CLASSIFIERS FOR
DETECTION OF GEAR DEFECTS**

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Mecânica da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do Grau de Mestre em Engenharia Mecânica.

Orientador: Prof. Júlio Apolinário Cordioli, Dr.
Eng.

Florianópolis

2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Sikora, Maria Vitoria
Comparison of Machine Learning Binary Classifiers for
Detection of Gear Defects / Maria Vitoria Sikora ;
orientador, Júlio Apolinário Cordioli, 2024.
106 p.

Dissertação (mestrado) - Universidade Federal de Santa
Catarina, Centro Tecnológico, Programa de Pós-Graduação em
Engenharia Mecânica, Florianópolis, 2024.

Inclui referências.

1. Engenharia Mecânica. 2. Monitoramento de condição. 3.
Detecção de defeito. 4. Processamento de sinais. 5. Machine
Learning. I. Cordioli, Júlio Apolinário. II. Universidade
Federal de Santa Catarina. Programa de Pós-Graduação em
Engenharia Mecânica. III. Título.

Maria Vitoria Sikora

Comparison of Machine Learning Binary Classifiers for Detection of Gear Defects

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Danilo Silva, Ph.D. Eng.
Universidade Federal de Santa Catarina

Prof. Márcio Holsbach Costa, Dr. Eng.
Universidade Federal de Santa Catarina

Lucas Costa Lobato, Dr. Eng.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Engenharia Mecânica.

Prof. Henrique Simas, Dr. Eng.
Coordenador do Programa

Prof. Júlio Apolinário Cordioli, Dr. Eng.
Orientador

Florianópolis, 02 de julho de 2024.

Dedico este trabalho à minha família.

ACKNOWLEDGEMENTS

First of all, I'd like to thank my family for all the support and encouragement. My parents, Ivone and Vitorio, from whom I feel loved. My brother Elvis, who inspires me, is always my go-to for advice and immensely impacted my decisions before and during my master's. And my extended family (uncles, aunties, and cousins), who were cheering me on. I extend my appreciation to Nicolle Andrade and Victor Dutra, friends who dearly listened to and supported me during the peaks and valleys of my master's journey.

I am thankful to the *Acoustics and Vibration Laboratory* (LVA) for the resources provided for this work. I would like to express my gratitude to Julio Apolinário Cordioli, my supervisor, for the guidance and support during the development of this work.

I would like to thank LVA's friends from Dynamox's project, Miguel Oliva, Lucas Leichtweiss Santos, Lucas Barbeiro, Luiz Fonseca, Fabian Abreu, João Pedro Garcia, Airton Schmitt, and Guilherme Miron for the help. Special thanks to the friends Racquel Knust and Victor Bauler, who greatly contributed to this thesis. Special thanks to all who were directly involved in the measurements and data collection presented in this work.

I am grateful to my friends from *Aeroacoustics*, Gabriel Caldeira (a.k.a. Gabs), Augusto Barth Beck (a.k.a. Gutinho), João Burigo, Isabela Canello Resener, Joanna Oliveira, Gustavo Kirschnick, Nicolas Quintino (a.k.a. Recruta), Lucas Bonomo (a.k.a. Bobs), Lucas Meirelles, Emanoela Teodoro (a.k.a. Manu) and Pedro Oliveira.

Many thanks to the folks from the Multidisciplinary Optimization Group (MOPT), especially Olavo M. Silva for all the support during my masters and the laughs. I would like to thank professor Arcanjo Lenzi (a.k.a. Chefe), for the teachings, kindness and great discussions. I extend my appreciation to LVA's staff, Mariana Farias, Sonia Pereira (a.k.a. Dona Sonia) and Maria Fernanda Vaz. I would like to thank the friends from the laboratory, Ricardo Brum (a.k.a. Ricardinho) and Gildean do Nascimento (a.k.a. Gil). I would also like to acknowledge the people who were not directly mentioned here, but also contributed to my personal and professional growth at LVA.

I am grateful to the Federal University of Santa Catarina (UFSC) and Postgraduate Program in Mechanical Engineering (POSMEC) for the opportunity to study and develop this work. I would like to thank the National Council for Scientific and Technological Development (CNPq) for the scholarship provided for this work. I would also like to express my gratitude to Dynamox, for supporting this work.

“All models are wrong, but some are useful.”

(George Box, 1976)

RESUMO

O monitoramento de condição baseado em vibração destaca-se como uma abordagem de manutenção preditiva devido à sua resposta rápida e relação custo-benefício. Consequentemente, a necessidade de modelos robustos capazes de distinguir entre sinais de vibração saudáveis e defeituosos é de extrema importância. Neste trabalho, avaliamos a eficácia de vários classificadores para diagnosticar estados de saúde de engrenagens usando *features* de sinais de vibração. Duas configurações de caixa de engrenagens são consideradas: uma conectada ao eixo do motor por meio de um sistema de correias e polias e outra diretamente ligada ao eixo do motor. Várias rotações e cargas são estudadas. Um acelerômetro triaxial é posicionado no mancal de rolamento do pinhão e outro no mancal de rolamento da engrenagem. A extração de *features* dos sinais de aceleração abrange *features* estatísticas no domínio do tempo, amplitudes das harmônicas da frequência de engrenamento (GMF) e bandas laterais associadas nos domínios de frequência e ordem, e amplitude das quefrências associadas às bandas laterais da GMF por meio da análise de Cepstrum, além de outras *features* como o FM0, ou o pico espectral e cepstral. Essas *features* servem como entradas para diferentes classificadores de aprendizado de máquina: Logistic Regression, SVM, Random Forest Classifier e XGBoost. A otimização de hiperparâmetros é feita usando um algoritmo de busca aleatória com a Área Sob a Curva ROC (AUC) como parâmetro de otimização. Três diferentes divisões de treino-teste são feitas: (A) uma aleatória, (B) treinamento com dados do sistema de correias e polias e teste com o sistema diretamente ligado ao motor, e (C) treinamento com o sistema diretamente ligado ao motor e teste com o de correias e polias. Os modelos são comparados pelo seu valor de validação AUC, duração do treinamento em segundos, acurácia e acurácia balanceada e valor de AUC de teste. No geral, o XGBoost apresentou os melhores resultados. Na divisão aleatória, alcançou 90% de TPR com 7% de FPR. Isso implica que modelos de árvore podem ser suficientes para descrever o problema, não sendo necessários modelos mais complexos, como redes neurais. Foi realizada uma análise SHAP (SHapley Additive exPlanations) para todas as divisões do XGBoost. *Features* que mostraram algum aspecto da forma do sinal se demonstraram mais importantes na análise SHAP. Isso pode ser devido à natureza dos defeitos analisados, que geram sinais periódicos de impacto. O FM0 também apareceu como muito importante em todas as divisões. *Features* com informações semelhantes apareceram como importantes tanto pelo método de Welch quanto pelo FFT no domínio da frequência. Análises adicionais implicam que elas não são necessárias para a tarefa de classificação, para os defeitos analisados. Investigações mostraram que, embora os modelos treinados com a divisão (B) falhassem mais na classificação do sinal saudável, eles apresentaram desempenho similar ao classificador treinado com a divisão (A) para a detecção de defeitos. O classificador treinado com o conjunto de dados (C) teve o maior número de falsos negativos, ou seja, classificou sinais como saudáveis quando na verdade eram de um pinhão defeituoso. Isso sugere que o conjunto de dados de polia-correia é melhor para generalizar o domínio do que a divisão de acionamento direto.

Palavras-chave: Classificação Binária, Cepstrum, Monitoramento de Condição, Detecção de Defeitos, Engrenagens, Regressão Logística, Machinery Fault Simulator, Aprendizado de Máquina, Sistema de Correia e Polia, Floresta Aleatória, Análise SHAP, SVM, Média Síncrona no Tempo, TSA, XGBoost.

ABSTRACT

Vibration-based monitoring stands out as a predictive maintenance approach in view of its rapid response and cost-effectiveness. Consequently, the need for robust models capable of distinguishing between healthy and defective vibrational signals is of the utmost importance. In this study, we assess the efficacy of various classifiers for diagnosing gear health states using vibration signal features. Two gearbox configurations are considered: one connected to the motor shaft via a pulley belt system and the other directly linked to the motor shaft. Various rotations and loads are studied. One triaxial accelerometer is positioned at the pinion's bearing housing and another at the gear's bearing housing. Feature extraction from the acceleration signals encompasses statistical features in the time-domain, amplitudes of the Gear Meshing Frequency (GMF) harmonics and associated sidebands in the frequency and order-domain, and amplitude of quefrequencies associated with GMF sidebands through Cepstrum analysis, and other features such as the FM0, or the spectral and cepstral peak. These features serve as inputs for different machine learning classifiers: Logistic Regression, SVM, Random Forest Classifier and XGBoost. Hyperparameter tuning is done using a randomized search with the Area Under the ROC Curve (AUC) as the optimization parameter. Three different divisions of train-test are made: (A) randomized one, (B) training with data from the pulley-belt system and testing with the direct-driven system and (C) training with the direct driven system and testing with the pulley-belt one. The models are compared by their validation AUC score, training duration, test accuracy, balanced accuracy and AUC score. Overall XGBoost had the best results. At the random division, it achieved 90% TPR at 7% FPR. It implies that tree models can be sufficient to describe the problem, not requiring more complex models, such as neural networks. A SHAP (SHapley Additive exPlanations) analysis was conducted for all divisions of XGBoost. Features that showed some shape aspect of the signal were more important in the SHAP analysis. This may be due to the nature of the analyzed defects, which results in periodic impact signals. The FM0 also appeared as very important in all divisions. Features with similar information appeared as important both from Welch and FFT's method at the frequency-domain. Further analysis implies that they are not necessary for the classification task given the analysed defects. Additional investigation showed that although the models trained with the (B) division failed more at the healthy signal classification, they had similar performance as the classifier trained with the (A) division for defect detection. The classifier trained with the (C) dataset had the highest number of false negatives, or classified signals as healthy when they were actually from a defective pinion. This suggests that the pulley-belt dataset is better at generalizing the domain than the direct-driven division.

Keywords: Binary Classification, Cepstrum, Condition Monitoring, Defect Detection, Gears, Logistic Regression, Machinery Fault Simulator, Machine Learning, Pulley-belt system, Random Forest Classifier, SHAP Analysis, SVM, Time Synchronous Averaging, TSA, XGBoost.

LIST OF FIGURES

Figure 2.1 – Pitch circle illustration. Source: Author	28
Figure 2.2 – A few types of gears. Adapted from: (COLLINS; BUSBY; STAAB, 2009)	29
Figure 2.3 – Gear trains. Adapted from: (COLLINS; BUSBY; STAAB, 2009)	30
Figure 2.4 – Spectra and cepstra for two truck gearboxes, one with a fault. Source: (RANDALL, 2011)	38
Figure 2.5 – Comparison of data sampling schemes (Δ , uniform Δt vs \square , uniform $\Delta\theta$). Source: (FYFE; MUNCK, 1997)	39
Figure 2.6 – ROC curve. Source: (WIKIPEDIA, 2024b)	42
Figure 2.7 – Standard logistic function representation. Source: (WIKIPEDIA, 2024a)	42
Figure 2.8 – SVM example. Source: (SCIKIT-LEARN, 2024d)	43
Figure 2.9 – Decision tree. Source: (WENIG, 2024)	44
Figure 2.10–Machine learning workflow. Source: (SCIKIT-LEARN, 2024a)	45
Figure 2.11–K-fold validation illustration. Source: (SCIKIT-LEARN, 2024a)	46
Figure 3.1 – Machinery fault simulator with the pulley-belt configuration	47
Figure 3.2 – Opened gearbox	48
Figure 3.3 – Defective straight tooth pinions	48
Figure 3.4 – Pulley-belt configuration details	49
Figure 3.5 – Direct driven gearbox mounting	49
Figure 3.6 – Time-domain acceleration signals at the pinion’s accelerometer vertical direction	53
Figure 4.1 – FFT acceleration signals at the pinion’s accelerometer vertical direction	58
Figure 4.2 – FFT acceleration signals at the pinion’s accelerometer vertical direction	60
Figure 4.3 – Comparison Welch and FFT in a log scale in y-axis plot at the pinion’s accelerometer vertical direction	61
Figure 4.4 – Cepstrum acceleration signals at the pinion’s accelerometer vertical direction	62
Figure 4.5 – Order acceleration signals at the pinion’s accelerometer vertical direction	63
Figure 5.1 – ROC curve for the (A) division.	70
Figure 5.2 – SHAP analysis for the (A) division.	72
Figure 5.3 – Confusion matrix for (A) division and XGBoost.	73
Figure 5.4 – ROC curve for the (B) division.	75

Figure 5.5 – SHAP analysis for the (B) division.	76
Figure 5.6 – Confusion matrix for (B) division and XGBoost.	77
Figure 5.7 – ROC curve for the (C) division.	79
Figure 5.8 – SHAP analysis for the (B) division.	80
Figure 5.9 – Confusion matrix (C) division and XGBoost.	81
Figure C.1 – Confusion matrix for (A) division and Logistic Regression.	97
Figure C.2 – Confusion matrix for pinion condition for (A) division and SVM.	97
Figure C.3 – Confusion matrix for (A) division and Random Forest Classifier.	98
Figure C.4 – Confusion matrix for (B) division and Logistic Regression.	98
Figure C.5 – Confusion matrix for (B) division and SVM.	99
Figure C.6 – Confusion matrix for (B) division and Random Forest Classifier.	99
Figure C.7 – Confusion matrix for (C) division and Logistic Regression.	100
Figure C.8 – Confusion matrix for (C) division and SVM.	100
Figure C.9 – Confusion matrix for (C) division and Random Forest Classifier.	101

LIST OF TABLES

Table 2.1 – Nomenclature of gear failure modes. Source: (AGMA, 1995)	32
Table 2.2 – Summary of the health indicators used for diagnostics of various types of gear failure modes. Adapted from: (KUNDU; DARPE; KULKARNI, 2020)	33
Table 3.1 – Triaxial DeltaTron Accelerometers with TEDS Types 4525-B characteristics. Source: (B&K, 2023)	50
Table 3.2 – PBC 352C33 characteristics. Source: (PCB, 2002)	50
Table 3.3 – Test matrix	51
Table 4.1 – Extracted features from their respective domains.	56
Table 4.2 – Time-domain feature analysis for the direct driven configuration.	57
Table 4.3 – Time-domain feature analysis for the pulley-belt configuration.	57
Table 4.4 – Default hyperparameters search space	65
Table 5.1 – Summary statistics for the train test split variability in (A) division.	69
Table 5.2 – Summary statistics for the hyperparameter search variability in (A) division.	69
Table 5.3 – Summary statistics for the hyperparameter search variability in (B) division.	74
Table 5.4 – Summary statistics for the hyperparameter search variability in (C) division.	78
Table 5.5 – Comparison of fft vs welch method for Random Forest Classifier and XGBoost (A) division.	82
Table A.1 – Influence of number of iterations on Logistic Regression and SVM for the (A) division.	93
Table A.2 – Influence of number of iterations on Random Forest Classifier and XGBoost for the (A) division.	93
Table A.3 – Influence of number of iterations on Logistic Regression and SVM for the (B) division.	93
Table A.4 – Influence of number of iterations on Random Forest Classifier and XGBoost for the (B) division.	94
Table A.5 – Influence of number of iterations on Logistic Regression and SVM for the (C) division.	94

Table A.6–Influence of number of iterations on Random Forest Classifier and XG-Boost for the (C) division.	94
Table B.1–Default hyperparameters for the (A) division.	95
Table B.2–Default hyperparameters for the (B) division.	95
Table B.3–Default hyperparameters for the (C) division.	96
Table D.1–Comparison of FFT vs Welch’s method for Logistic Regression and SVM (A) division.	103
Table D.2–Comparison of FFT vs Welch’s method for Logistic Regression and SVM (B) division.	103
Table D.3–Comparison of FFT vs Welch’s method for Random Forest Classifier and XGBoost (B) division.	103
Table D.4–Comparison of FFT vs Welch’s method for Logistic Regression and SVM (C) division.	104
Table D.5–Comparison of FFT vs Welch’s method for Random Forest Classifier and XGBoost (C) division.	104

CONTENTS

1	INTRODUCTION	23
1.1	OBJECTIVES	25
1.2	SPECIFIC OBJECTIVES	25
2	LITERATURE REVIEW	27
2.1	GEARS	27
2.1.1	Gear types	28
2.1.2	Gearset and gear trains	29
2.1.3	Gear defects	30
2.2	CONDITION MONITORING OF GEARS	31
2.2.1	Vibration-based condition monitoring of gears	31
2.3	SIGNAL PROCESSING TECHNIQUES AND HEALTH INDICATORS	34
2.3.1	Time-domain	34
2.3.2	Frequency-domain	35
2.3.2.1	Fast Fourier Transform (FFT)	36
2.3.2.2	Power spectrum	36
2.3.2.3	Welch's method	37
2.3.2.4	Other frequency-domain health indicators	37
2.3.3	Quefrequency-domain	37
2.3.3.1	Cepstrum	37
2.3.4	Order-domain	38
2.3.5	TSA and order analysis	39
2.4	BINARY CLASSIFICATION	40
2.4.1	Binary classification metrics	40
2.4.1.1	Accuracy and balanced accuracy	41
2.4.1.2	ROC curve and AUC	41
2.4.2	Logistic Regression	42
2.4.3	SVM	43
2.4.4	Decision trees and Random Forest Classifier	43
2.4.5	XGBoost	44

2.5	MACHINE LEARNING WORKFLOW	44
2.5.1	Train test split and cross-validation	45
2.5.2	SHapley Additive exPlanations (SHAP)	46
3	EXPERIMENTAL MATERIALS & METHODS	47
3.1	MACHINERY FAULT SIMULATOR	47
3.1.1	Machine configurations: pulley-belt and direct driven	48
3.2	SENSORS	50
3.3	TEST MATRIX AND EXPERIMENTAL PROCEDURE	50
3.4	SIGNAL REPRESENTATION IN TIME-DOMAIN	52
4	MACHINE LEARNING METHODS	55
4.1	FEATURE EXTRACTION	55
4.1.1	Time-domain	56
4.1.2	Frequency-domain	57
4.1.2.1	Fast Fourier Transform	58
4.1.2.2	Welch’s method	60
4.1.3	Quefrequency-domain: Cepstrum	61
4.1.4	Order-domain: TSA	62
4.2	TRAIN TEST SPLIT	64
4.2.1	Hyperparameter tuning and cross-validation	64
4.3	MODEL PIPELINES	65
4.3.1	Metrics	66
4.3.2	SHAP analysis	66
5	RESULTS & DISCUSSIONS	67
5.1	DEFAULT MODEL ANALYSIS	67
5.1.1	(A) Random division	68
5.1.1.1	Train-test split variability	68
5.1.1.2	Randomized search variability	69
5.1.1.3	ROC curve	70
5.1.1.4	SHAP Analysis	70
5.1.1.5	Confusion matrix	71
5.1.2	(B) Pulley-belt division	74

5.1.2.1	Randomized search variability	74
5.1.2.2	ROC curve	74
5.1.2.3	SHAP Analysis	74
5.1.2.4	Confusion matrix	77
5.1.3	(C) Direct driven division	77
5.1.3.1	Randomized search variability	77
5.1.3.2	ROC curve	78
5.1.3.3	SHAP Analysis	78
5.1.3.4	Confusion matrix	79
5.2	FFT VS WELCH VS NEITHER	82
6	CONCLUSION	83
6.1	FUTURE WORK	84
	Bibliography	87
	APPENDIX A – NUMBER OF ITERATIONS	93
	APPENDIX B – HYPERPARAMETERS	95
	APPENDIX C – CONFUSION MATRICES	97
	APPENDIX D – RESULTS FFT VS WELCH	103

1 INTRODUCTION

According to the Oxford English Dictionary a machine is “*an apparatus using or applying mechanical power and having several parts, each with a definite function and together performing a particular task*”. The parts are bearings, shafts, keys, couplings, gears, etc, which are known as machine elements. A particular task could be the rotation of a component. A motor, for example, generates power and rotates a shaft.

Normally in a machine, there is a need to transfer motion from one shaft to another. Several options are available: flat belts, V-belts, toothed timing belts, chain drives, friction wheel drives and gear drives. While belt and chain drives are usually less pricey, gearsets are compact, slip-free, efficient, light weight, precise in timing, smooth in motion and, therefore, competitive (COLLINS; BUSBY; STAAB, 2009). However, like any component, wear and tear or unexpected loads deviates its behaviour and performance. If not dealt with properly, a catastrophic failure could happen. To avoid this or, less severely, to optimize a component’s useful life, machinery maintenance comes as a solution.

There are mainly four approaches when it comes to machinery maintenance: corrective, preventive, predictive and prescriptive maintenance (DYNAMOX, 2021). Each of these come with its advantages and disadvantages. Nevertheless, predictive maintenance is recognized as the best strategy for the majority of the cases. It anticipates potential failures, optimizes time and financial resources by enabling a planned stop for maintenance, eliminates unnecessary revisions and increases employee safety (RANDALL, 2011). This approach bases itself on the monitoring of several machine parameters, such as temperature, particles in lubricant, acoustic emission or vibration. Among these, the vibration-based condition monitoring is one of the most cost effective, because it can detect the type and location of a defect and responds rapidly to changes in the machine (KUNDU; DARPE; KULKARNI, 2020).

It is possible to measure vibration with proximity probes, velocity transducers, accelerometers, dual vibration probes and laser vibrometers. The most common sensors in condition monitoring are piezoelectric accelerometers, because of their wide frequency and dynamic range. Up until recently, industry used accelerometers for condition monitoring in two arrangements: either a worker went intermittently from machine to machine with an accelerometer and vibration collector, or the accelerometer was fixed on a machine, with continuous or frequent measurements and long carefully placed cables. The first

arrangement risks employees safety and is not suitable to monitor machine's sudden breakdown. The second can continuously monitor machines, but it has the downside of maintenance and cost of cables ([RANDALL, 2011](#)).

Vibration-based condition monitoring is an area of study that has been growing in the past few decades. It is possible to reach conclusions while analysing raw vibration data. However, the use of signal processing techniques considerably eases and enhances the detection and diagnosis of defects ([KUNDU; DARPE; KULKARNI, 2020](#)).

Health indicators can be constructed to assess an asset's health state. They can be classified in five domains: time, frequency, quefrequency, order and time-frequency. In time, some indicators are RMS, kurtosis, crest factor, etc. The frequency-domain can be obtained through Fourier transform, spectral kurtosis, kurtogram, etc. The order domain is a frequency-domain that is scaled by the rotational speed of the machine. The quefrequency domain is obtained by the cepstrum – which will be later explained. The time-frequency domain can be obtained through the short-time Fourier transform, empirical mode decomposition, wavelet transform, etc. In these domains, we can extract health indicators via the amplitude of harmonics or some constructed metrics.

Traditionally, vibration analysts compare vibration signals from a machine, looking for patterns that indicated some type of fault. They employed signal processing techniques to enhance certain signal aspects, took into consideration aspects such as the machine's surrounding, the element's type of coupling, and checked if health indicators were within a threshold. Therefore, research focused on the performance of signal processing techniques and inventing new health indicators. However, the process is time consuming and requires a skilled analyst.

Advances in research prompted the understanding that problems of damage identification are fundamentally ones of statistical pattern recognition ([FARRAR; WORDEN, 2007](#)). [Rytter \(1993\)](#) describes a five step process to assess the damage state of a system: (i) existence, (ii) location, (iii) type, (iv) extent and (v) prognosis. The first step, existence or detection, can be seen as a binary classification problem, the task of categorizing the elements of a set into one of two groups. For example, answering the question “Is it a signal from a healthy component or a defective one?”. When applying a model to classify a health state of machines or components, the health indicators become features – measurable properties – of the models.

Statistical binary classification is a type of supervised learning, where the labels of

the data are well known and defined. There are many tools for binary classification. They can be divided into two categories: shallow learning and deep learning. Shallow learning is a subset of machine learning algorithms that are easy to implement and interpret. They are usually used for simpler tasks, smaller datasets, or situations where interpretability is crucial. Deep learning is a subset of machine learning algorithms that are more complex and require more computational power. They are typically more suitable for complex tasks, large datasets, and scenarios where accuracy and automatic feature learning are essential. Examples of shallow learning algorithms are decision trees, random forests, support vector machines (SVM), logistic regression, probit model, genetic programming, multi-expression programming, or linear genetic programming, etc. Examples of deep learning algorithms are convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM), deep belief networks (DBN), deep Boltzmann machines (DBM), deep neural networks (DNN), etc ([GERÓN, 2019](#)).

This work focuses on defect detection on gears. There are many reviews on the diagnosis (study of type and extent of defect) and prognosis (study of remaining useful life) of gears and rotating machinery, such as the works of [Kumar et al. \(2020\)](#), [Kundu, Darpe and Kulkarni \(2020\)](#) and [Singh et al. \(2021\)](#). These articles usually focus on deep learning approaches – ([HAN; JIANG, et al., 2018](#)), ([HAN; YANG, et al., 2019](#)) and ([ELFORJANI, 2020](#)) – and focus on diagnosis and not detection.

1.1 OBJECTIVES

The objective of this work is to develop and compare the performance of binary classifiers – logistic regression, SVM, random forest classifier, and XGBoost – to detect defects in gears with features extracted from vibration signals' time, frequency, quefreny and order domains.

1.2 SPECIFIC OBJECTIVES

This work's specific objectives are:

- Measuring vibrational signals of healthy and defective gears in a test bench in several configurations;

- Extracting features from the signals in the time, frequency, quefreny and order-domains;
- Implementing binary classification algorithms to detect defects in gears, such as Logistic Regression, SVM, Random Forest Classifier, and XGBoost;
- Comparing the performance of the algorithms in terms of area under the curve, accuracy, balanced accuracy and training time.

2 LITERATURE REVIEW

This Chapter covers an introduction on gears, gear types, gear defects, condition monitoring of gears and a review on vibration-based approaches for diagnostics on gear defects. It also covers signal processing techniques, binary classification, machine learning workflow and SHAP (SHapley Additive exPlanations).

2.1 GEARS

According to (MOTT; VAVREK; WANG, 2017) “Gears are toothed, cylindrical wheels used for transmitting motion and power from one rotating shaft to another”. Gears can change the RPM and movement direction of a shaft with the aid of another gear. In a gearset, there is typically a smaller “gear” that is called pinion and a larger one which is called “gear”. In general, the pinion is the input (driver) and the gear is the output (driven member) (COLLINS; BUSBY; STAAB, 2009).

The Gear Meshing Frequency (GMF), also found as “toothmeshing frequency” in literature, is the rate at which gear and pinion teeth periodically engage:

$$\text{GMF} = f_p N_p = f_g N_g, \quad (2.1)$$

where f_p is the rotation frequency of the pinion, f_g is the rotational frequency of the gear, N_p is the number of teeth in the pinion and N_g is the number of teeth in the gear (COLLINS; BUSBY; STAAB, 2009).

There is a standardized nomenclature for two mating gears geometries’ encounter. The pitch circle is an imaginary circle that passes across the contact points of two mating gears. The pitch point is where the pitch circles are tangent to each other. Figure 2.1 shows the pitch circles of two mating gears, where P is the pitch point and O_1 and O_2 are the center of each gear.

For a gearset to work, the angular velocity ratio of two mating gears must be constant at all instants. That is known as the **fundamental law of gearing** which states that *the angular velocity ratio between the gears of a gearset must remain constant throughout the mesh* (NORTON, 2010). It can be expressed with the following relations:

$$\frac{f_1(t)}{f_2(t)} = \frac{D_2}{D_1} = \frac{N_2}{N_1}, \quad (2.2)$$

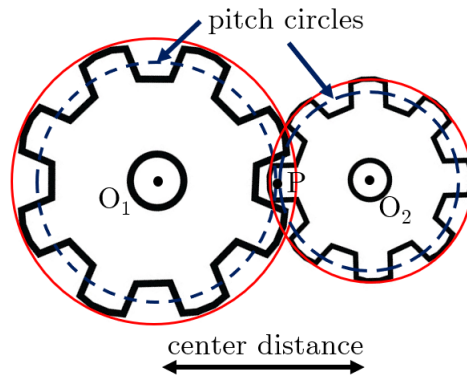


Figure 2.1 – Pitch circle illustration. Source: Author

where the subscripts 1 and 2 represent two mating gears, f is the angular velocity, D is the pitch circle diameter and N is the number of teeth.

2.1.1 Gear types

The type of gear depends on its application scenario and the limitation they have. Some of the design constraints are reduction ratio required, power to be transmitted, rotational speeds, budget, geometric and noise-level limitations. On these accounts, there are three shafting arrangements encountered:

1. Parallel shafts axes

- Straight-tooth spur gears: with an *involute profile*¹, they are easy to design, manufacture and check for precision. It imposes radial load only on supporting bearings. Their speed is usually limited because of noise-levels.
- Helical gears: their teeth are angled to the axis of rotation, forming parallel helical spirals. It imposes both radial and axial loads on supporting bearings because of its angled teeth.

2. Intersecting shafts axes

- Straight bevel gears: the pitch surface is conical frustum². Normal to the tooth axis, the tooth profile resembles an involute. They impose both radial and axial load on supporting bearings.

¹ The involute curve was proposed by L. Euler. It is the trajectory from completely stretched string that is being unwrapped around a circle. It ensures the **fundamental law of gearing**.

² Frustum (pl. frusta) is a truncated solid, usually a cone.

- Spiral bevel gears: also have conical frustum as pitch surface, but the teeth are spiral.

3. Shaft axes are neither parallel nor do they intersect

- Hypoid gears: similar to the spiral bevel gears, but with an offset between the gear axes.

Figure 2.2 gives an example of the above mentioned gear types. Other gear types are face, zero bevel, spiroid, crossed-helical and worm gearsets.

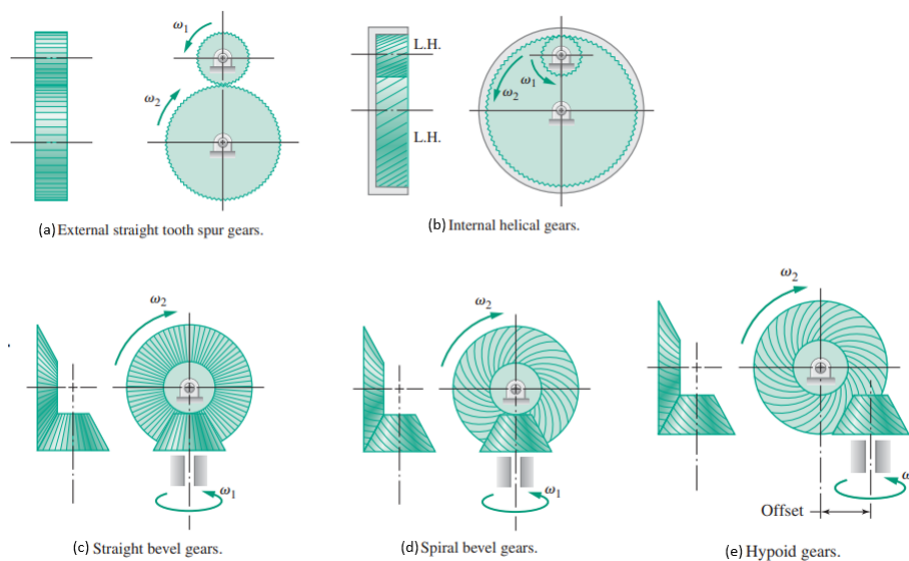


Figure 2.2 – A few types of gears. Adapted from: (COLLINS; BUSBY; STAAB, 2009)

2.1.2 Gearset and gear trains

A gear must always be in contact with another gear in order to fulfill its purpose. A pair of meshing gears is called gearset. A gear train is a series of gearsets positioned in a way to produce a desired output speed, torque and direction of rotation.

Figure 2.3 shows a simple gear train and two compound gear trains. In a simple gear train, such as in Figure 2.3 (a), all gears are mounted on parallel shafts. The iddle gear (2) functions only to reverse the direction of rotation in opposition to a direct mesh of gears (1) and (3), but not the magnitude of the angular velocity ratio.

In a compound gear train, there are at least two gears mounted the same shaft. This imposes the same velocity and direction of rotation on the two gears, which are called compound gear. Figure 2.3 (b) and (c) show reverted and nonreverted compound gear

trains, respectively. Figure 2.3 (b) is reverted because the input shaft and output shaft are colinear (they are the same shaft). Figure 2.3 (c) is nonreverted because the input shaft and output shaft are not colinear (COLLINS; BUSBY; STAAB, 2009). A special case of gear trains is the planetary or epicyclic gear train showed on Figure 2.3 (d), usually applied at wind turbines and automatic transmissions.

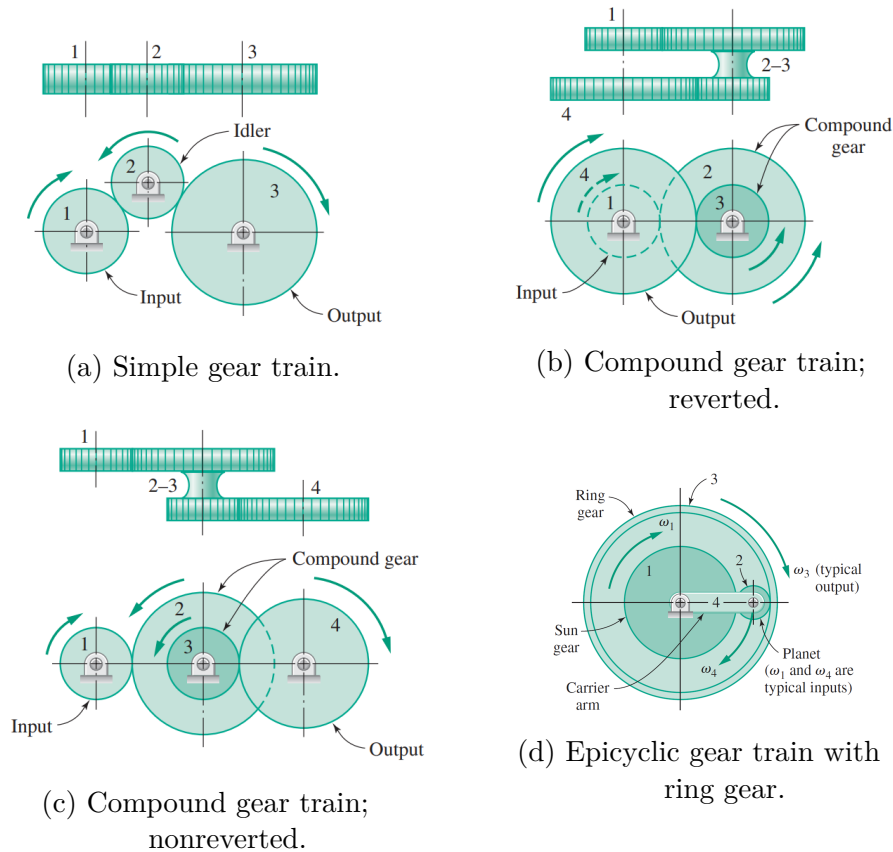


Figure 2.3 – Gear trains. Adapted from: (COLLINS; BUSBY; STAAB, 2009)

2.1.3 Gear defects

To discuss gear defects, let us consider the type of load is imposed on gears. Typically, gears rotate in a single direction. Consequently, the teeth experience bending in only one direction each time they go through the mesh. Due to this non-zero cyclic stresses, the gear root are prone to fatigue failure (COLLINS; BUSBY; STAAB, 2009).

Likewise, the teeth surface receive cyclic loads. When elastic materials come in contact, they deform and produce what is known as *Hertzian contact stresses*³. The subsurface's stresses are higher, consequently cracking the subsurface and generating surface pits. Usually the pinion pits first than the gear, since it has a smaller diameter,

³ Its theory comes from the areas of tribology and contact mechanics.

faster rotation and more frequent toothmesh (COLLINS; BUSBY; STAAB, 2009). Other failures like adhesive and abrasive wear, ghost components⁴, etc may happen.

Alban (1985) divides failure modes in four groups of decreasing occurrence frequency: fatigue, impact, wear and stress rupture. Fatigue can come from tooth bending or surface contact. Impact is random, and, the tooth gets fractured within a few cycles. Wear comes from lack of lubricant or abrasive particles in oil. When internal stresses increase to a magnitude beyond the strength of the material, the tooth ruptures.

Table 2.1 shows part of the “Appearance of Gear Teeth: Terminology of Wear and Failure” from the *American Gear Manufacturers Association Technical Committee*, which classifies and standardizes nomenclature for gear failures.

2.2 CONDITION MONITORING OF GEARS

Diagnostics of gears can be physics-based, data-driven and hybrid approaches (KUNDU; DARPE; KULKARNI, 2020). Diagnosis techniques are different for constant or varying angular speed gears (KUMAR et al., 2020). A gear train is usually immersed in oil, increases its temperature during operation, vibrates and produces noise. As stated at the introduction, it is possible to monitor assets by vibration, oil analysis, noise, acoustic emission (high frequency vibrations), temperature, etc. This work focuses on vibration-based approaches for gears with constant angular speed.

2.2.1 Vibration-based condition monitoring of gears

According to Kundu, Darpe and Kulkarni (2020), in general the vibration signal is quite responsive and contains the most information related to gear dynamics compared to other sensors, like microphones for noise. On the other hand, it requires expert knowledge to extract health indicators, it is direction dependent, and the signal is affected by structural response and mechanical background noise. Raw vibration signal acquired in a gearbox may have the following elements: (1) periodic components due to the meshing pair(s) of gear teeth, (2) periodic impact due to tooth fault and (3) background noise.

It is common to construct some health indicators to distinguish the signal’s elements and assess a gearbox’s health state. This health indicator extraction can be done in different domains: time, frequency, quefrequency, order and time-frequency. It is established

⁴ Because of inadequate manufacturing, gear vibration signal may present a *ghost component*, which appear like GMF components but corresponds to a different number of teeth to those actually cut.

Table 2.1 – Nomenclature of gear failure modes. Source: (AGMA, 1995)

Class	General mode	Specific mode or degree	Not preferred
Wear	Adhesion	Mild Moderate Severe (see scuffing)	Running-in wear
	Abrasion Polishing Corrosion Fretting corrosion Scaling Cavitation Erosion Electrical discharge Rippling	Mild, Moderate, Severe Mild, Moderate, Severe	Scoring Scratching Cutting Burnishing
Scuffing	Scuffing	Mild, Moderate, Severe	Scoring Cold scuffing Hot scuffing Welding Galling Seizing
Plastic deformation	Plastic deformation	Indentation	Bruising Peening Denting Brinelling Permanent deformation Overheating
		Cold flow Hot flow Rolling Tooth hammer Rippling Ridging Burr Root fillet yielding Tip-to-root interference	Fish scaling
Contact fatigue	Pitting (Macropitting)	Initial Progressive Flake Spall	Destructive Arrow head
	Micropitting		Frosting Gray staining Peeling Case crushing
	Subcase fatigue		
Cracking	Hardening cracks Grinding cracks Rim and web cracks Case/core separation Fatigue cracks		Quenching cracks
			Internal rupture
Fracture	Brittle fracture Ductile fracture Mixed mode fracture Tooth shear Fracture after plastic de- formation		Fast fracture Smearing Semi-brittle
Bending fatigue	Low-cycle fatigue High-cycle fatigue	Root fillet cracks Profile cracks Tooth end cracks	

that increases in the amplitude of significant components, such as the GMF or its sidebands, indicate deterioration (RANDALL, 2011).

Table 2.2 shows a list of signal processing techniques and metrics for gear signals encountered in the literature. The acronyms are: root mean square (RMS), time-synchronous averaging (TSA), Short-time Fourier Transform (STFT), Correlation coefficient of residual vibration signal (CCR), gear mesh frequency (GMF), average logarithmic ratio (ALR), complementary ensemble empirical mode decomposition (CEEMD), variable mode decomposition (VMD), orthogonal empirical mode decomposition (OEMD). More about FM0 (which will be further discussed at subsection 2.3.5), NA4, NA4*, FM4, M6A, energy ratio, NB4, NP4 can be found at (SAIT; SHARAF-ELDEEN, 2011).

Table 2.2 – Summary of the health indicators used for diagnostics of various types of gear failure modes. Adapted from: (KUNDU; DARPE; KULKARNI, 2020)

Type	Indicator name	Fault identified	Data processing	
Time	RMS	General fault progression	Raw vibration signal	
	Kurtosis	Tooth breakage, wear		
	Crest factor (CF)	Localized tooth fault	TSA	
	Energy operator	Scuffing, severe pitting		
	Matched filtered RMS	Wear		
	FM0	Tooth breakage and heavy distributed wear	Residual TSA	
	NA4	Both single teeth pitting and multiple tooth pitting, progressing damage		
	NA4*	Progressing damage	Difference TSA	
	CCR	Pitting progression		
	FM4	Localized pitting or small crack on one or two teeth		
M6A	Surface damage			
Frequency	Energy ratio	Heavy uniform wear	Difference and harmonics TSA	
	NB4	Localized fault		
	GMF harmonics amplitude	Wear	Band pass TSA	
	Sidebands amplitudes	Pitting		
	Side band ratio	Pitting	Raw vibration signal	
	ALR	Crack		
	ALR	Wear	TSA signal	
	Cepstrum	For all kinds of fault		
	Spectral kurtosis	Pitting, crack		
	Phase modulation	Crack		
Time-frequency	NP4	For all kinds of fault	Raw vibration signal	
	Wavelet	For all kinds of fault		
	Empirical mode decomposition	For all kinds of fault	Raw vibration signal, TSA signal	
	CEEMD	For all kinds of fault		
	VMD	For all kinds of fault		
	OEMD	For all kinds of fault		
	Short-time Fourier transform (STFT)	Early-stage fault diagnostics	Band pass TSA signal	
	Winger-Ville Distribution (WVD)	Early-stage fault diagnostics		
			Early-stage fault diagnostics	TSA

2.3 SIGNAL PROCESSING TECHNIQUES AND HEALTH INDICATORS

This section will delve deeper into the signal processing techniques and health indicators mentioned earlier and utilized in this study.

2.3.1 Time-domain

Time-domain health indicators are calculated directly from the time-domain acceleration signal. Most of them are statistics of the signal. Definitions are taken from (MONTGOMERY; RUNGER, 2003), (TAYLOR, 2003) and (MCFADDEN; SMITH, 1985). We define the following:

- Peak to peak (pk_pk): difference between the maximum and minimum values of the signal.

$$\text{pk_pk} = \max(x) - \min(x), \quad (2.3)$$

where x is the signal. The peak to peak is a measure of the signal's amplitude.

- RMS: root mean square of the signal

$$\text{RMS}(x) = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}, \quad (2.4)$$

where x_i is the i -th sample of the signal and N is the number of samples of the signal. The RMS is a measure of the signal's energy.

- Kurtosis: a statistical measure used to describe the shape of a signal's distribution, particularly in terms of the "tailedness" or the presence of outliers.

$$\text{Kurtosis} = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma} \right)^4, \quad (2.5)$$

where x_i is the i -th sample of the signal, μ is the mean and σ is the standard deviation. For a given signal, kurtosis quantifies how much of the signal's variance is due to extreme values (tails) compared to a normal distribution.

- Skewness: measure that describes the asymmetry of a signal's distribution around its mean.

$$\text{Skewness} = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma} \right)^3, \quad (2.6)$$

where x_i is the i -th sample of the signal, μ is the mean and σ is the standard deviation.

- Shape factor: ratio of the RMS to the mean of the absolute signal

$$\text{Shape factor} = \frac{\text{RMS}(x)}{\frac{1}{N} \sum_{i=1}^n |x_i|}. \quad (2.7)$$

It is a measure of the signal's shape.

- Crest factor: ratio of the peak to the RMS of the signal

$$\text{CF} = \frac{\max |x|}{\text{RMS}}. \quad (2.8)$$

It is the extent to which a waveform's peak amplitude exceeds its average or RMS (root mean square) value. It is a dimensionless measure.

- Impulse factor: used to characterize the sharpness or spikiness of a signal. It is particularly useful in identifying signals that contain sudden, high-energy impulses or transients.

$$\text{Impulse factor} = \frac{\max |x|}{\frac{1}{N} \sum_{i=1}^n |x_i|}. \quad (2.9)$$

In summary, the impulse factor provides insight into the presence and severity of sudden, high-energy components within a signal, making it a valuable tool in various fields for detecting and analyzing transient events.

- Clearance factor: quantifies the sharpness or peakiness of a signal relative to the overall energy content, specifically emphasizing the presence of high-amplitude transients.

$$\text{Clearance factor} = \frac{\max |x_i|}{\left(\frac{1}{N} \sum_{i=1}^n \sqrt{|x_i|} \right)^2}. \quad (2.10)$$

2.3.2 Frequency-domain

This section presents the frequency-domain signal processing techniques and its health indicators.

2.3.2.1 Fast Fourier Transform (FFT)

The Fast Fourier Transform (FFT) is an efficient algorithm to compute the Discrete Fourier Transform (DFT). The DFT is a numerical tool to compute the Fourier transform, inverting the domain of a function – from x to $1/x$. For example, from time-domain to the frequency-domain or from distance to wavelength. It enables the analysis of the frequency content of a signal. The DFT is defined as:

$$X(f) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi fn/N}, \quad (2.11)$$

where $X(f)$ is the f -th frequency component of the signal, $x(n)$ is the n -th sample of the signal and j is the imaginary unit (SHIN; HAMMOND, 2008). The FFT is a complex-valued function, but it is common to take the magnitude of the FFT to analyze the signal's frequency content. The magnitude of the FFT is defined as

$$|X(f)| = \sqrt{\text{Re}(X(f))^2 + \text{Im}(X(f))^2}, \quad (2.12)$$

where $\text{Re}(X(f))$ is the real part of $X(f)$ and $\text{Im}(X(f))$ is the imaginary part of $X(f)$.

In the case of condition monitoring rotating machinery, periodic components related to the rotation of the shaft, bearings, gearbox, belts, etc., are present in the vibration signal. When a defect starts to occur at a given component, usually the amplitude of the corresponding frequency component increases. In the case of gearboxes, we pay attention to the amplitude of GMF, the rotation frequency of the gears and their harmonics (RANDALL, 2011).

2.3.2.2 Power spectrum

The power spectrum is the Fourier Transform of the autocorrelation of a signal. The autocorrelation is a measure of the similarity between samples of a variable as a function of the time delay between them. It helps to find repeating patterns in a signal and dismiss noise-like ones. The power spectrum gives a description of the energy distribution along frequency. It is useful to identify the signal's spectral content masked by noise (SHIN; HAMMOND, 2008). The power spectrum can be used in Welch's method and to calculate the cepstrum, which will be later introduced.

2.3.2.3 Welch’s method

The Welch’s method is a technique to estimate the power spectrum. It divides the signal into overlapping segments, computes the FFT of each segment and averages them to estimate the power spectrum. The Welch’s method is useful to reduce the variance of the power spectrum estimate (WELCH, 1967).

2.3.2.4 Other frequency-domain health indicators

Other frequency-domain health indicators are the spectral flatness and the RMS carpet. The RMS carpet (SYLVESTER; PEARCE, 2024) calculates the root mean square of a moving average in the frequency-domain. Spectral flatness (S. DUBNOV, 2004), also known as the tonality coefficient, is a metric designed to gauge the degree of noise-like qualities in a sound, as opposed to tone-like qualities. A spectral flatness of one indicates a white noise-like spectrum. It is calculated by the geometric average divided by the arithmetic mean of the spectrum magnitude:

$$\text{Spectral flatness} = \frac{\sqrt[N]{x_1 x_2 \cdots x_N}}{\frac{1}{N} \sum_{i=1}^N x_i}, \quad (2.13)$$

where x_i is the i -th sample of the spectral magnitude.

2.3.3 Quefrequency-domain

This section introduces the cepstrum and how health indicators can be obtained through it.

2.3.3.1 Cepstrum

First designed for speech analysis (RANDALL, 2017), cepstrum is a tool for detecting periodicity in a spectrum, such as uniform spaced families of harmonics. According to Randall (1973), “*the cepstrum is defined in a number of different ways, but all can be considered as a spectrum of a logarithmic spectrum (i.e. logarithmic amplitude, but linear frequency scale)*”. Its original mathematical definition was:

$$C(\tau) = |\mathcal{F} \{\log (F_{xx}(f))\}|^2, \quad (2.14)$$

where \mathcal{F} is the Fourier transform and F_{xx} is the power spectrum. The τ has a dimension of time but is named quefrequency.

The quefrency gives information about frequency spacing and not absolute frequency. Small uniform peak spacing in the spectrum are said to have high quefrency and large uniform frequency spacing between peaks are said to have low quefrency. For example, peaks in the cepstrum come from families of sidebands. The inverse quefrency value of a peak represents a uniform spacing at the frequency-domain.

Figure 2.4 shows (a) a faulty gearbox and (b) a gearbox in good condition. The left side shows the spectra and the right side shows the cepstra. The cepstrum immediately gives an accurate value of the average spacing of all members of a given family of sidebands without having to find the individual members of the family.

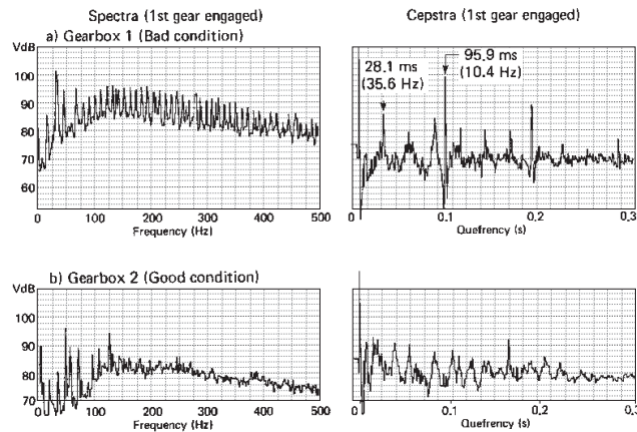


Figure 2.4 – Spectra and cepstra for two truck gearboxes, one with a fault. Source: (RANDALL, 2011)

As a consequence of the kinematics, vibration of gearboxes has modulations. Thus the spectra have sidebands around GMF and increases in these sidebands indicate deterioration. The spacing of the sidebands give valuable information of the source of the defect. The advantage of Cepstrum over Spectrum is that it is less sensitive to transmission path effects because the Cepstrum component corresponding to a given sideband is an average sideband height over the whole spectrum (RANDALL, 1973). It is important to emphasize that Cepstrum may not be the best choice when harmonics or sidebands are not well defined.

2.3.4 Order-domain

This section introduces the TSA and order analysis and its advantages to the frequency-domain methods.

2.3.5 TSA and order analysis

Regarding rotating machinery, a single rotation period encompasses all interactions between machinery components. Any noise, disturbance, or periodic signal content incongruent with the rotation is eliminated by averaging across uniform rotation angles or complete rotations, as opposed to arbitrary time segments. This method of averaging is known as time-synchronous averaging (TSA).

The TSA effectiveness depends on the equality in the number of samples in each segment corresponding to the rotation period. However, it is common for this equality not to be met due to various factors. For example, the analyzed signal may exhibit small fluctuations in the extraction frequency, resulting in variations in the value of the period such as illustrated in Figure 2.5. To address this issue, it is possible to manipulate the signal applying interpolation techniques to ensure an equal number of samples in each segment known as angular resampling (DOMINGUES, 2023).

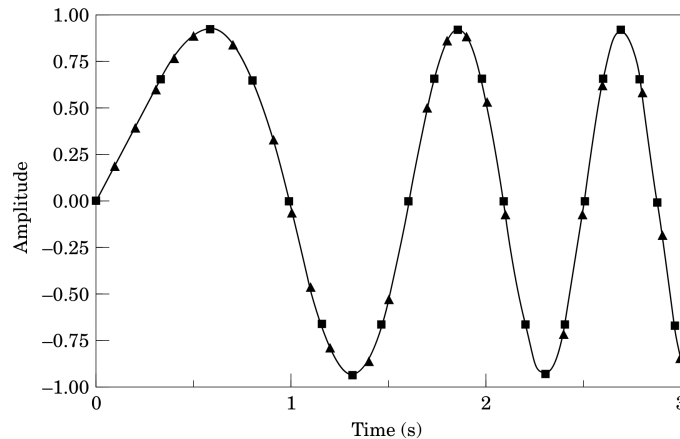


Figure 2.5 – Comparison of data sampling schemes (\triangle , uniform Δt vs \square , uniform $\Delta\theta$). Source: (FYFE; MUNCK, 1997)

In the context of gear analysis, Stewart (1977) first proposed the FM0 (Figure of Merit 0) as a health indicator. The FM0 is calculated as:

$$\text{FM0} = \frac{\text{pk_pk}}{\sum_{i=1}^N \text{GMF}(i)}, \quad (2.15)$$

where pk_pk is the peak-to-peak value, $\text{GMF}(i)$ is the amplitude of the i -th harmonic of the GMF on the spectrum of the TSA signal and N is the total number of harmonics. It increases when heavy wear occurs due to the the peak-to-peak value staying constant and the GMF shrinking. Conversely, it is not as sensitive for minor tooth damage.

It is possible to derive the difference signal and residual signal from the TSA signal. The residual signal removes the shaft frequency and its harmonics, gear mesh frequencies and their harmonics from the TSA signal. The difference signal removes shaft frequency and its harmonics, gear frequencies and their harmonics, first sidebands on gear frequencies and their harmonics from the TSA signal (KUNDU; DARPE; KULKARNI, 2020). They are most useful with geartrains with several gears. Sait and Sharaf-Eldeen (2011) detail the health indicators which are derived from them.

2.4 BINARY CLASSIFICATION

Classification involves determining the category or subpopulation to which an observation or observations belong. When there are only two categories to choose from, it is known as binary classification. In the context of this study, the categories are defective and healthy components. The goal is to predict the category of a new observation based on the features extracted from the vibration signal.

2.4.1 Binary classification metrics

A variety of metrics are available for evaluating and comparing models. It falls upon the modeler to select a metric that highlights the problem’s scope and definition. In binary classification, we typically delineate the categories as positive and negative, and the samples in these categories will be named P and N, respectively. In the context of detection, we designate “Positive” for defective components and “Negative” for healthy components. This study primarily concentrates on accuracy, balanced accuracy, the ROC curve, and the AUC score (area under the ROC curve).

In order to construct metrics, it proves beneficial to define certain concepts. True Positive (TP) represents the quantity of data correctly classified as positive, while True Negative (TN) indicates the quantity of data correctly classified as negative. False Positive (FP) signifies the quantity of data erroneously classified as positive, and False Negative (FN) denotes the quantity of data erroneously classified as negative.

Further definitions encompass the True Positive Rate and the False Positive Rate. The True Positive Rate (TPR), also recognized as sensitivity, recall, or probability of detection, represents the ratio of True Positive to the sum of True Positive and False Negative. Conversely, the False Positive Rate (FPR), also termed the probability of false

alarm, reflects the ratio of False Positive to the sum of False Positive and True Negative (GRANDINI; BAGLI; VISANI, 2020). They are defined as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{P}}, \quad \text{and} \quad (2.16)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{\text{N}}. \quad (2.17)$$

2.4.1.1 Accuracy and balanced accuracy

Accuracy and balanced accuracy are commonly used metrics to assess the performance of binary classifiers. Accuracy measures the proportion of correctly classified instances among all instances. It is calculated as

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (2.18)$$

Balanced accuracy, on the other hand, is a metric that considers the class imbalance in the dataset. It calculates the TPR and TNR to provide a more balanced assessment of classifier performance. It is defined as

$$\text{BA} = \frac{\text{TPR} + \text{TNR}}{2}. \quad (2.19)$$

Balanced accuracy is particularly useful when dealing with imbalanced datasets, where one class significantly outnumbers the other. In summary, accuracy considers all predictions equally, while balanced accuracy gives equal weight to both classes, providing a more fair evaluation when dealing with skewed datasets.

2.4.1.2 ROC curve and AUC

The receiver operating characteristic (ROC) curve depicts the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis, while varying the threshold of classification. This threshold is determined by the probability distribution of both True Positive and False Positive outcomes (FAWCETT, 2006). As exemplified in Figure 2.6, proximity to the top-left corner (TPR=1 and FPR=0), indicates superior classifier performance (GERÓN, 2019). To summarize the information derived from the ROC curve, which has two components (TPR and FPR), the Area Under the Curve (AUC) is defined as the area under the ROC curve.

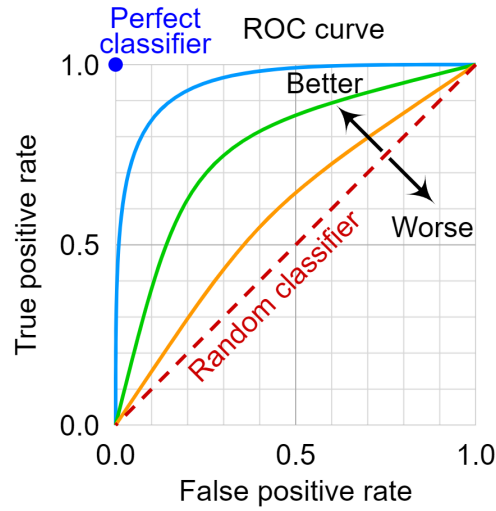


Figure 2.6 – ROC curve. Source: ([WIKIPEDIA, 2024b](#))

2.4.2 Logistic Regression

Logistic regression applies a linear combination of variables (features, in this context) to a sigmoid function:

$$\sigma(t) = \frac{1}{1 + e^{-(x-\mu)/s}} \quad (2.20)$$

where μ is a location parameters and s is a scale parameter. A standard logistic function ($\mu = 0$ and $s = 1$) is depicted at [Figure 2.7](#). Although it is a continuous function, it finds utility in binary classification by establishing a threshold along the curve. In [Figure 2.7](#) we could designate 0.5 as the threshold, where values equal to or greater than it would be classified as defective, while those lower would be regarded as healthy.

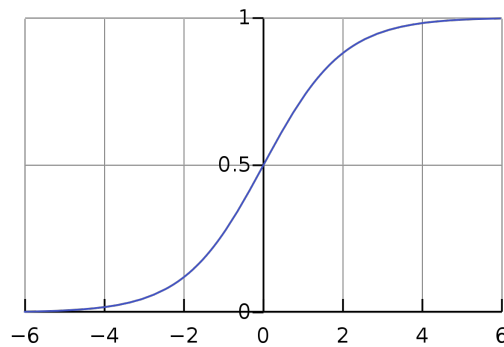


Figure 2.7 – Standard logistic function representation. Source: ([WIKIPEDIA, 2024a](#))

2.4.3 SVM

Support Vector Machine (SVM) is a classification model capable of operating both linearly and nonlinearly. It functions by identifying hyperplanes that effectively separate data points by minimizing the distance between the classification borders. Consider the illustration in [Figure 2.8](#): various hyperplanes can potentially separate the brown and blue classes. In essence, a separation that will broadly generalize the classifier is the one in which the hyperplane maximizes the distance to the nearest training data points (known as the functional margin) ([SCIKIT-LEARN, 2024d](#)). Moreover, SVM can accommodate nonlinear kernels and exhibits sensitivity to scaling.

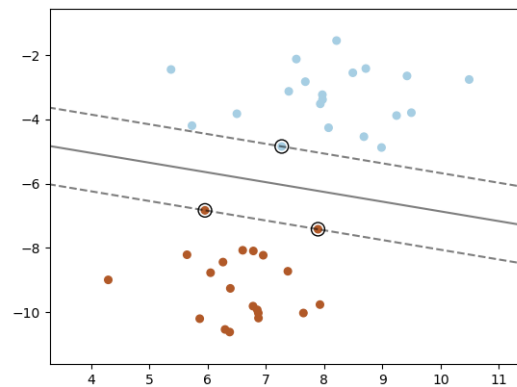


Figure 2.8 – SVM example. Source: ([SCIKIT-LEARN, 2024d](#))

2.4.4 Decision trees and Random Forest Classifier

The Random Forest model, utilized for classification or regression tasks, employs decision trees as its foundational components. Decision trees originate from a guiding question or premise, such as "should I accept this new job offer?" as depicted in [Figure 2.9](#). Alternatively, they may address inquiries like "does this signal originate from a defective component?". To address such queries, a decision tree commences with a root node. Subsequently, branches are formed from this root node based on available features (e.g., "is the RMS > 2.5?"), which are designated as decision nodes. The evaluations conducted at each decision node aim to generate homogeneous subsets. At the terminus of each branch lies a leaf node, representing potential outcomes of the model. This structure emulates human decision-making processes, enhancing interpretability while necessitating minimal data preprocessing ([JAMES et al., 2023](#)). However, it should be noted that Random Forest

employs a greedy search strategy, which can entail higher training costs compared to other algorithms (IBM, 2024).

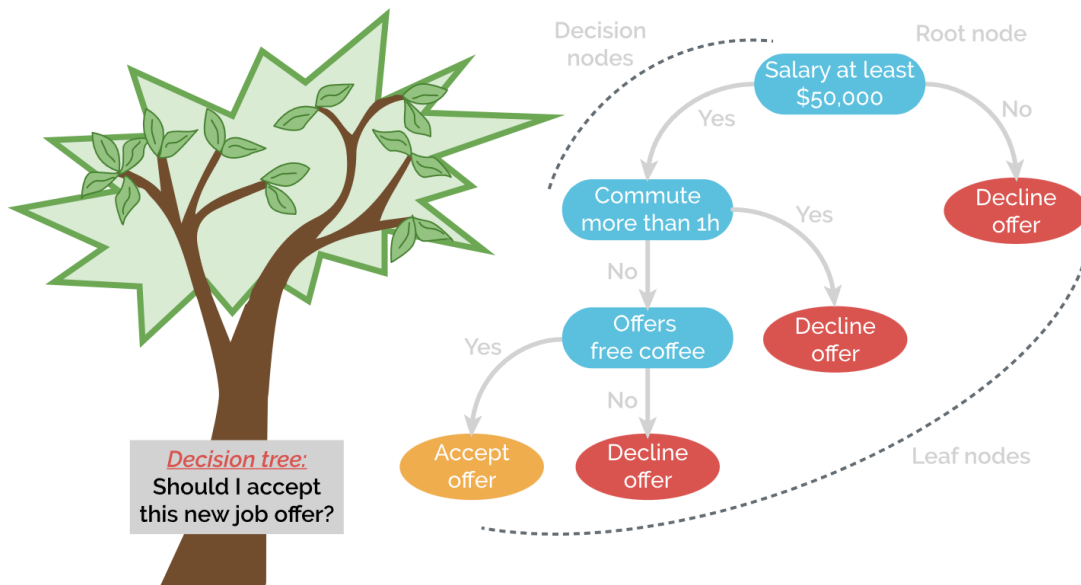


Figure 2.9 – Decision tree. Source: (WENIG, 2024)

2.4.5 XGBoost

XGBoost, abbreviated from "eXtreme Gradient Boosting," represents a decision tree model that employs a boosting algorithm. Deviating from the parallel training approach of Random Forests, the boosting algorithm sequentially trains models, iteratively correcting errors and refining performance with each iteration. It leverages gradient descent optimization techniques to ascertain the direction of error reduction. XGBoost is renowned for its efficiency, scalability, and high performance, making it a popular choice for various machine learning tasks (CHEN; GUESTRIN, 2016).

2.5 MACHINE LEARNING WORKFLOW

A typical machine learning methodology follows a prescribed workflow, as illustrated in Figure 2.10. Initially, a dataset (in our case the health indicators extracted from vibration signals) is partitioned into training and test subsets. Subsequently, the model undergoes training using a cross-validation strategy, which will be further explained at subsection 2.5.1. In summary, it subdivides the training data-set into N groups, trains a model with a set of hyperparameters (denoted as "Parameters" in Figure 2.10 and will be further explained in the next section) with data from $N - 1$ groups and then predicts

and calculates the chosen metrics (such as accuracy, balanced accuracy or AUC score) in the group that was left out of training – which is called validation. This process is done iteratively at the groups. After all groups were used in validation, the process starts again with a different set of hyperparameters. After iterating this process multiple times, the hyperparameters yielding the highest performance scores are selected. The model is then retrained using the entire training dataset. Finally, the trained model is applied to the test data, where its performance is evaluated to assess its efficacy.

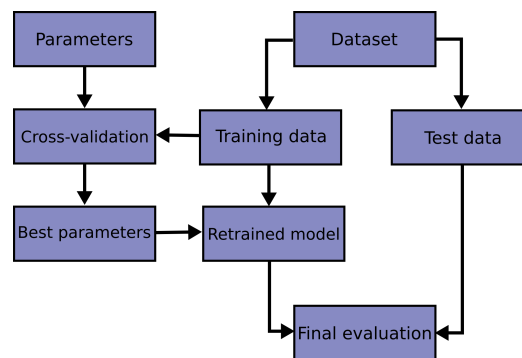


Figure 2.10 – Machine learning workflow. Source: (SCIKIT-LEARN, 2024a)

2.5.1 Train test split and cross-validation

Each model is trained with a specific set of hyperparameters. Hyperparameters serve to regulate the learning process, determining aspects such as the number of splits a tree is permitted to make (referred to as max depth in tree models). This is in contrast to model parameters, which are derived from the learning process itself, such as the coefficients of a logistic regression. Given the infinite possibilities of hyperparameters, which can significantly impact model results, hyperparameter tuning is commonly employed as a crucial step in the model development process.

According to the Oxford dictionary, prediction entails "saying or estimating that (a specified thing) will happen in the future or will be a consequence of something". In the context of predictive modeling, the objective is to assess the performance of a model that has been trained, or fitted, on one dataset and tested on an unseen dataset. To mitigate the risks of overfitting and selection bias, cross-validation is employed. Typically, all available data is partitioned into training, validation, and test sets.

Figure 2.11 depicts an example of k-fold cross-validation, where the training set is divided into $k=5$ folds. Utilizing a set of hyperparameters, a model is trained on 4

of the folds and tested on the 5th fold, with this process iterated across all folds. The average of the results obtained on the tested folds constitutes the performance metric for the validation set. Subsequently, after training the model multiple times with various hyperparameter combinations, the best model is selected based on a specified evaluation metric. This optimal model is subsequently evaluated on the test set.

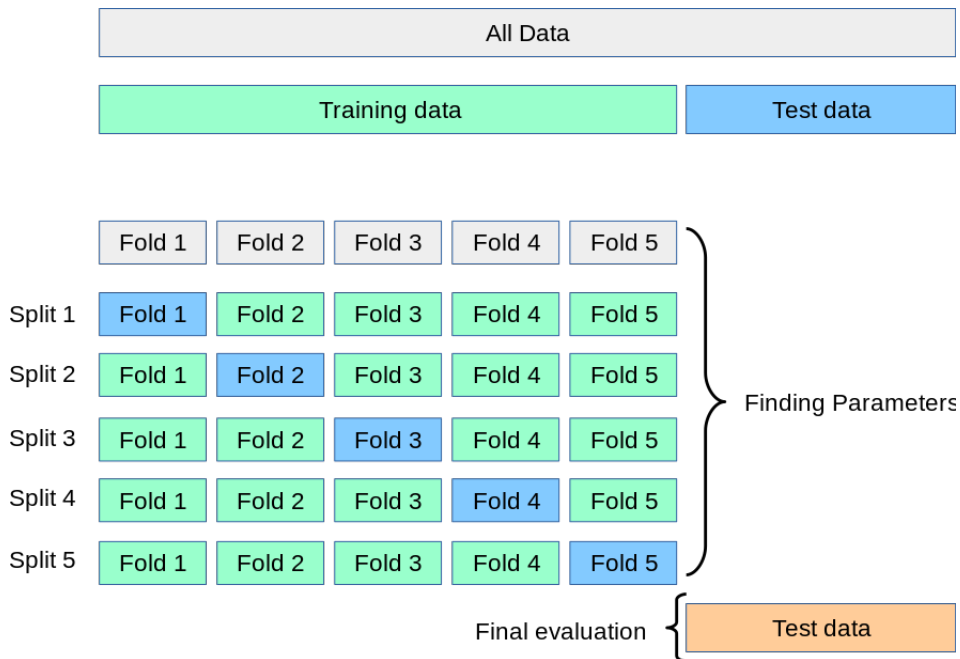


Figure 2.11 – K-fold validation illustration. Source: (SCIKIT-LEARN, 2024a)

2.5.2 SHapley Additive exPlanations (SHAP)

After implementation, Machine Learning models often function similarly to a black box. To enhance comprehension of these models, evaluating the importance of each feature proves insightful. Various methods exist for this purpose. For instance, interpreting the coefficients within a regression model. However, such approaches can potentially lead to misinterpretations. Additionally, they solely consider the overall importance of features, neglecting the potential impact of varying feature values on model outcomes.

Drawing from game theory, SHAP (SHapley Additive exPlanations) values offer a means to address these limitations. These values estimate the influence of feature inputs on the model's outcomes, thereby facilitating the ranking of feature importance. They also provide insights into how higher or lower feature values may affect the model's predictions. Nevertheless, it is important to note that SHAP values do not directly assess the quality of model predictions (SHAP, 2024).

3 EXPERIMENTAL MATERIALS & METHODS

In this Chapter, we list the workflow and equipment that was used in this study.

3.1 MACHINERY FAULT SIMULATOR

This work utilizes *SpectraQuest's Machinery Fault Simulator* (SPECTRAQUEST, 2023) (see Figure 3.1) as a test bench that simulates industrial machinery in a controlled environment, allowing for an in-depth understanding of machine behavior. Several configurations are possible, enabling study of different machine components and their faults. For operation, it uses a motor driven by a frequency inverter, allowing users to control the rotational speed.

To study gear fault detection, our research group opted to use a gearbox with a pair of straight bevel gears, as shown in Figure 3.2. The pinion has 20 teeth and the gear, 30. There is a healthy pinion, a pinion with a missing tooth, see Figure 3.3 (a), and one with a chipped tooth, see Figure 3.3 (b). The gear remained healthy and unchanged throughout the experiments.

The gearbox input shaft is coupled with the pinion. Additionally, there is a magnetic brake associated with the gear shaft, from Magtork, model MTL 10-5/8 (MAGTORK, 2023). Its function is to generate load in the system. The load is variable, varying from 0,5 lb – in to 10 lb – in (0,06 Nm to 1,2 Nm) of torque. The magnetic brake's scale can be set from 0 (minimum load) to 5 (maximum load). Based on the experience of the

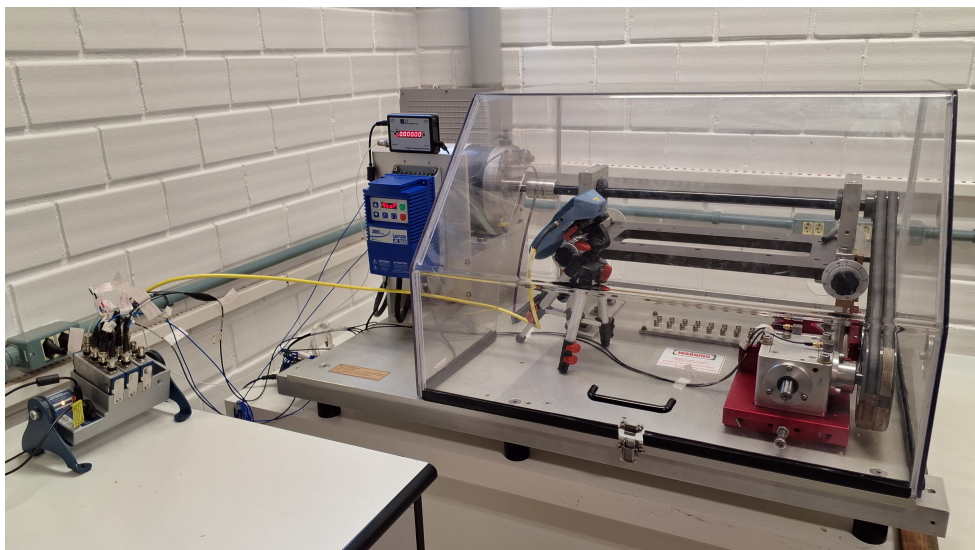


Figure 3.1 – Machinery fault simulator with the pulley-belt configuration

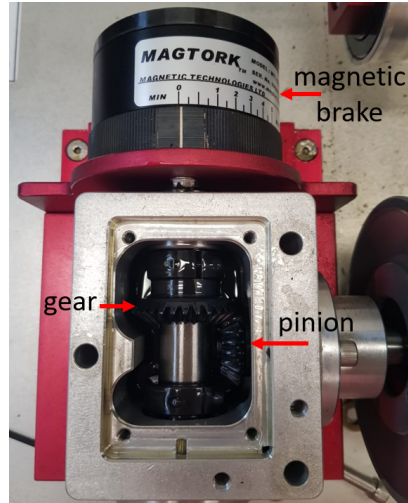


Figure 3.2 – Opened gearbox



(a) Pinion with a chipped tooth



(b) Pinion with a missing tooth

Figure 3.3 – Defective straight tooth pinions

authors, the magnetic brake's adjustment scale has a nonlinear relation to the torque. This information will be later used to properly define the loads applied to the system.

3.1.1 Machine configurations: pulley-belt and direct driven

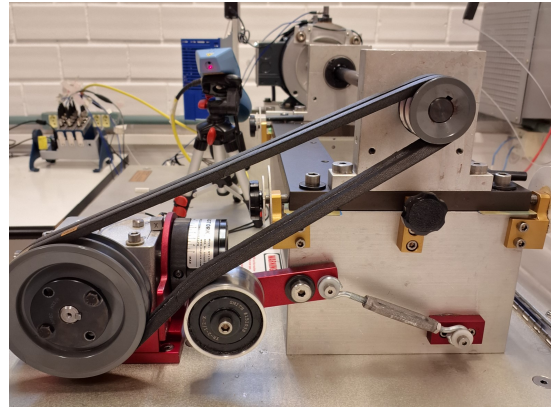
Among the several possible machine configurations, this work adopts two configurations: the pulley-belt (see [Figure 3.4](#)) and the direct driven (see [Figure 3.5](#)).

At the pulley-belt configuration, the motor is coupled by a beam coupling to a shaft supported by two bearing housings. At the shaft's opposite side of the motor, there is a pulley-belt system, as depicted in [Figure 3.4](#) (b), which is linked to the gearbox. The pulleys from the motor shaft and pinion shaft have diameters of 5 and 12.5 cm, respectively. They share the same linear velocity, but the angular velocity takes into consideration the pulleys' radii. Therefore, the f_{pinion} at the pulley-belt system is $40\%f_{motor}$. For the pulley-belt configuration, there were a tachometer pointed at the motor's shaft and another pointed at the pinion's pulley, as shown in [Figure 3.4](#) (a).

At the direct driven configuration, the motor is coupled directly to the gearbox by a



(a) Tachometer pointed at the pinion pulley



(b) Close-up of the pulley-belt

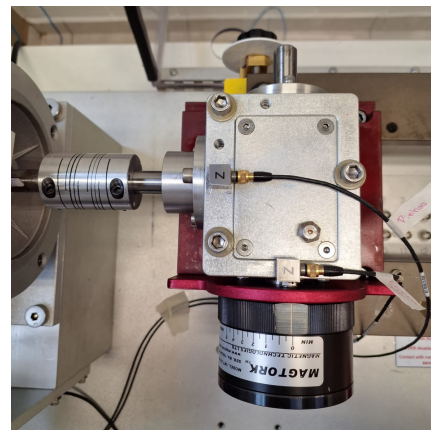
Figure 3.4 – Pulley-belt configuration details

beam coupling, as shown in Figure 3.5 (a). A tachometer was directed at the motor's shaft. Unlike the pulley-belt configuration, the direct driven configuration enables the study of gearbox vibration signatures without any belt, shaft, or bearing imprint, which will have their own vibration signature.

Another difference between the direct driven and the pulley-belt systems is the higher slippage in the pulley-belt system. This means that the rotational variation is higher in the pulley-belt system. The setups mimic two different machines. Therefore, it allows for the study of the model's generalization capability. It is also possible to investigate how each feature is able to differentiate vibration signals from different machines.



(a) Isometric view



(b) Top view

Figure 3.5 – Direct driven gearbox mounting

3.2 SENSORS

This work utilized tachometers, as well as uniaxial and triaxial accelerometers. Triaxial accelerometers 4525-B (characteristics and information are available on [Table 3.1](#)) from *Brüel & Kjaer* were placed on top of the pinion and gear’s bearing housing, as shown in [Figure 3.5](#) (b). The accelerometers were calibrated before utilization.

Table 3.1 – Triaxial DeltaTron Accelerometers with TEDS Types 4525-B characteristics. Source: ([B&K, 2023](#))

Characteristics	
Dimension	21,1 mm (cable entrance) × 12,2 mm × 12,2 mm
Weight	6 g
Operational temperature	$-54\text{ °C} \leq T \leq 121\text{ °C}$
Sensitivity X- axis	9,897 mV/g (1,009 mV/m/s ²)
Sensitivity Y- axis	10,58 mV/g (1,079 mV/m/s ²)
Sensitivity Z- axis	10,19 mV/g (1,039 mV/m/s ²)
Frequency range ($\pm 10\%$)	up to 10 kHz
Dynamic Range	$\pm 500\text{ g}$

At the motor, we used *PCB Piezotronics* model 352C33 (characteristics and information are available on [Table 3.2](#)) uniaxial accelerometers – one at the drive end and one at the non drive end. For the data analysis, we did not use the motor’s accelerometers data. Nevertheless, it will be stated here to possibly guide future works.



Table 3.2 – PBC 352C33 characteristics. Source: ([PCB, 2002](#))

Characteristics	
Dimension	15,7 mm (height) × 12,2 mm (hex)
Weight	5,8 g
Operational temperature	$-54\text{ °C} \leq T \leq 93\text{ °C}$
Sensitivity	100 mV/g (10,2 mV/m/s ²)
Frequency range ($\pm 10\%$)	up to 15 kHz
Dynamic Range	$\pm 50\text{ g pk}$

3.3 TEST MATRIX AND EXPERIMENTAL PROCEDURE

Considering the different parameters available (configuration, load, rotational speed, and pinion condition), we decided on the test matrix described in [Table 3.3](#). Initially we assembled a configuration with a given load. The screws were tightened with a screwdriver with torque control to ensure the same mounting conditions. For the pulley-belt configuration, we adjusted the belt tightening screws up to a mark.

Table 3.3 – Test matrix

Configuration	Load	RPM	Condition	Replicate
Pulley-belt system 	0	500, 1000,	Healthy	3x
	2.5	1500,	Chipped tooth	2x
	4	2000,		
	5	2500, 3000, 3500	Missing tooth	2x
Direct-driven 	0	200, 400, 500, 600,	Healthy	3x
	2.5	800, 1000, 1200,	Chipped tooth	2x
	4	1400, 1500,		
	5	2000, 2500, 2000, 2500, 3000, 3500	Missing tooth	2x

For the measurements, we turned on the motor, waited for the RPM to stabilize at the tachometer display, and then collected the data. An user interface was developed at the *Acoustics and Vibration Laboratory (LVA)* to configurate and save the metadata for each measurement. We used a *National Instruments* acquisition board. The measurement and its metadata was then saved in a database. This was done for a set of RPMs. Then, we changed the load and repeated the process. The signals have 30 seconds and a sampling rate of 25 600 Hz. The backend system employed in the user interface measures an additional duration of approximately five seconds. However, the initial and final 2.5 of the measurement are excluded from the saved signal.

Replicates are important to evaluate the repeatability of the data acquisition system and the machine behavior. To ensure the repeatability and to try to balance our data distribution, we collected three replicates for the healthy pinion condition and two for the defective ones. The replicate recordings were collected under different mountings (disassembling and assembling the machine), and same parameters.

We selected seven different RPMs – 500, 1000, 1500, 2000, 2500, 3000, and 3500 – at the motor for the pulley-belt configuration. Due to speed reduction from the motor to the pinion shaft in the pulley-belt configuration, we chose thirteen RPMs – 200, 400, 500, 600, 800, 1000, 1200, 1400, 1500, 2000, 2500, 3000, and 3500 – at the motor for the direct-driven configuration. We selected four different loads, including minimum, maximum, and two intermediate levels. Some combinations of parameters were not collected either because the motor did not endure the configuration or due to problems when storing the data. These include:

- From the direct driven configuration:
 - Missing pinion tooth load 4 and 5 for 200 rpm and 2nd replicate;
 - Chipped pinion tooth load 4 and 5 200 rpm and 2nd replicate;
 - Healthy pinion tooth load 4 and 5 200 rpm and 3rd replicate;
 - All rpms for the healthy condition and load 5.

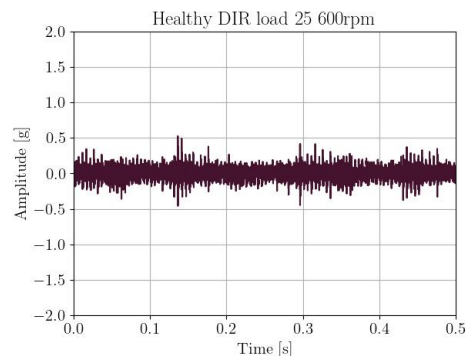
As we solely analyze the accelerometers fixed to the gearbox, we obtained six signals (comprising three channels per triaxial accelerometer) for a set of parameters. Following the dataset cleansing, we retained 3234 signals, consisting of 1890 defective and 1344 healthy ones. It is noteworthy that the dataset utilized exhibits an imbalance, which could potentially influence the performance of the models.

3.4 SIGNAL REPRESENTATION IN TIME-DOMAIN

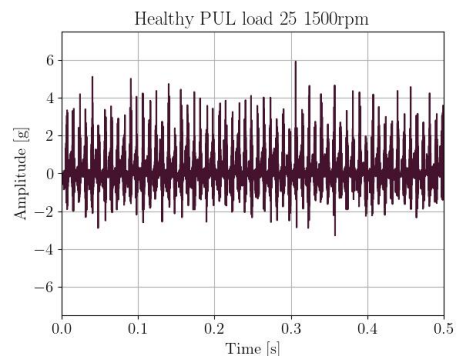
[Figure 3.6](#) shows the signals in the time-domain in the vertical direction on top of the pinion. The signals are from the direct-driven configuration on the left and pulley-belt configuration on the right. From top to bottom we have the healthy, chipped, and missing conditions consecutively. The rotational speed is 600 rpm for the direct driven configuration and 1500 rpm for the pulley-belt configuration, so it is possible to compare vibration signature with the same pinion rotation and load 2.5 on both. The signals are from the triaxial accelerometer fixed to the gearbox. It is important to pay attention to the plot's scale. Adjustments were made to improve signal visualization and comparison, while preserving information.

In the direct driven configuration, the healthy condition shows a relatively constant amplitude. At the chipped condition, there seems to be some modulation in the same periodicity as the pinions rotation, which is 10 Hz. The missing pinion presents peaks at the same periodicity as the pinion rotation, probably coming from the impact caused by the missing tooth.

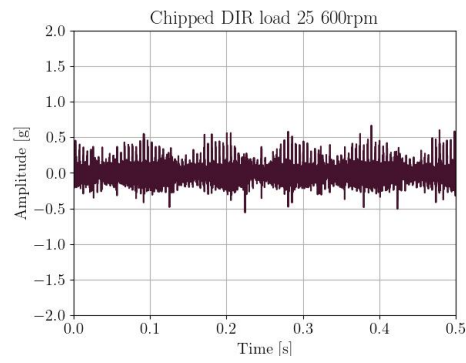
Visually, the signals from the pulley-belt configuration are not as visually differentiable. They seem however to have higher peaks above zero, than bellow. The amplitudes are greater at the healthy and pinion condition, when compared to the direct driven configuration.



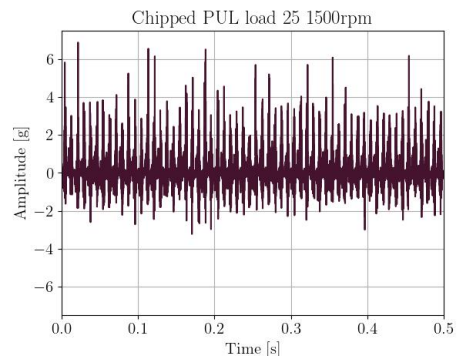
(a) Direct driven healthy pinion



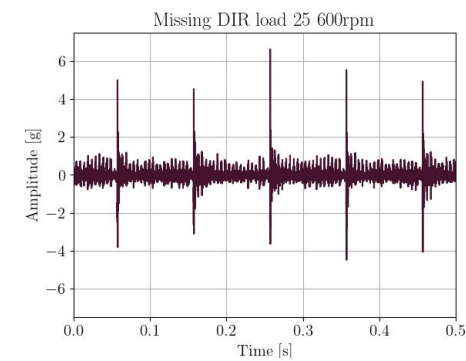
(b) Pulley-belt healthy pinion



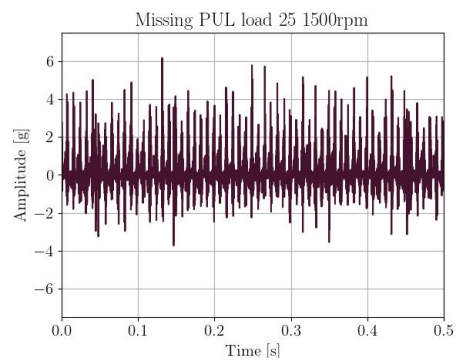
(c) Direct driven chipped pinion



(d) Pulley-belt chipped pinion



(e) Direct driven missing pinion



(f) Pulley-belt missing pinion

Figure 3.6 – Time-domain acceleration signals at the pinion’s accelerometer vertical direction

4 MACHINE LEARNING METHODS

This Chapter details the machine learning methods used in this work. Feature extraction, train-test split, hyperparameter tuning, cross-validation, model selection, and pipelines are described. The metrics used to evaluate the models are also presented. Finally, the SHAP analysis is explained.

4.1 FEATURE EXTRACTION

This section details the process of feature extraction. Each section delves the signal processing methods and analyses the values of some features from that domain. In the plots DIR and PUL represent the direct driven and pulley-belt configurations, respectively. The conditions are represented by H, C, and M, which stand for healthy, chipped, and missing tooth, respectively. To avoid overloading the text, the plots and tables showed in this section are only from the vertical direction of the accelerometer at the pinion and first mounting assembly. It is important to remind that the conical form of the gears apply load in the three directions at the mesh.

The signals were accessed at the database to then be processed. The signals were processed in the time, frequency, quefrequency, and order domains. In the frequency-domain, we used the Fast Fourier Transform (FFT) and Welch's method. Cepstrum was used to analyse the signals in the quefrequency-domain and TSA for the order-domain.

Table 4.1 shows the features extracted from each domain, where GMF, f_p , and f_g represent gear meshing frequency's, pinion frequency's, and gear frequency's amplitudes respectively. The numbers (1, 2, 3) represent the extracted harmonics (multiples) of those quefrequencies. So, for example, 2X GMF represents the amplitude of twice the quefrequency associated with the GMF (1/GMF, or the inverse of GMF). The crest factor plus is a variation of the crest factor. The calculation method for this parameter will remain confidential.

Similarly, a reasoning is made with the metrics on the frequency and order-domain. So, for example, (1, 2, 3)X f_p means that the amplitude associated with the first three harmonics of the pinion frequency was extracted. Some GMF sidebands associated with the gear and pinion frequency were also extracted. To account for possible rotation speed variations, we utilized the average rotation speed from the tachometer to calculate the

Table 4.1 – Extracted features from their respective domains.

Time	Frequency	Quefrequency	Order
peak-to-peak, kurtosis, RMS, crest factor, crest factor plus, skewness, shape factor, impulse factor, clearance factor	RMS carpet, spectral flatness, (1, 2, 3)X GMF, (1, 2, 3)X f_p , (1, 2, 3)X f_g , (1, 2, 3)X GMF \pm (1, 2, 3)X f_p , (1, 2, 3)X GMF \pm (1, 2, 3)X f_g	(1, 2, 3)X GMF, (1, 2, 3)X f_p , (1, 2, 3)X f_g , cepstral peak	(1, 2, 3)X GMF, (1, 2, 3)X f_p , FM0

metrics.

There are 9 features from the time-domain, 48 features from the frequency-domain which were extracted from the FFT and the Welch method, 10 features from the quefrequency-domain, and 7 features from the order-domain. This totals 122 features analyzed. The next section will give a more detailed explanation in how these metrics are extracted from each domain.

4.1.1 Time-domain

The time-domain features come from the raw signal. The features extracted from this domain are peak-to-peak, kurtosis, RMS, crest factor, crest factor plus, skewness, shape factor, impulse factor, and clearance factor. In this section, we will delve into the RMS, kurtosis, shape factor and skewness.

Table 4.2 shows the time-domain feature analysis for the direct driven configuration at 1000 rpm. In general, they seem to increase with the severity of the defect. This behaviour is not observed in the skewness for loads 4 and 5. The kurtosis for missing condition seems to decrease in value as the load increases. The shape factor is higher for the missing condition, and the skewness is higher for the chipped condition.

Table 4.3 shows the time-domain feature analysis for the pulley-belt configuration at 1000 rpm. In the case of this division, there is no clear pattern of increase in the kurtosis values with the severity of the defect. This is expected since their plots differed significantly from each other. Furthermore, there seems to be two outliers in the RMS: missing condition for load 0 and 5. The skewness even yielded negative values for the chipped condition at 0 load and missing condition at load 5.

Table 4.2 – Time-domain feature analysis for the direct driven configuration.

config.	cond.	rpm	load	RMS	kurtosis	shape_factor	skewness
DIR	H	1,000	0.0	0.10	4.35	1.30	0.36
DIR	C	1,000	0.0	0.19	12.25	1.42	0.87
DIR	M	1,000	0.0	0.86	54.21	1.78	1.71
DIR	H	1,000	2.5	0.14	5.38	1.33	0.16
DIR	C	1,000	2.5	0.21	8.01	1.43	0.26
DIR	M	1,000	2.5	0.87	42.02	1.67	0.89
DIR	H	1,000	4.0	0.15	4.57	1.32	0.26
DIR	C	1,000	4.0	0.28	9.73	1.44	0.54
DIR	M	1,000	4.0	0.83	25.15	1.54	0.23
DIR	H	1,000	5.0	0.18	5.83	1.36	0.35
DIR	C	1,000	5.0	0.29	9.54	1.48	0.20
DIR	M	1,000	5.0	0.88	21.67	1.52	0.30

Table 4.3 – Time-domain feature analysis for the pulley-belt configuration.

config.	cond.	rpm	load	RMS	kurtosis	shape_factor	skewness
PUL	H	1,000	0.0	0.74	10.32	1.79	1.61
PUL	C	1,000	0.0	8.22	7.60	1.45	-0.06
PUL	M	1,000	0.0	0.72	11.33	1.80	1.72
PUL	H	1,000	2.5	0.73	13.43	1.81	2.01
PUL	C	1,000	2.5	0.71	15.15	1.82	2.15
PUL	M	1,000	2.5	0.70	15.94	1.84	2.32
PUL	H	1,000	4.0	0.76	14.21	1.80	2.17
PUL	C	1,000	4.0	0.76	13.98	1.80	2.06
PUL	M	1,000	4.0	0.72	16.07	1.84	2.36
PUL	H	1,000	5.0	0.62	15.62	1.83	2.37
PUL	C	1,000	5.0	0.77	13.73	1.79	2.05
PUL	M	1,000	5.0	8.15	3.37	1.26	-0.25

4.1.2 Frequency-domain

At the frequency-domain, we extracted the RMS carpet, spectral flatness as well as the first three harmonics associated with the GMF, pinion frequency, represented as f_p , gear frequency, represented as f_g . The pinion frequency is acquired from the tachometer, and the gear frequency and GMF are calculated by the gearing ratio and number of teeth. The features were extracted from “pure” FFT and from Welch’s method. More details on the methods can be found in the next subsections.

The harmonics from the pinion and gear frequency were estimated as the maximum value in the range of $\pm 5\%$ of the expected associated frequency. The GMF was estimated by the maximum value in the range $\pm 0,5$ Hz around its frequency value. The different

range applied to the GMF was due to the fact that the GMF is close to sidebands, and could be mistakenly identified as a sideband. Specially at higher frequencies, the range of range of $\pm 5\%$ could be too wide and include peaks greater than but that were not a GMF.

We also extracted the sidebands associated with pinion and gear frequencies around the GMF. We extracted sidebands around the first three harmonics of the GMF. We considered sidebands $\pm (1, 2, 3)X$ from the pinion and gear frequencies. So, for example, the second GMF's harmonic sideband associated with a "negative" third pinion harmonic would be represented as $2 * \text{GMF} - 3X f_p$. The sidebands were extracted in a similar way as the GMF harmonics: by the maximum value in the range $\pm 0,5 \text{ Hz}$ around the sideband frequency value.

4.1.2.1 Fast Fourier Transform

At the FFT, we utilized a rectangular window. [Figure 4.1](#) shows the full range of the frequency spectrum for the direct driven and pulley-belt configurations. High peaks around 6kHz and 12kHz are frequencies associated with the motor's inverter. It is noticeable that the pulley-belt configuration has a higher amplitude in general than the direct driven configuration. It is noteworthy since they have the same rpm at the motor, therefore, the pinion's should be rotating at 400 rpm.

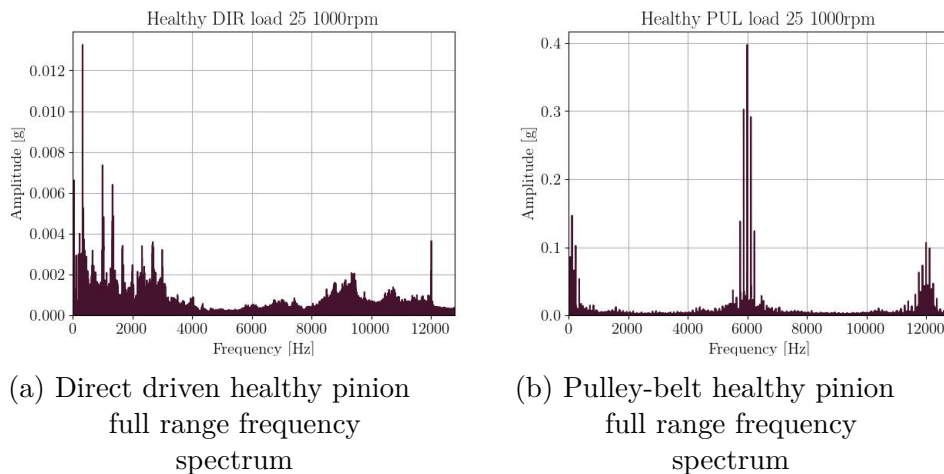


Figure 4.1 – FFT acceleration signals at the pinion's accelerometer vertical direction

In order to better visualize the signals, we limited the frequency range to a bit more than 3 times the GMF. This way, we could better visualize the GMF and its harmonics. [Figure 4.2](#) shows the FFT signals for the direct driven and pulley-belt configurations.

The first three harmonics of the GMF are 333.3, 666.6 and 1000 Hz for the direct driven configuration and 133.3, 266.6 and 400 Hz for the pulley-belt configuration.

The amplitude of the pulley-belt configuration is higher than the direct driven configuration. There is not a clear differentiation in general amplitude for the healthy and chipped conditions at the direct driven configuration. The missing condition, however, seems to have a higher amplitude than the other conditions. The highest peak in both conditions seem to be between 350 and 400 Hz. Sidebands around the GMF's third harmonic are more prominent.

At the pulley-belt configuration, it seems that there are some frequencies with much more influence when compared to others. There seems to be peaks near first and second harmonics of the GMF, but only the second harmonic seems to increase with defect severity.

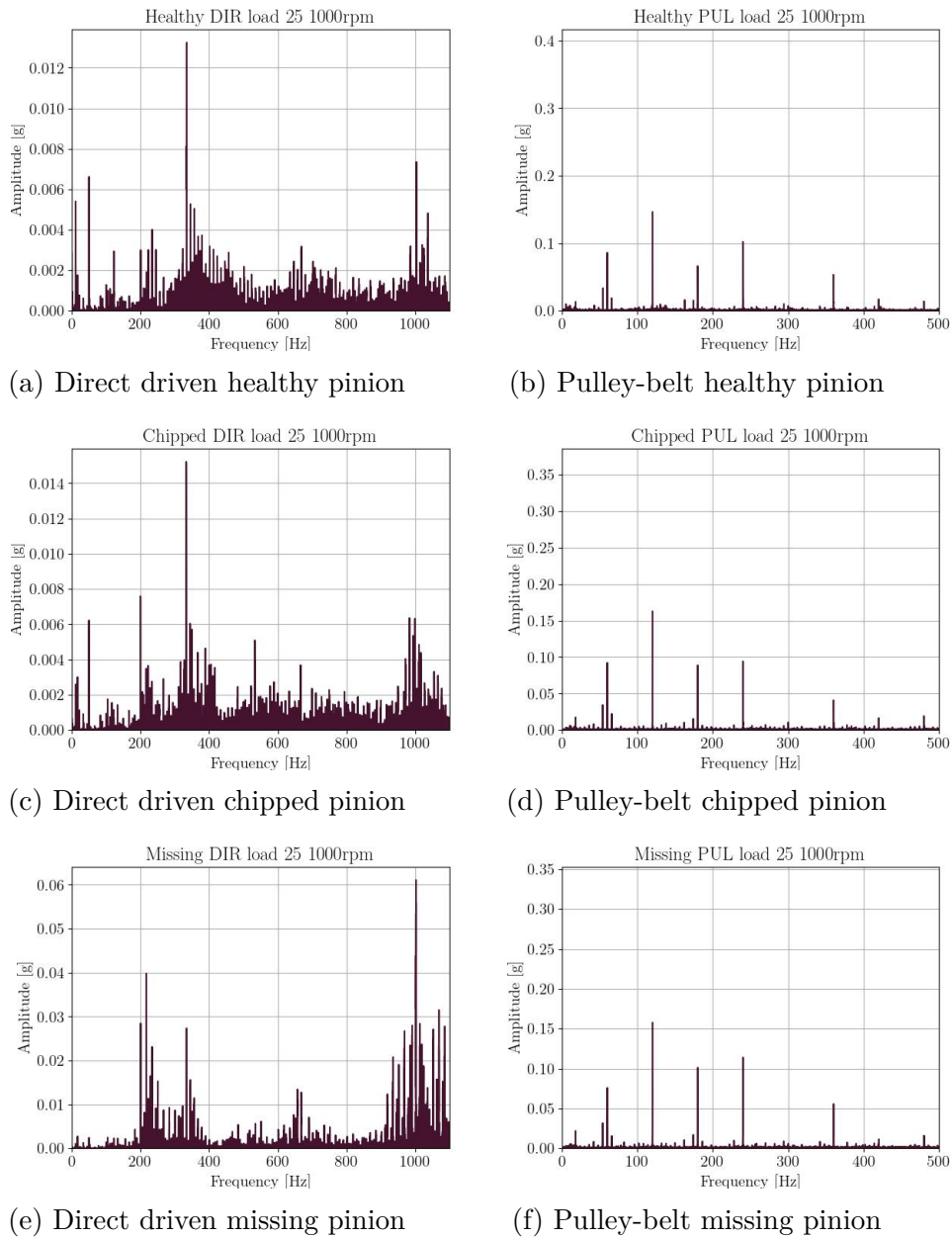


Figure 4.2 – FFT acceleration signals at the pinion’s accelerometer vertical direction

4.1.2.2 Welch’s method

For the Welch’s method we utilized a Hanning window, a number of segments to achieve 0,5 Hz resolution at the frequency and 90 percent overlap between segments. Since FFT and Welch carry similar information, we will not delve into the same details as we did for the FFT. Nevertheless, to emphasize their difference, we compare both methods with a log scale in y-axis plot (Figure 4.3). The noise is significantly reduced while still preserving the curve’s shape. Note that the welch method shows the plot in power ($[g^2]$) while the FFT shows the amplitude ($[g]$).

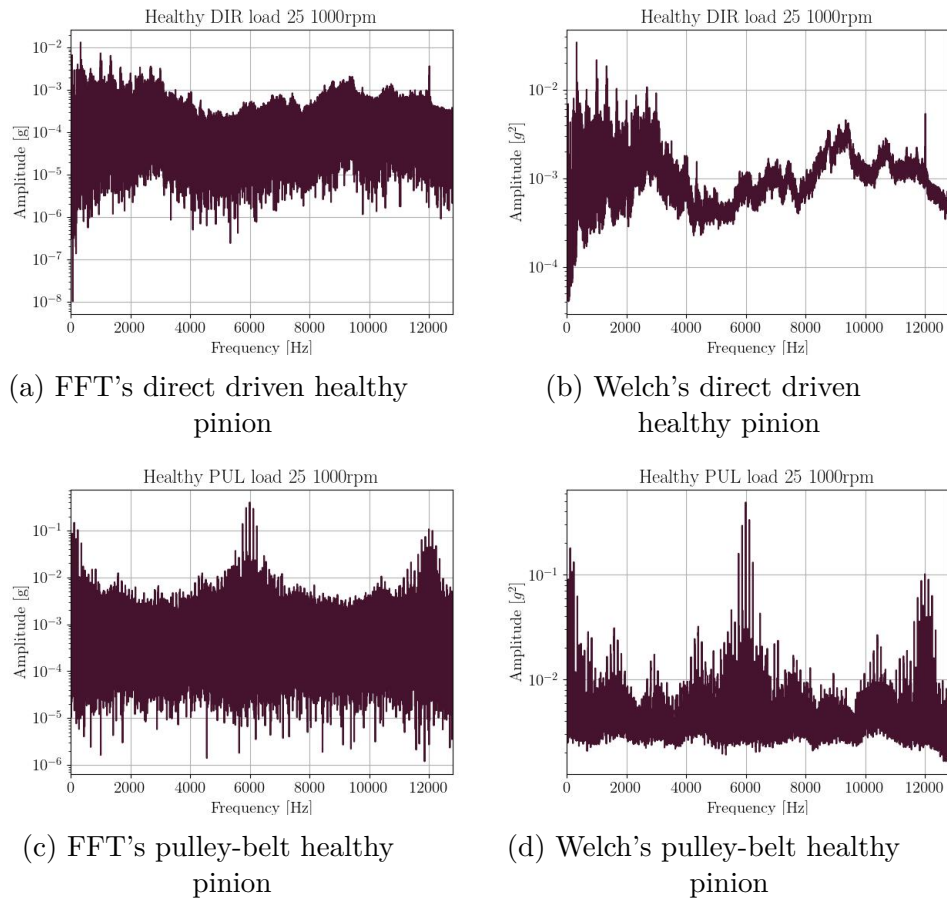
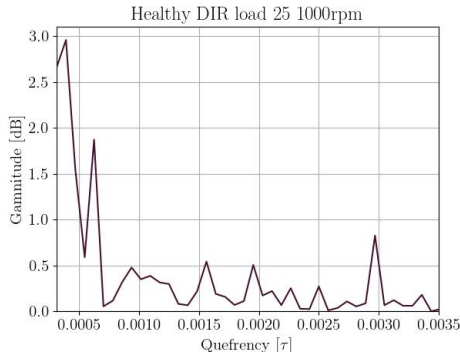


Figure 4.3 – Comparison Welch and FFT in a log scale in y-axis plot at the pinion's accelerometer vertical direction

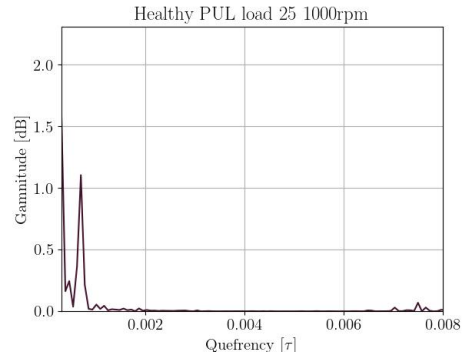
4.1.3 Quefrequency-domain: Cepstrum

In the quefrequency-domain, we have extracted amplitudes (or gamnitudes) of quefrequencies which are the inverse of frequencies of interest – first three harmonics of GMF, pinion frequency, and gear frequency. These gamnitudes were calculated based on the maximum value within a bandwidth of $\pm 5\%$ of the associated quefrequency. The cepstral peak was also extracted.

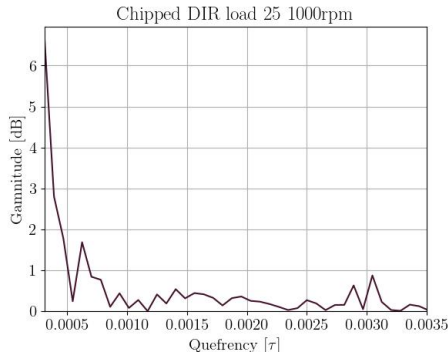
In order to better visualize the signals, we limited the quefrequency range to a bit less than 3 times the inverse of the GMF, as can be seen in [Figure 4.4](#). It seems that for the pulley-belt configuration, cepstrum shows little information, at least on the analised plots. At the direct driven configuration, the peak at 0.003, around the inverse of the frequency associated with the first GMF harmonic, seems to decrease with the severity of the defect.



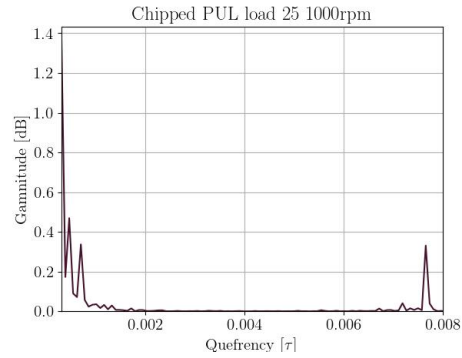
(a) Direct driven healthy pinion



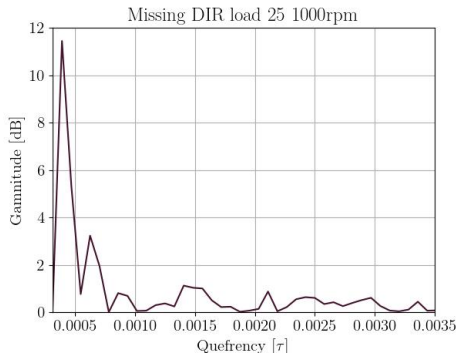
(b) Pulley-belt healthy pinion



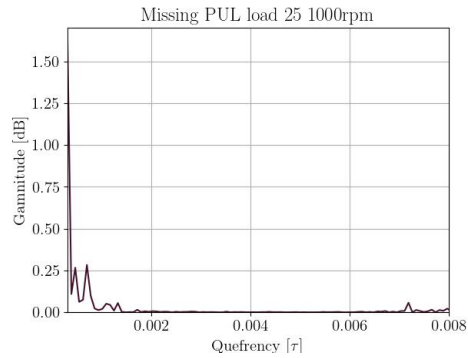
(c) Direct driven chipped pinion



(d) Pulley-belt chipped pinion



(e) Direct driven missing pinion



(f) Pulley-belt missing pinion

Figure 4.4 – Cepstrum acceleration signals at the pinion's accelerometer vertical direction

4.1.4 Order-domain: TSA

For the order-domain, first we apply an angular resampling to the signal, then we apply the TSA. After these steps, we apply the FFT to transform the TSA signal, then divide the frequency vector by the rotation of the pinion to see spectral content into the order-domain. At the order-domain, we extracted the FM0 and the first three harmonics associated with the GMF and f_p . The harmonics was extracted from the maximum value in the range of $\pm 0.001\%$ of the expected associated order.

The reason why f_g was not extracted is because it was completely filtered out from

the signals, as can be seen on [Figure 4.5](#). This may be due to a few reasons. The gearbox is quite simple, containing only one stage. Another reason is that the signal has 30 seconds, which encompasses lot of rotations. The TSA signal is obtained by averaging the signal over a number of rotations. This procedure may have filtered out any frequency that was not a direct multiple of the rotation of the pinion. The curves' shape, however, did not differ much from the FFT and Welch.

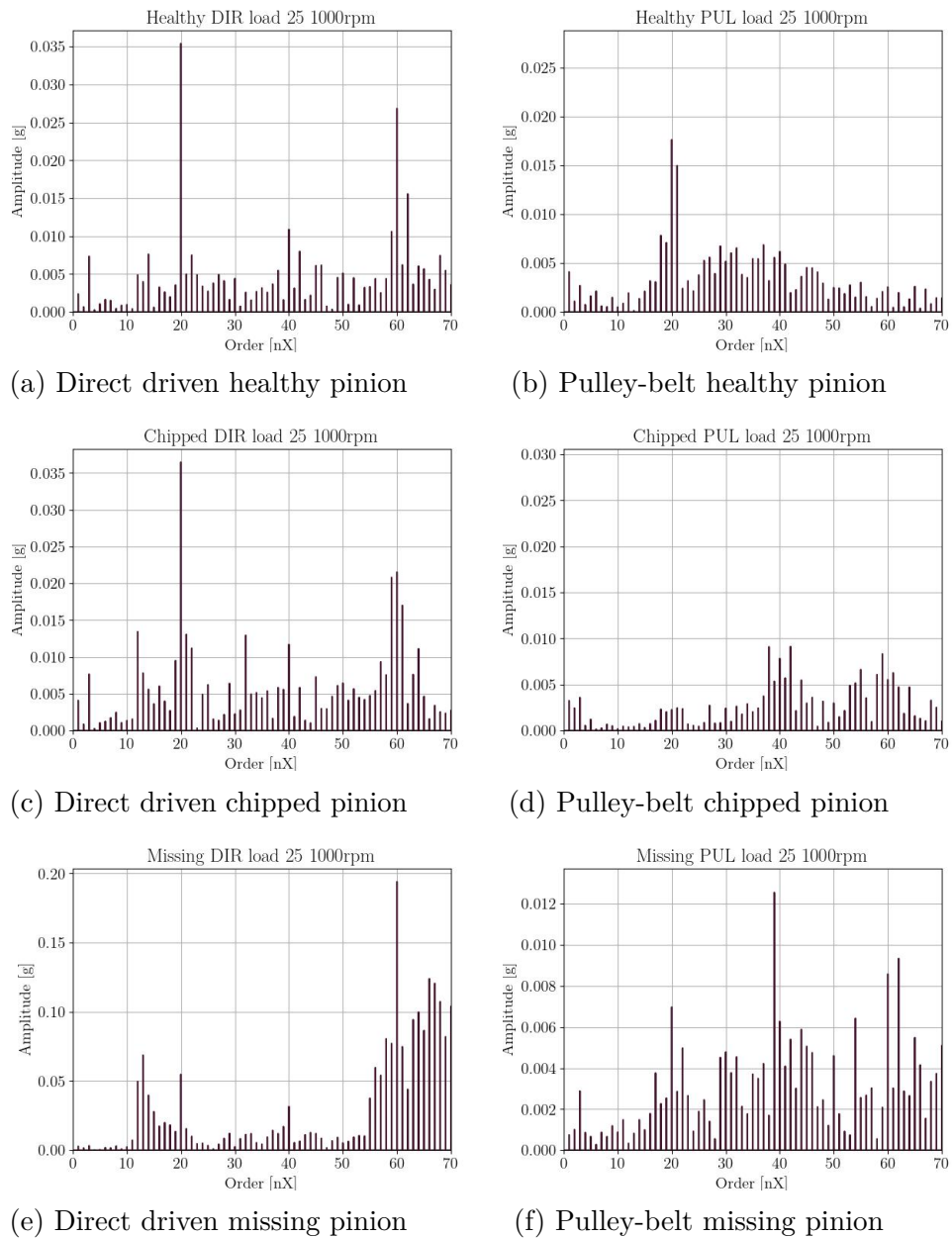


Figure 4.5 – Order acceleration signals at the pinion's accelerometer vertical direction

4.2 TRAIN TEST SPLIT

We decided on three different train-test divisions: (A) random division with one-third of the data for the test set, (B) training with data from the pulley-belt configuration and testing with data from the direct driven configuration, and (C) training with data from the pulley-belt and direct driven configurations, and testing with data from the pulley-belt configuration. Train-test splits (B) and (C) were intended to mimic real-world applications, where the models are usually trained in a particular configuration, and may be applied to different configurations.

4.2.1 Hyperparameter tuning and cross-validation

For the hyperparameter optimization, we carried out a cross-validation with three folds, which will be detailed in the next paragraph. With a specific set of hyperparameters, a model was trained using two of the folds and tested on the third one. This process was done interchangeably for all the folds. The mean of the AUC score from the tested fold represented the result of the validation set. After training the model several times with different combinations of hyperparameters, the best model was selected according to a given metric. The best model was retrained using the entire training set and subsequently tested on the test set.

For the cross-validation strategy, we chose a GroupKFold with three splits. GroupKFold is a variant of K-fold cross-validation that guarantees each group within the dataset is exclusively present either in the training or validation sets, but not both simultaneously (SCIKIT-LEARN, 2024a). In this case, the groups were chosen based on the combination of their experimental configuration. For example, all the replicates from a combination of load 0, 500 RPM, healthy pinion at the pulley-belt configuration were present either in the training or testing sets. The replicates with the same combination were never divided in the sets. This procedure was applied in the train and validation division and also in the train and test. This was done in order to avoid data leakage from training to testing.

Table 4.4 shows the default hyperparameters search range for each model. *Loguniform* and *uniform* stand for the probability distribution functions with the same name. The function `randint` returns a random integer. The values inside the functions stand for the search range. We chose the `RandomizedSearchCV` (SCIKIT-LEARN, 2024b) from the `scikit-learn` library. An analysis on the number of iterations is performed and can be

seen on [Appendix A](#). In about 25 iterations, there’s already a convergence on the models’ performance. Nevertheless, the number of iterations for the “default” models is set to 100.

Table 4.4 – Default hyperparameters search space

Model	Hyperparameter	Search Space
Logistic Regression	C	loguniform(10^{-6} , 10^5)
SVM	C	loguniform(10^{-6} , 10^5)
	gamma	loguniform(10^{-6} , 10^5)
Random Forest Classifier	n_estimators	randint(50,400)
	max_depth	randint(1,64)
	max_features	[None, “sqrt”, “log2”]
	min_samples_split	randint(2, 16)
	min_samples_leaf	randint(1, 10)
	ccp_alpha	uniform(10^{-6} , 1)
XGBoost	max_depth	randint(1, 64)
	learning_rate	uniform(0.01, 0.1)
	subsample	uniform(10^{-6} , 1)
	n_estimators	randint(50, 400)

4.3 MODEL PIPELINES

In total, four classifiers were implemented: Logistic Regression, SVM, Random Forest and XGBoost. We only considered two classes: healthy and defective, where the latter encompasses the chipped and missing tooth pinions’ condition. To deal with the data, some pipelines were created. The first step was to remove the columns with metadata associated. In the case of SVM and Logistic Regression, which greatly benefit from scaling, another pipeline step was added, which entailed the data standardization performed by the StandardScaler algorithm from the scikit-learn library ([SCIKIT-LEARN, 2024c](#)). The standardization is given by the formula:

$$z = \frac{x - \mu}{\sigma} \quad (4.1)$$

where z is the standardized value, x is the original value, μ is the mean of the feature in the train data-set, and σ is the standard deviation. This step is not necessary for the tree models, as they are not sensitive to the scale of the features as explained in “Decision trees and Random Forest Classifier” ([subsection 2.4.4](#)).

4.3.1 Metrics

After the models were trained, their performance was evaluated by a few parameters: validation AUC, represented as “Val. AUC”, test accuracy, represented as “ACC test”, test balanced accuracy, represented as “BA test”, the AUC from the test, represented as “AUC test”, the FPR from the test for a TPR of 90%, represented as “FPR (TPR \geq 90%)”, and the time it took to train the model with 11 processors/threads, represented as “Time [s]”. Then, ROC curves for the test group were plotted for all divisions.

4.3.2 SHAP analysis

After implementation, machine learning models often work like a black box. To better understand models, it is beneficial to evaluate the importance of each feature to the final result. There are several ways to do this. For example, interpreting the coefficients in a regression model. However, this can lead to misinterpretations. Furthermore, this method only takes into account the feature’s overall importance, but not how higher or lower values may impact the model’s outcome. From game theory, SHAP (SHapley Additive exPlanations) values are a way of overcoming these limitations. It can estimate how much an input from a feature impacted the model’s outcome, and then rank the feature’s importance. Nonetheless, it does not evaluate the quality of the prediction itself (SHAP, 2024).

5 RESULTS & DISCUSSIONS

This Chapter is dedicated to the results and discussions of this work. It is divided into two sections. The first section divided into subsections, each one corresponding to a different division of the dataset. For each division we present: variability analysis, ROC curve, SHAP analysis, and confusion matrix.

The second section is dedicated to the comparison between classifiers that were trained with all features except the ones generated from the FFT and Welch methods, classifiers that were trained with all the features but the ones from FFT and, classifiers that were trained with all the features but the ones from Welch. The models names were shortened to LR, SVM, RFC, and XGB for Logistic Regression, Support Vector Machine, Random Forest Classifier, and XGBoost, respectively.

5.1 DEFAULT MODEL ANALYSIS

This section analyses the variability of the default models, their ROC curves, SHAP analysis, and “confusion matrix” for each division. The “confusion matrices” are modified to show the model’s mistakes (FP and FN) and successes (TP and TN) but they are segregated by the pinion’s condition – unlike the usual confusion matrix. Nevertheless, for simplicity sake, it will be labeled “confusion matrix” in this work’s scope. The default models are trained with all features, and 100 iterations, as stated in machine learning methods ([chapter 4](#)).

There are basically two possible ways in which we can introduce randomness in our model’s training. The first one is the train-test split, and the second one is the hyperparameter search. The hyperparameter search is performed by the `RandomizedSearchCV` method from the `scikit-learn` library. This method randomly selects hyperparameters from a given range and performs a cross-validation to evaluate the model’s performance. The variability analysis aims to evaluate how the model’s performance varies with the train-test split and hyperparameter search. The models are trained and tested 15 times for each division. The mean and standard deviation of the AUC_{val} , ACC_{test} , BA_{test} , AUC_{test} , FPR (for a $TPR \geq 90\%$), and time in seconds are the evaluated metrics.

Following this analysis, we illustrate the models performance with their ROC curves. To enhance our comprehension of the models, summaries of the SHAP analysis were gener-

ated, plotting the top 20 features of each model. Since the XGBoost model demonstrated the best performance in most cases, considering also the training time, we selected it for the analysis. The SHAP library is a unified approach to explain the output of any machine learning model.

The SHAP summary plot comprises three elements: the y-axis, the x-axis, and the colorbar. The y-axis exhibits the features, while the x-axis displays the SHAP value. A positive SHAP value signifies that the feature contributes positively to the prediction, leading the model to predict 1 (indicating a defect in our case). Conversely, a negative SHAP value suggests that the feature negatively influences the prediction, guiding the model towards predicting 0, which indicates a healthy state. The colorbar represents the feature values, with blue indicating lower feature values, and red indicating higher feature values. The “thickness” of a feature’s plot along the axis of the SHAP value (x-axis) illustrates the amount of features that have that value.

We also analyse the XGBoost model mistakes and successes in a confusion matrix, to gain a better insight on the model. The confusion matrix plot for other models are illustrated at [Appendix C](#).

The default model hyperparameters can be checked at [Appendix B](#). It is noteworthy that the results presented at the ROC curve, SHAP analysis and confusion matrix were generated with a pseudo-random algorithm, with a “*random seed*” set to 5, so that they can be reproducible.

5.1.1 (A) Random division

This section presents the results for the (A) division.

5.1.1.1 Train-test split variability

[Table 5.1](#) shows the summary statistics for the train-test split variability in the (A) division. Overall, the standard deviation is less than 0.01, indicating that the data distribution is fair among groups. It is also important to remember that the hyperparameter search algorithm performs a randomized exploration. Since the model’s performance is not being highly affected by a randomized search, it may suggest that the optimization problem has many local minima which are close to the global minimum. This same conclusion can be extended to the next results of variability analysis.

The results show that the XGBoost model has the best performance in all metrics. Its time is around 202 seconds, about half of the Random Forest Classifier’s time. Furthermore, the XGBoost model presented a higher AUC_{test} than AUC_{val} . This may indicate that the quantity of information in the train and test datasets are different. It is interesting to note that SVM performed better than the Random Forest Classifier. The Logistic Regression model has the worst performance in all metrics. The FPR is the metric with highest variance among the models. This is expected since the FPR is a metric that is highly dependent on the threshold.

Table 5.1 – Summary statistics for the train test split variability in (A) division.

	LR		SVM		RFC		XGB	
	mean	std	mean	std	mean	std	mean	std
AUC_{val}	0.93	7.8e-03	0.94	6.7e-03	0.94	8.9e-03	0.96	6.6e-03
ACC_{test}	0.85	1.7e-02	0.89	1.5e-02	0.87	1.7e-02	0.91	1.6e-02
BA_{test}	0.85	1.9e-02	0.89	1.6e-02	0.86	1.8e-02	0.91	1.9e-02
AUC_{test}	0.93	1.1e-02	0.95	1.1e-02	0.95	9.7e-03	0.98	7.5e-03
FPR	0.22	4.5e-02	0.13	3.7e-02	0.18	4.1e-02	0.07	3.0e-02
Time	3.14	5.5e-02	76.18	2.8e+00	539.18	7.6e+00	202.54	7.0e+00

5.1.1.2 Randomized search variability

Table 5.2 shows the summary statistics for the hyperparameter search variability in the (A) division. It is interesting to note that the FPR was higher for the Random Forest Classifier (0.34 against 0.18) as well as for the XGBoost model (0.07 against 0.08) when comparing to Table 5.1. This may indicate that the hyperparameter search variability is finding local minima.

Table 5.2 – Summary statistics for the hyperparameter search variability in (A) division.

	LR		SVM		RFC		XGB	
	mean	std	mean	std	mean	std	mean	std
AUC_{val}	0.93	3.7e-05	0.94	3.1e-03	0.92	1.9e-02	0.97	5.4e-04
ACC_{test}	0.83	2.6e-03	0.85	1.9e-02	0.81	4.1e-02	0.91	4.3e-03
BA_{test}	0.83	2.2e-03	0.86	1.7e-02	0.81	4.5e-02	0.91	4.3e-03
AUC_{test}	0.92	5.1e-04	0.94	1.1e-02	0.89	4.4e-02	0.98	1.2e-03
FPR	0.27	8.0e-04	0.25	4.9e-02	0.34	1.3e-01	0.08	6.9e-03
Time	1.50	8.7e-02	34.55	1.2e+00	315.58	6.0e+01	88.24	7.7e+00

5.1.1.3 ROC curve

Figure 5.1 illustrates the performance of the models for the (A) division. XGBoost is the best performing model with a TPR of 1, for an FPR of 0.3 to 0.4. The rest of the models present the plateau at the FPR about at 0.6. SVM was the second best performing model. Random Forest Classifier has a very similar ROC curve to the Logistic Regression's, which is a simpler model.

Since the performance SVM was greater than the Logistic Regression's, it implies that there are some non-linearities at the analyzed problem. The SVM algorithm used in this work considers a Radial Basis Function (RBF) kernel, which computes a similarity score between data points based on their distance in the input space.

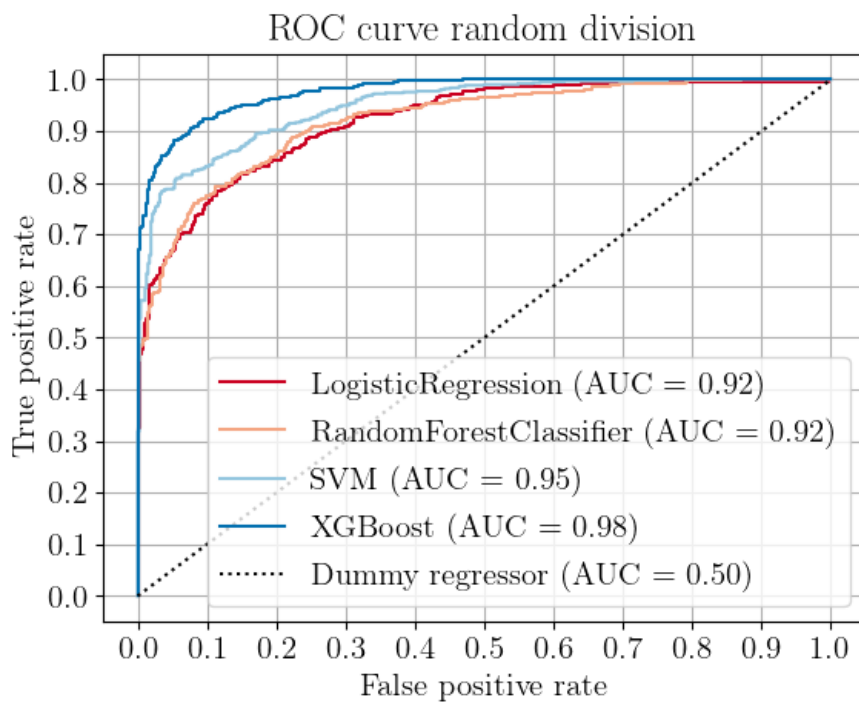


Figure 5.1 – ROC curve for the (A) division.

5.1.1.4 SHAP Analysis

Figure 5.2 shows the SHAP analysis for the XGBoost model and (A) division. The first seven most important features are from the cepstrum, order and time-domain. The most important being the first rahmonic of the gear rotation and the fourth feature, the third rahmonic. More specifically, high values of the feature predict a healthy signal, while

low values predict defects. This may be because the defective pinion transmits less load to the gear, which then vibrates less.

Although FM0, which comes from the TSA, appeared at the second place, other features derived from this transformation appeared on 11th and 19th place only. One should observe that it needs tachometer information and requires far more computational resources than the other transformation or features presented. Therefore, its cost-performance ratio raises debate about an extended real world application.

Kurtosis, shape factor and crest factor plus ranked third, fifth, and seventh respectively. All three provide insights into the signal's shape, peaks, or impulsiveness. The SHAP analysis suggests that higher values of kurtosis predict defective signals. This finding aligns with existing literature, as defects such as a missing gear tooth often manifest as impulsive behavior in the signal. The same trend applies to shape factor and crest factor plus.

It stands out that the first six features have the most spread values. Also, features from the frequency-domain appeared from the eighth importance on. This may be because the information from the frequency-domain is embedded in the other features.

5.1.1.5 Confusion matrix

Figure 5.3 shows a confusion matrix for the XGBoost model. There were only four mistakes for the missing pinion condition, sixty one for the chipped one and thirty three for the healthy condition. This suggests that the model is better at predicting the missing pinion condition than the chipped one. This is expected since the missing pinion condition is a more severe defect than the chipped one, which would imply in higher differentiation to the healthy condition.

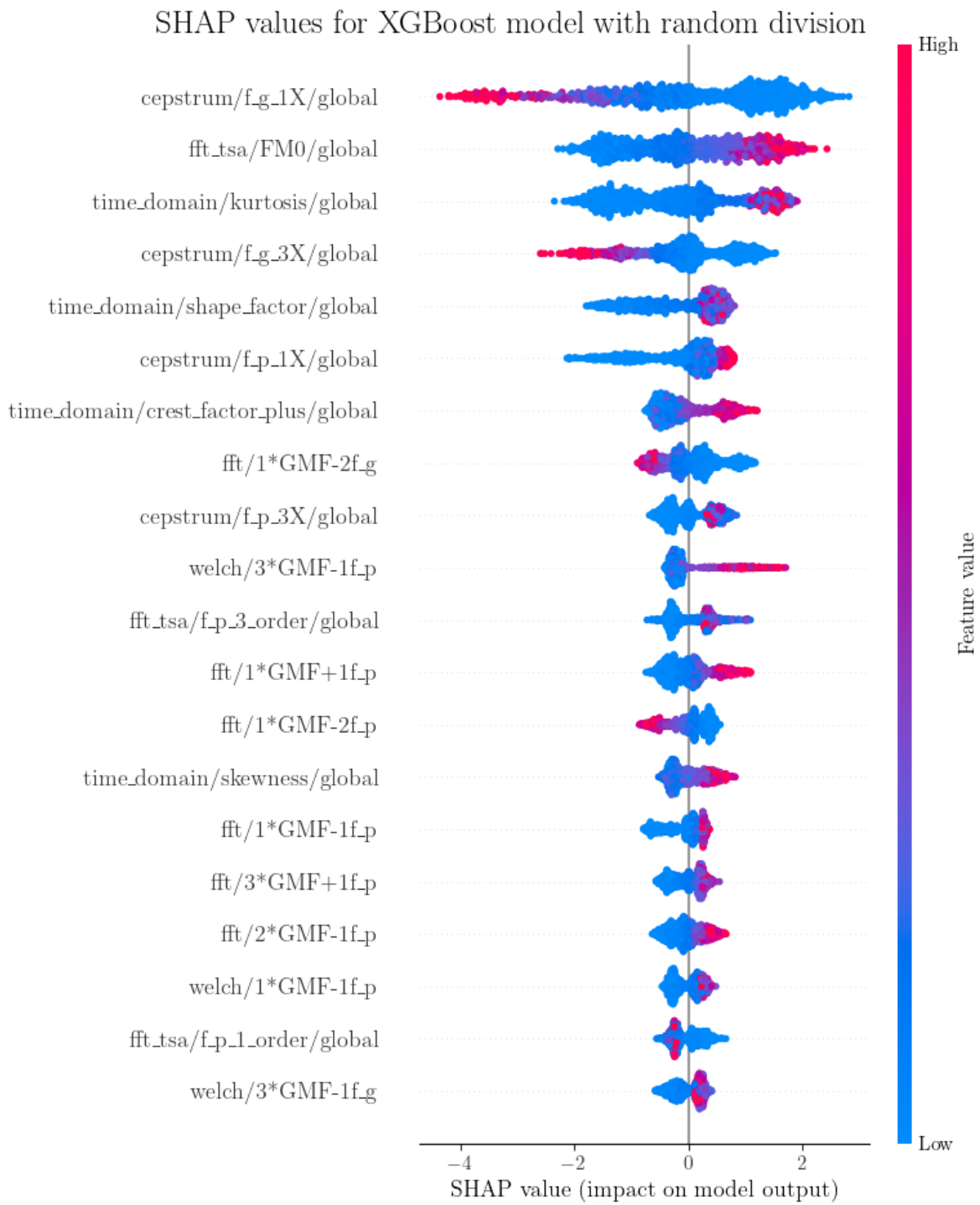


Figure 5.2 – SHAP analysis for the (A) division.

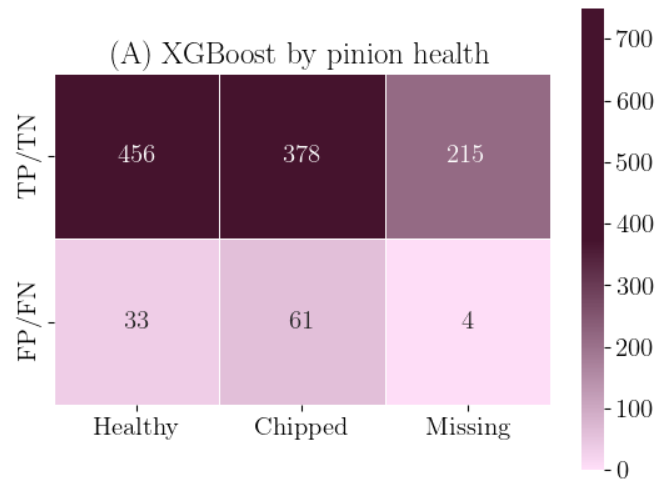


Figure 5.3 – Confusion matrix for (A) division and XGBoost.

5.1.2 (B) Pulley-belt division

This section presents the results for the (B) division.

5.1.2.1 Randomized search variability

Table 5.3 shows the summary statistics for the hyperparameter search variability in the (B) division. The standard deviation is less than 0.05 for all metrics. The AUC_{val} has similar values, when comparing to the (A) division. The tests scores performed worse than the (A) division, presenting FPR values higher than 0.70. It was expected, since the model was tested in a condition in which it was not trained.

Table 5.3 – Summary statistics for the hyperparameter search variability in (B) division.

	LR		SVM		RFC		XGB	
	mean	std	mean	std	mean	std	mean	std
AUC_{val}	0.90	3.1e-05	0.92	8.9e-03	0.89	2.0e-02	0.93	4.8e-04
ACC_{test}	0.65	8.4e-04	0.64	5.0e-03	0.69	1.8e-02	0.68	3.9e-03
BA_{test}	0.61	7.3e-04	0.59	7.9e-03	0.64	2.1e-02	0.62	5.5e-03
AUC_{test}	0.66	3.7e-04	0.61	2.0e-02	0.78	2.0e-02	0.81	5.2e-03
FPR	0.73	2.1e-03	0.74	5.6e-03	0.60	4.8e-02	0.61	1.8e-02
Time	0.96	3.4e-02	10.22	2.9e-01	162.96	2.9e+01	62.08	3.5e+00

5.1.2.2 ROC curve

Figure 5.4 illustrates the performance of the models for the (B) division. XGBoost and Random Forest Classifier had similar performance, with XGBoost having the highest AUC score and lower training time. Logistic Regression overall performed better than the SVM model. For a TPR lower than 0.5, SVM performed similarly to a dummy regressor, which guesses randomly the input's label. Logistic Regression has a linear kernel, meanwhile SVM has a non-linear one. Since the models scored better at training, it seems that SVM had more overfitting than Logistic Regression at this division.

5.1.2.3 SHAP Analysis

Figure 5.5 shows the SHAP analysis for the XGBoost model and (B) division. The most important feature is the FM0, followed by the shape factor. These two show the most spread values. High values of the first rahmonic of the GMF, the third most important feature, indicate most certainly a defective signal. The same can be said for the eighth

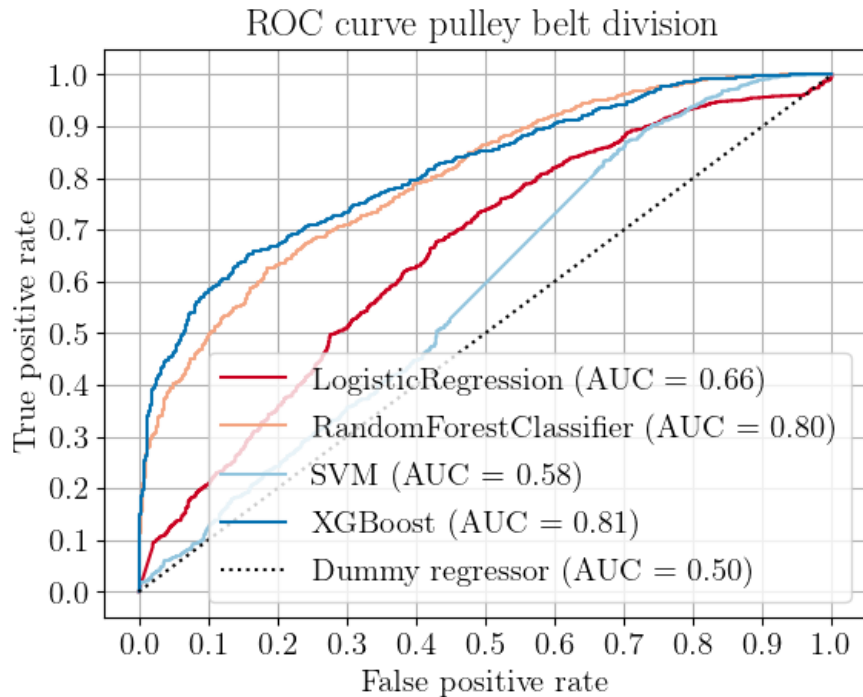


Figure 5.4 – ROC curve for the (B) division.

most important feature, the second harmonic of the GMF. As in (A) division, kurtosis is one of the most important features and high values of it predict defective signals.

In this division, the third and first harmonics of the pinion rotation frequency appear as the fourth and fifth most important features, exceeding in importance order compared to the features from the frequency-domain. This may be because the TSA not only reduces noise in the signal, but also included the angular resampling, which reduces the influence of the pulley-belt system slippage.

There are nine features of harmonics from the frequency-domain. Six of them are related to GMF sidebands. Sidebands are associated with modulation in the signal. The pulley-belt system as overall more variation in rpm, which may smear the rpm peak generating sidebands. There are “repeated” features that come from both Welch and FFT methods, such as the sidebands of minus one gear frequency in the third GMF ($3*GMF-1f_p$). Although they differ, since Welch’s method is a statistical one, and FFT is a deterministic one, they both provide similar information. In that sense, they may be concurring in the model’s decision.

Spectral flatness, the tenth most important feature, serves as a reliable predictor of defects when its values are low. However, it consistently predicts defects, regardless of the signal characteristics. Therefore, when spectral flatness values are high, other features

should be considered to evaluate the component's health. On the other hand, low values of the first rahmonic of the pinion rotation predict healthy component. High values, however, are not as conclusive.

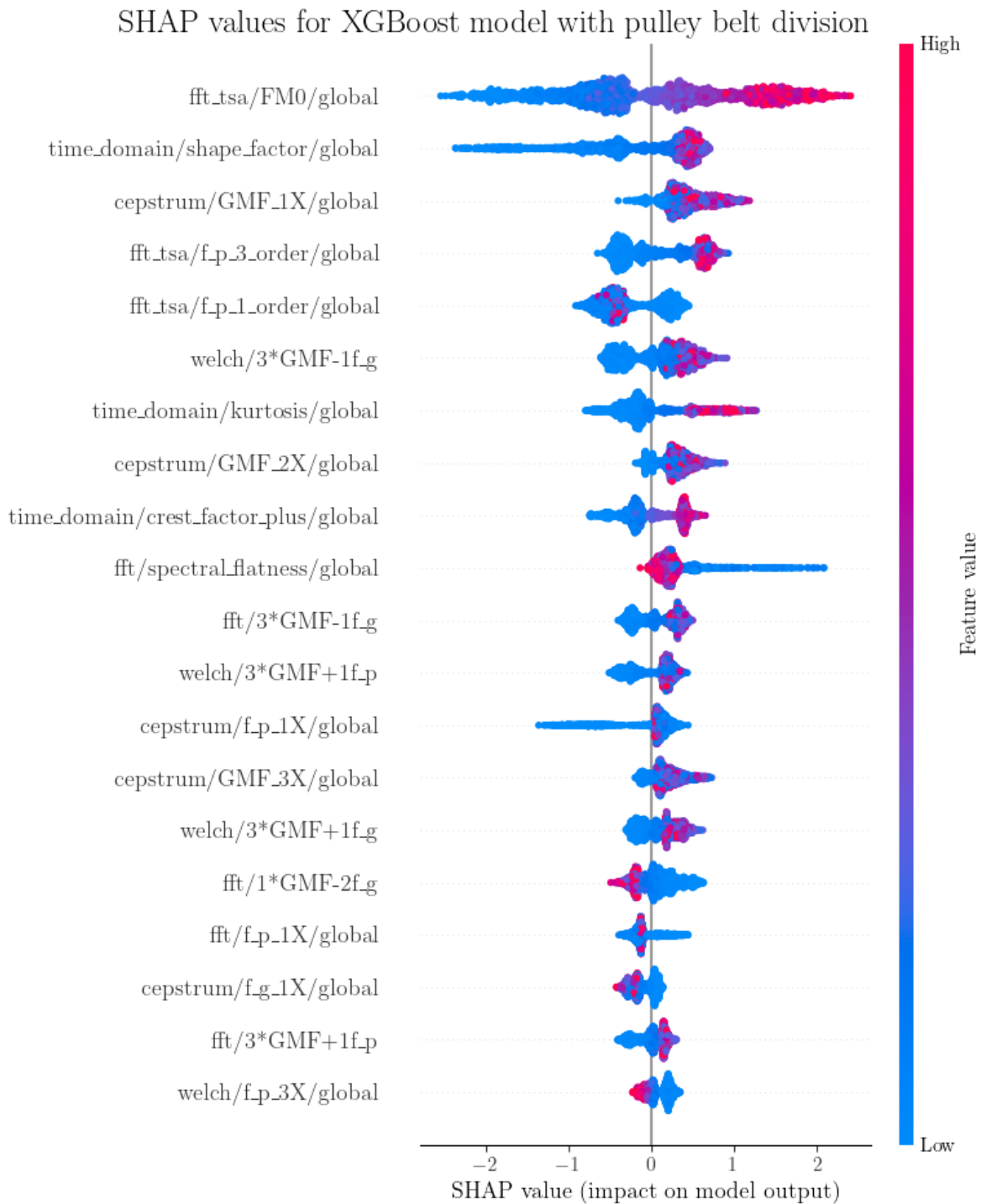


Figure 5.5 – SHAP analysis for the (B) division.

5.1.2.4 Confusion matrix

Figure 5.6 shows a confusion matrix for the XGBoost model. The majority of mistakes were of false positives. There were only four mistakes for the missing pinion condition, sixty three for the chipped one and 591 for the healthy condition. One should observe that, although the mistakes for the healthy condition were more than one order higher when compared to the random division, the mistakes for the defective conditions were about the same in number.

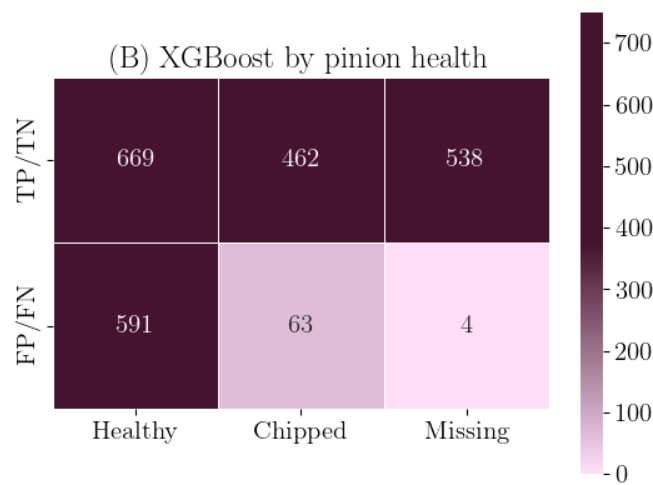


Figure 5.6 – Confusion matrix for (B) division and XGBoost.

5.1.3 (C) Direct driven division

This section presents the results for the (C) division.

5.1.3.1 Randomized search variability

Table 5.4 shows the summary statistics for the hyperparameter search variability in the (C) division. The metrics also presented low standard deviation, with values less than 0.05.

The (C) division had by far the best score at validation, reaching up to a hundred percent with XGBoost. On the other hand, it generally had the worst results in the test. There are a few differences between the configurations. One of them is that the pulley-belt system has a higher slippage. This means that the rotational variation is higher in the pulley-belt system. The pulley-belt system also has imprinting of other rotating component

– such as the belt, the main shaft and its bearings, etc. The direct driven division, on the other hand, consists only of the motor, the coupling, and the gearbox. This implies that the direct driven signals are “cleaner” than the pulley-belt signals.

Another difference between the divisions themselves is that there were more RPMs analyzed in the (C) division than in the (B) division (13 compared to 7). This means that there were more signals acquired from (C) than from (B). Even though the (C) division had more data to fit to, it yielded worse results when tested in another configuration. In both (C) and (B) divisions, there was some overfitting, as the train and test scores differed considerably from each other.

Table 5.4 – Summary statistics for the hyperparameter search variability in (C) division.

	LR		SVM		RFC		XGB	
	mean	std	mean	std	mean	std	mean	std
AUC_{val}	0.99	7.5e-06	0.98	2.4e-03	0.96	1.2e-02	1.00	1.8e-04
ACC_{test}	0.61	6.3e-04	0.64	1.6e-02	0.70	1.0e-02	0.66	9.6e-03
BA_{test}	0.56	6.7e-04	0.59	1.9e-02	0.67	1.5e-02	0.63	1.2e-02
AUC_{test}	0.68	8.4e-04	0.70	2.8e-02	0.78	4.9e-03	0.77	7.1e-03
FPR	0.74	3.4e-03	0.66	2.5e-02	0.64	7.7e-02	0.71	2.3e-02
Time	1.42	5.5e-02	31.01	7.1e-01	259.41	4.9e+01	60.01	5.7e+00

5.1.3.2 ROC curve

Figure 5.7 illustrates the performance of the models for the (C) division. Random Forest Classifier and XGBoost had similar performances, with Random Forest Classifier outperforming XGBoost after FPR of 0.5. SVM had a better performance at (C) division than at (B) division, but still equal or worse than the Logistic Regression model.

5.1.3.3 SHAP Analysis

Figure 5.8 shows the SHAP analysis for the XGBoost model and (C) division. The first five features present the most spread in feature values. The first two are the first and third rahmonic of the gear rotation. Again, this may be because the defective pinion transmits less load to the gear. It is essential to thoroughly analyze the signals affecting this behavior. The fifth feature, the first rahmonic of the pinion rotation, and the tenth, the third rahmonic of the pinion rotation, and eighteenth, cepstral peak, are also features from the quefrequency-domain. It seems that no division presented a second rahmonic as an important feature.

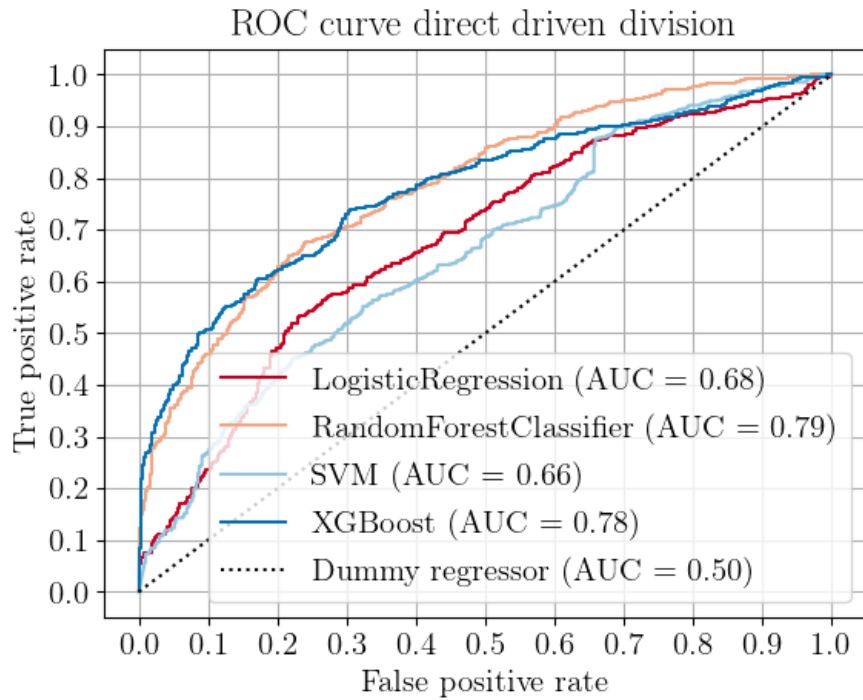


Figure 5.7 – ROC curve for the (C) division.

Kurtosis and FM0 were at the most important features, just as in the other divisions. The shape factor also appeared as important. Eleven features are harmonic amplitudes from the frequency-domain, mostly associated with sidebands.

5.1.3.4 Confusion matrix

Figure 5.9 shows a confusion matrix for the XGBoost model. The (C) division mistook less at the healthy condition than the (B) division, but more at the defective conditions compared to all divisions. The (C) division mistook thirteen missing tooth pinion condition, more than three times than the other divisions.

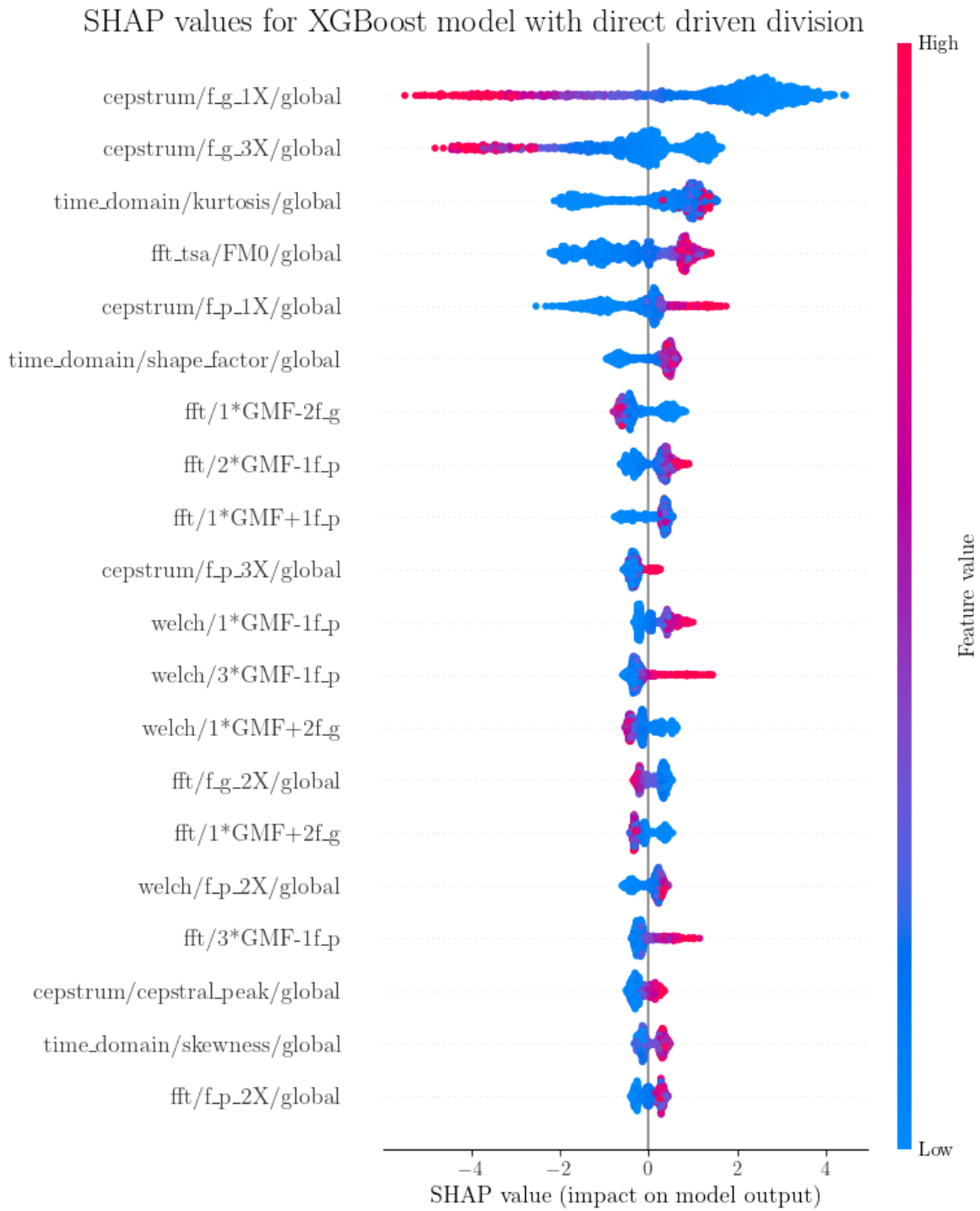


Figure 5.8 – SHAP analysis for the (B) division.

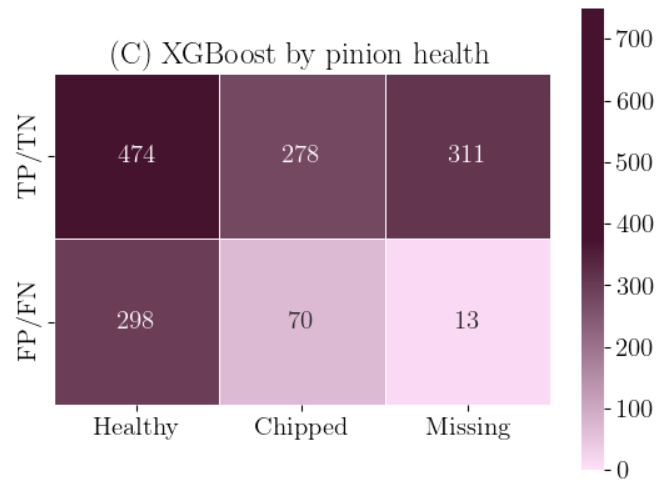


Figure 5.9 – Confusion matrix (C) division and XGBoost.

5.2 FFT VS WELCH VS NEITHER

After noticing that there were features with similar information from the frequency-domain, we decided to test models where the features came solely from either Welch’s method or the FFT method. We also tested models that excluded frequency-domain features altogether. Since the results were repetitive, we decided to display [Table 5.5](#). Tables for the other conditions can be found at [Appendix D](#).

Table 5.5 – Comparison of fft vs welch method for Random Forest Classifier and XGBoost (A) division.

method	RFC			XGB		
	welch	fft	none	welch	fft	none
AUC_{val}	0.93	0.93	0.94	0.96	0.97	0.96
ACC_{test}	0.82	0.83	0.82	0.90	0.90	0.88
BA_{test}	0.83	0.83	0.82	0.91	0.91	0.88
AUC_{test}	0.92	0.92	0.92	0.98	0.97	0.96
FPR	0.24	0.24	0.27	0.09	0.09	0.13
Time	343.32	344.67	147.99	115.73	115.49	43.89

Overall, the three conditions yielded similar results, with the models excluding frequency-domain features performing better at times. This contradicts the literature, which emphasizes the importance of frequency-domain features in fault diagnosis. This may be due to the analysed defects, which are more severe types of defects, and the signals are more impulsive than modulated. Impulsive signals affect more broadband frequencies, than tonal or specific ones. These defects are better captured by time-domain features.

6 CONCLUSION

Among the existing machine elements, gear defects are critical not only for their malfunctioning, but also for the possibility of damaging other components. Nevertheless, there are few studies that focus on comparing binary classifiers for gear defect detection. This work focuses on defect detection and not diagnosis (locating and identifying the type of defect) or prognosis. This step can be seen as one of binary classification: “defective or healthy signal?”. To answer this question, the literature suggests different features. We implemented statistical features from the time domain, amplitudes from the frequency and order-domain, gamnitudes from the cepstrum analysis, and a few others, such as the FM0.

Our configuration setups comprised a gearbox connected directly to the motor and another with a pulley-belt system. There were three investigated train-test divisions: (A) random division, (B) training with pulley-belt data and testing with the direct driven, and (C) training with the direct driven data and testing with the pulley-belt one. The latter two mimicked real-world applications, where a model is trained in a machine and we try to reproduce it on other machines.

We compared four different classifiers, Logistic Regression, SVM, Random Forest and XGBoost. To train and optimize hyperparameters, we used the AUC score as the main metric. To evaluate the models thoroughly, we used the validation and test AUC, the FPR for a TPR of 90% at the test, test accuracy and balanced accuracy, and the time spent training the models.

In all divisions, the training time, from faster to slower, was as follows: Logistic Regression, SVM, XGBoost, and Random Forest Classifier. XGBoost had, overall, the best results. The random division performed the best with all classifiers. The pulley-belt train division displayed better results than the direct driven division, even though it had less data for training. In contrast, the direct-driven configuration includes only the motor, the coupling, and the gearbox, which means the direct-driven signals are “cleaner” compared to the pulley-belt signals. Another difference is that the pulley-belt system experiences higher slippage, resulting in greater rotational variation. A model trained with data with more variability is more likely to perform better with data with less variability. Models require domain generalization techniques to be applied at a larger scale.

The random division had a satisfying performance with all models. For the XGBoost model, a 90% TPR was achieved at 7% FPR. Although SVM had the second best

performance at the random division, it is not as applicable in real-world context, since the data was standardized in the pipeline. Random Forest Classifier had a similar performance to the Logistic Regression model, but it was the slowest to train. It seems that more complex models, such as neural networks, could be an overkill for these requirements, especially when thinking of a real-world application, which demand low latency and computational costs. This may be true if data is diverse in training conditions, just as the random division. Moreover, shallow learning models are easier to interpret and explain than deep learning ones.

Ranked by SHAP values, the twentieth most important features from XGBoost were analyzed. Kurtosis and FM0 were deemed as important in all divisions. The SHAP analysis suggests that features that are a measure of the signal's shape are, in general, the best to classify the signal as defective. This may be because of the nature of the analyzed defects: chipped and missing tooth, which give rise to periodic impact. A worn gear is unlikely to produce such an impulsive response.

Looking at the model's mistakes and successes based on the pinion's condition, we observed that the (C) division failed more often to classify the defective pinion. Although (B) division erred more often in classifying the healthy pinion, it had a similar performance to the (A) division. It seems that (B) is better at generalizing the domain than (C).

Analysis excluding features from the frequency-domain showed similar performance to the full feature set, contradicting literature on their importance. This may be because the other features already entailed the information. Another reason could be due to the dynamic response style of the defective signals.

6.1 FUTURE WORK

This work opens up a few possibilities for future research. We could investigate the following topics with data already available:

- Investigate the influence of direction (vertical, horizontal or axial) on feature values and classification performance for the studied defects;
- Do a multiclass classification and investigate which features describe better each defect individually;

- Check model’s performance without TSA features, to see if they are really necessary, since they are not applicable extensively in many “real world” context;
- Apply domain generalization techniques to the models and investigate their performance to the different configurations;
- There were important features which gave similar information (e.g. kurtosis, shape factor or crest factor plus). It would be interesting to train a model with few features with little overlapping information. To evaluate overlapping features, we could evaluate the correlation between them and the importance of each feature in the model. This could be done with a correlation matrix and a SHAP analysis, respectively.

Moreover, other investigations with new data could be:

- Investigate a geartrain with more stages;
- Investigate other gear defects, such as wear;
- Cross defects that are commonly occurring with gears, such as bearing defects, and evaluate the model’s hability to classify the signal as defective or healthy and differentiate the defects.

BIBLIOGRAPHY

- AGMA. Appearance of Gear Teeth: Terminology of Wear and Failure. **American Gear Manufacturers Association Technical Committee**, v. 40, 1995.
- ALBAN, L. E. **Systematic Analysis of Gear Failures**. American Society for Metals, 1985. ISBN 0-87170-200-2.
- B&K. **Product Data – Triaxial DeltaTron Accelerometers with TEDS – Types 4525-B and 4525-B-001**. 2023. Last visited on 10/11/2023. Available from: <https://www.bksv.com/media/doc/bp2203.pdf>.
- CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: p. 785–794. Last visited on 14/04/2024. ISBN 9781450342322. DOI: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785). Available from: <https://doi.org/10.48550/arXiv.1603.02754>.
- COLLINS, J. A.; BUSBY, H. R.; STAAB, G. H. **Mechanical Design of Machine Elements and Machines: A Failure Prevention Perspective**. Second Edition: John Wiley & Sons, 2009. P. 1–912. ISBN 0470413034.
- DOMINGUES, R. K. **Avaliação de metodologia de detecção de falhas em mancais de rolamento utilizando análise de ordem e média síncrona no tempo**. 2023. Universidade Federal de Santa Catarina. Last visited on 28/04/2024. Available from: <https://repositorio.ufsc.br/handle/123456789/251520>.
- DYNAMOX. **4 types of Industrial Maintenance**. 2021. Last visited on 31/01/2023. Available from: <https://dynamox.net/en/blog/4-types-of-industrial-maintenance>.
- ELFORJANI, M. Diagnosis and prognosis of real world wind turbine gears. **Renewable Energy**, v. 147, p. 1676–1693, Mar. 2020. ISSN 0960-1481. DOI: [10.1016/j.renene.2019.09.109](https://www.sciencedirect.com/science/article/pii/S0960148119314478). Available from: <https://www.sciencedirect.com/science/article/pii/S0960148119314478>.
- FARRAR, C. R.; WORDEN, K. An introduction to structural health monitoring. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 365, n. 1851, p. 303 – 315–303 –315, 2007. Publisher: Royal Society. DOI: [10.1098/rsta.2006.1928](https://doi.org/10.1098/rsta.2006.1928). Available from:

<<https://www.scopus.com/inward/record.uri?eid=2-s2.0-33846995979&doi=10.1098%2frsta.2006.1928&partnerID=40&md5=045ec6404f278a98240d44b060e04705>>.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006. ROC Analysis in Pattern Recognition. ISSN 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2005.10.010>. Available from:

<<https://www.sciencedirect.com/science/article/pii/S016786550500303X>>.

FYFE, K.; MUNCK, E. ANALYSIS OF COMPUTED ORDER TRACKING.

Mechanical Systems and Signal Processing, v. 11, p. 187–205, 2 Mar. 1997. ISSN 08883270. DOI: [10.1006/mssp.1996.0056](https://doi.org/10.1006/mssp.1996.0056).

GERÓN, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2nd Edition: O'Reilly, 2019. ISBN 1492032646.

GRANDINI, M.; BAGLI, E.; VISANI, G. Metrics for Multi-Class Classification: an Overview. **ArXiv**, abs/2008.05756, 2020. Available from:

<<https://api.semanticscholar.org/CorpusID:221112671>>.

HAN, B.; YANG, X., et al. Comparisons of different deep learning-based methods on fault diagnosis for geared systems. **International Journal of Distributed Sensor Networks**, v. 15, n. 11, p. 1550147719888169, Nov. 2019. Publisher: SAGE Publications. ISSN 1550-1329. DOI: [10.1177/1550147719888169](https://doi.org/10.1177/1550147719888169). Available from:

<<https://doi.org/10.1177/1550147719888169>>. Visited on: 27 Feb. 2024.

HAN, T.; JIANG, D., et al. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. **Transactions of the Institute of Measurement and Control**, v. 40, n. 8, p. 2681–2693, May 2018. Publisher: SAGE Publications Ltd STM. ISSN 0142-3312. DOI:

[10.1177/0142331217708242](https://doi.org/10.1177/0142331217708242). Available from:

<<https://doi.org/10.1177/0142331217708242>>. Visited on: 27 Feb. 2024.

IBM. **Decision Trees**. 2024. Available from:

<<https://www.ibm.com/topics/decision-trees>>.

JAMES, G. et al. **An Introduction to Statistical Learning: With Applications in Python**. 1st Edition: Springer, 2023. ISBN 3031387465. Available from:

<<https://www.statlearning.com/>>.

KUMAR, A. et al. Latest developments in gear defect diagnosis and prognosis: A review. **Measurement: Journal of the International Measurement Confederation**, Elsevier B.V., v. 158, 2020. Cited by: 84. ISSN 02632241. DOI:

10.1016/j.measurement.2020.107735. Available from:

<<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081993599&doi=10.1016%2fj.measurement.2020.107735&partnerID=40&md5=355e097d05c1de619088ffc1fb7479da>>.

KUNDU, P.; DARPE, A. K.; KULKARNI, M. S. A review on diagnostic and prognostic approaches for gears. **Structural Health Monitoring**, SAGE Publications Ltd, 2020. cited By 17. ISSN 14759217. DOI: 10.1177/1475921720972926. Available from: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097980097&doi=10.1177%2f1475921720972926&partnerID=40&md5=1981274e07aebc92c4949b50fee64b3e>>.

MAGTORK. **Magtork Model MTL 10 Standard Hysteresis Clutch**. 2023. Last visited on 10/11/2023. Available from:

<<https://www.magtork.com/model-mtl-10-standard-hysteresis-clutch/>>.

MCFADDEN, P. D.; SMITH, J. D. Statistical Parameters for Vibration Analysis in Condition Monitoring. **Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science**, SAGE Publications, v. 199, n. 4, p. 287–292, 1985. DOI: 10.1243/PIME_PROC_1985_199_110_02.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engineers**. 6th Edition: John Wiley and Sons, 2003.

MOTT, R. L.; VAVREK, E. M.; WANG, J. **Machine Elements in Mechanical Design**. Ed. by Pearson. 6th Edition, 2017. ISBN 978-0-13-444118-4.

NORTON, R. L. **Machine Design: An Integrated Approach**. 4th Edition: Pearson, 2010. ISBN 0136123708.

PCB. **PCB 352C33**. 2002. Last visited on 27/02/2023. Available from:

<<https://www.pcb.com/products?m=352c33>>.

RANDALL, R. B. Cepstrum Analysis and Gearbox Fault Diagnosis. **Brüel and Kjær Application Note No. 13-150**, 1973. Last visited on 31/01/2023. Available from:

<<https://www.bksv.com/doc/233-80.pdf>>.

RANDALL, R. B. A history of cepstrum analysis and its application to mechanical problems. **Mechanical Systems and Signal Processing**, v. 97, p. 3–19, Dec. 2017. ISSN 08883270. DOI: [10.1016/j.ymssp.2016.12.026](https://doi.org/10.1016/j.ymssp.2016.12.026).

RANDALL, R. B. **Vibration-based Condition Monitoring**. Wiley, Jan. 2011. ISBN 9780470747858. DOI: [10.1002/9780470977668](https://doi.org/10.1002/9780470977668).

RYTTER, A. *Vibrational Based Inspection of Civil Engineering Structures*, 1993. Place: Aalborg Publisher: Dept. of Building Technology and Structural Engineering, Aalborg University.

S. DUBNOV. Generalization of spectral flatness measure for non-Gaussian linear processes. **IEEE Signal Processing Letters**, v. 11, n. 8, p. 698–701, Aug. 2004. ISSN 1558-2361. DOI: [10.1109/LSP.2004.831663](https://doi.org/10.1109/LSP.2004.831663).

SAIT, A. S.; SHARAF-ELDEEN, Y. I. A Review of Gearbox Condition Monitoring Based on vibration Analysis Techniques Diagnostics and Prognostics. In: p. 307–324. DOI: [10.1007/978-1-4419-9428-8_25](https://doi.org/10.1007/978-1-4419-9428-8_25).

SCIKIT-LEARN. **Cross-validation: evaluating estimator performancen**. 2024. Last visited on 14/02/2024. Available from: https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation.

SCIKIT-LEARN. **RandomizedSearchCV**. 2024. Last visited on 08/03/2024. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html.

SCIKIT-LEARN. **StandardScaler**. 2024. Last visited on 08/03/2024. Available from: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>.

SCIKIT-LEARN. **SVM - Mathematical formulation**. 2024. Last visited on 02/02/2024. Available from: <https://scikit-learn.org/stable/modules/svm.html#svm-mathematical-formulation>.

SHAP. SHAP library references, 2024. Last visited on 08/03/2024. Available from: <https://shap.readthedocs.io/en/latest/overviews.html>.

SHIN, K.; HAMMOND, J. **Fundamentals of Signal Processing for Sound and Vibration Engineers**. 1st Edition: Wiley, 2008. ISBN 0470511885.

SINGH, V. et al. Artificial intelligence application in fault diagnostics of rotating industrial machines: a state-of-the-art review. **Journal of Intelligent Manufacturing**, 2021. Publisher: Springer. DOI: [10.1007/s10845-021-01861-5](https://doi.org/10.1007/s10845-021-01861-5). Available from:

<<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118973616&doi=10.1007%2fs10845-021-01861-5&partnerID=40&md5=e9b2a94f8145a25161f4b5b28a016bcd>>.

SPECTRAQUEST. **Machinery Fault Simulator**. 2023. Last visited on 16/06/2023.

Available from: <<https://spectraquest.com/simulators/details/mfs/>>.

STEWART, R. M. **Some useful analysis techniques for gearbox diagnostics**.

Technical Report MHM/R/10/77, Machine Health Monitoring Group, Institute of Sound and Vibration Research, University of Southampton, July, 1977.

SYLVESTER, J.; PEARCE, J. **Book of Gold: Practical Condition Monitoring Case Studies**. JPS Reliability. Available from:

<<https://rms-reliability.com/product/book-of-gold/>>. Visited on: 19 Feb. 2024.

TAYLOR, J. I. **The Vibration Analysis Handbook**. Kingwood, Texas: Vibration Consultants, 2003.

WELCH, P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. **IEEE Transactions on Audio and Electroacoustics**, v. 15, n. 2, p. 70–73, 1967. DOI:

[10.1109/TAU.1967.1161901](https://doi.org/10.1109/TAU.1967.1161901).

WENIG, B. **Decision Trees Presentation**. 2024. Last visited on 31/01/2024. Available from: <<https://brookewenig.com/DecisionTrees.html#/1>>.

WIKIPEDIA. **Logistic Regression**. 2024. Last visited on 02/02/2024. Available from:

<https://en.wikipedia.org/wiki/Logistic_regression>.

WIKIPEDIA. **Receiver operating characteristic**. Jan. 2024. Last visited on 30/01/2024. Available from:

<https://en.wikipedia.org/wiki/Receiver_operating_characteristic>.

APPENDIX A – NUMBER OF ITERATIONS

Table A.1 – Influence of number of iterations on Logistic Regression and SVM for the (A) division.

n_{iter}	LR				SVM			
	10	25	50	100	10	25	50	100
AUC_{val}	0.93	0.93	0.93	0.93	0.92	0.93	0.94	0.95
ACC_{test}	0.82	0.82	0.82	0.82	0.77	0.80	0.84	0.86
BA_{test}	0.83	0.83	0.83	0.83	0.79	0.82	0.84	0.87
AUC_{test}	0.92	0.92	0.92	0.92	0.90	0.91	0.94	0.95
FPR	0.30	0.28	0.27	0.27	0.36	0.36	0.24	0.19
Time	2.16	0.53	0.92	1.55	4.75	8.72	17.74	35.52

Table A.2 – Influence of number of iterations on Random Forest Classifier and XGBoost for the (A) division.

n_{iter}	RFC				XGB			
	10	25	50	100	10	25	50	100
AUC_{val}	0.83	0.93	0.93	0.93	0.96	0.96	0.96	0.96
ACC_{test}	0.57	0.83	0.84	0.83	0.90	0.90	0.90	0.91
BA_{test}	0.50	0.83	0.84	0.83	0.91	0.91	0.90	0.91
AUC_{test}	0.77	0.92	0.92	0.92	0.97	0.97	0.97	0.98
FPR	1.00	0.23	0.24	0.24	0.09	0.09	0.09	0.08
Time	21.74	69.53	113.10	228.66	12.28	27.08	49.30	95.58

Table A.3 – Influence of number of iterations on Logistic Regression and SVM for the (B) division.

n_{iter}	LR				SVM			
	10	25	50	100	10	25	50	100
AUC_{val}	0.90	0.90	0.90	0.90	0.89	0.91	0.93	0.93
ACC_{test}	0.64	0.64	0.65	0.65	0.65	0.64	0.63	0.63
BA_{test}	0.60	0.60	0.61	0.61	0.60	0.60	0.58	0.58
AUC_{test}	0.65	0.65	0.66	0.66	0.64	0.63	0.58	0.58
FPR	0.75	0.74	0.73	0.73	0.71	0.72	0.74	0.74
Time	0.31	0.42	0.62	0.97	1.64	2.75	5.54	10.73

Table A.4 – Influence of number of iterations on Random Forest Classifier and XGBoost for the (B) division.

n_{iter}	RFC				XGB			
	10	25	50	100	10	25	50	100
AUC_{val}	0.62	0.91	0.90	0.91	0.93	0.93	0.93	0.93
ACC_{test}	0.59	0.71	0.71	0.71	0.68	0.69	0.69	0.68
BA_{test}	0.50	0.66	0.66	0.65	0.62	0.63	0.63	0.62
AUC_{test}	0.70	0.81	0.81	0.80	0.79	0.82	0.82	0.81
FPR	1.00	0.54	0.54	0.56	0.62	0.59	0.59	0.60
Time	11.03	33.91	55.34	113.48	8.92	19.00	34.47	67.87

Table A.5 – Influence of number of iterations on Logistic Regression and SVM for the (C) division.

n_{iter}	LR				SVM			
	10	25	50	100	10	25	50	100
AUC_{val}	0.99	0.99	0.99	0.99	0.96	0.98	0.98	0.99
ACC_{test}	0.61	0.61	0.61	0.61	0.70	0.66	0.66	0.62
BA_{test}	0.56	0.56	0.56	0.56	0.67	0.62	0.62	0.57
AUC_{test}	0.68	0.68	0.68	0.68	0.78	0.74	0.74	0.66
FPR	0.74	0.74	0.74	0.74	0.58	0.66	0.66	0.70
Time	0.40	0.55	0.87	1.45	4.12	8.17	15.94	31.80

Table A.6 – Influence of number of iterations on Random Forest Classifier and XGBoost for the (C) division.

n_{iter}	RFC				XGB			
	10	25	50	100	10	25	50	100
AUC_{val}	0.86	0.98	0.98	0.98	1.00	1.00	1.00	1.00
ACC_{test}	0.57	0.70	0.70	0.70	0.68	0.68	0.68	0.68
BA_{test}	0.50	0.67	0.67	0.67	0.64	0.64	0.64	0.64
AUC_{test}	0.77	0.78	0.78	0.79	0.78	0.78	0.78	0.78
FPR	0.63	0.59	0.58	0.60	0.68	0.68	0.68	0.68
Time	19.72	61.00	99.15	195.27	8.44	18.64	33.95	65.49

APPENDIX B – HYPERPARAMETERS

Table B.1 – Default hyperparameters for the (A) division.

Model	Hyperparameter	Value
Logistic Regression	C	0.0390
SVM	C	14.3278
	gamma	0.0106
Random Forest Classifier	ccp_alpha	0.0022
	max_depth	27
	max_features	sqrt
	min_samples_leaf	7
	min_samples_split	2
	n_estimators	223
XGBoost	learning_rate	0.0930
	max_depth	26
	n_estimators	355
	subsample	0.3653

Table B.2 – Default hyperparameters for the (B) division.

Model	Hyperparameter	Value
Logistic Regression	C	0.1052
SVM	C	8.4008
	gamma	0.0342
Random Forest Classifier	ccp_alpha	0.0022
	max_depth	27
	max_features	sqrt
	min_samples_leaf	7
	min_samples_split	2
	n_estimators	223
XGBoost	learning_rate	0.0527
	max_depth	40
	n_estimators	309
	subsample	0.6536

Table B.3 – Default hyperparameters for the (C) division.

Model	Hyperparameter	Value
Logistic Regression	C	2.0630
SVM	C	14.3278
	gamma	0.0106
Random Forest Classifier	ccp_alpha	0.0022
	max_depth	27
	max_features	sqrt
	min_samples_leaf	7
	min_samples_split	2
	n_estimators	223
XGBoost	learning_rate	0.0866
	max_depth	31
	n_estimators	258
	subsample	0.2968

APPENDIX C – CONFUSION MATRICES

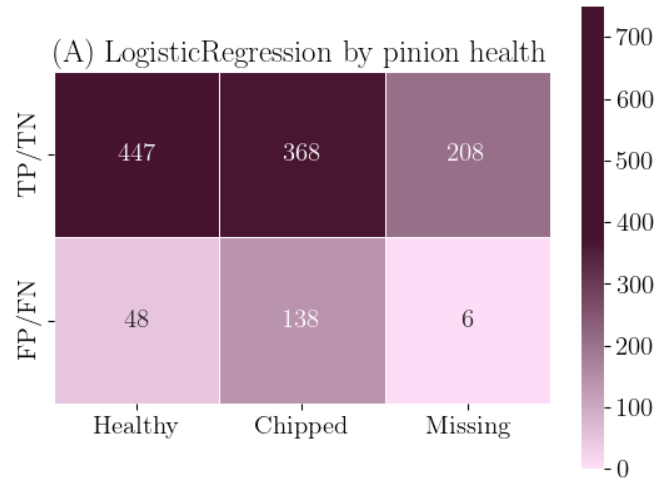


Figure C.1 – Confusion matrix for (A) division and Logistic Regression.

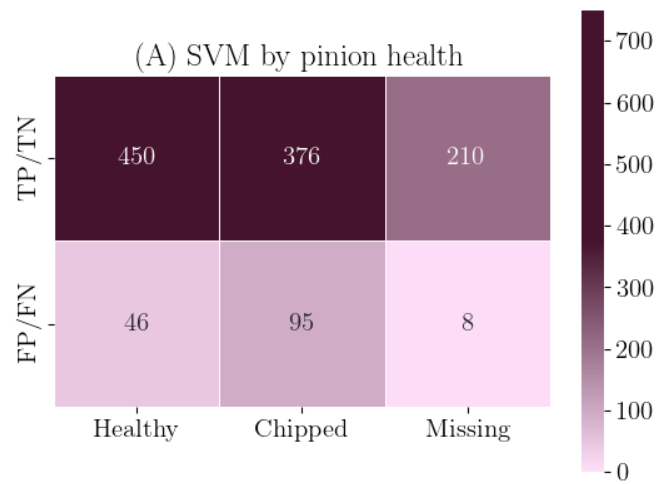


Figure C.2 – Confusion matrix for pinion condition for (A) division and SVM.

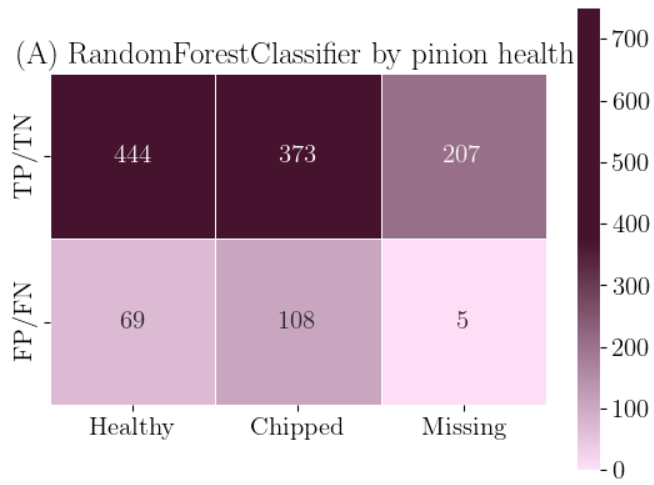


Figure C.3 – Confusion matrix for (A) division and Random Forest Classifier.

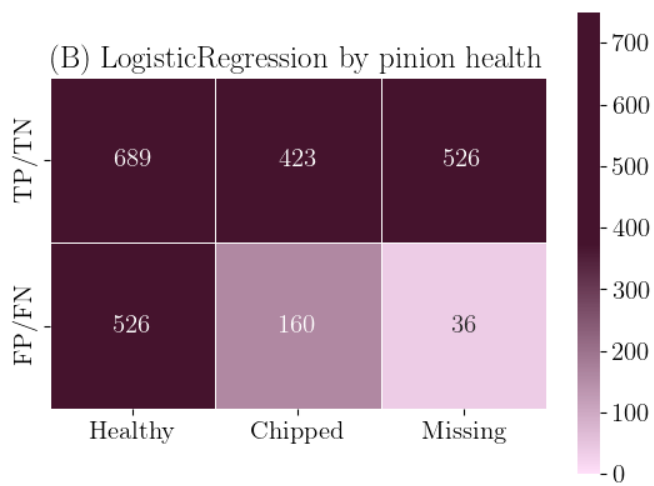


Figure C.4 – Confusion matrix for (B) division and Logistic Regression.

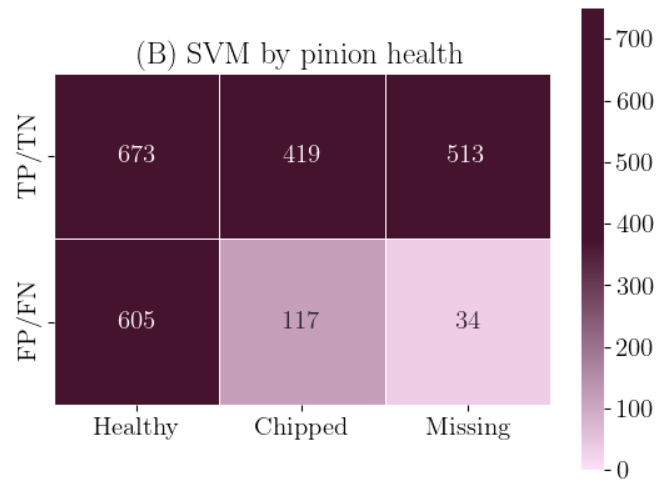


Figure C.5 – Confusion matrix for (B) division and SVM.

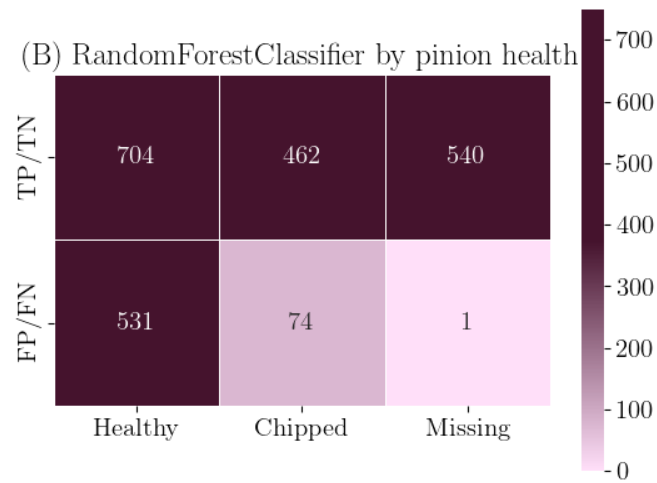


Figure C.6 – Confusion matrix for (B) division and Random Forest Classifier.

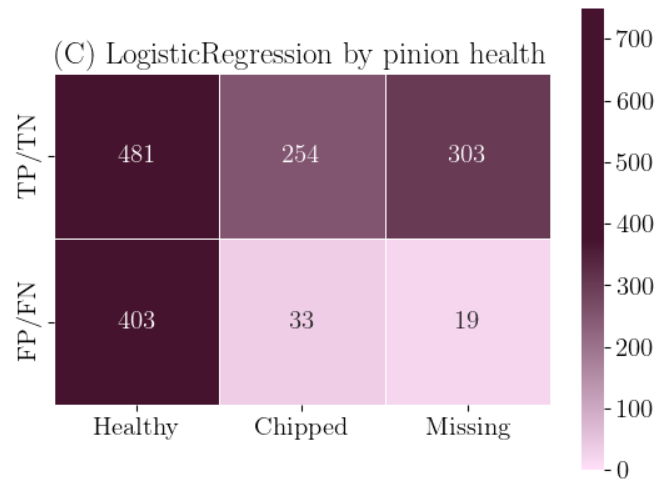


Figure C.7 – Confusion matrix for (C) division and Logistic Regression.

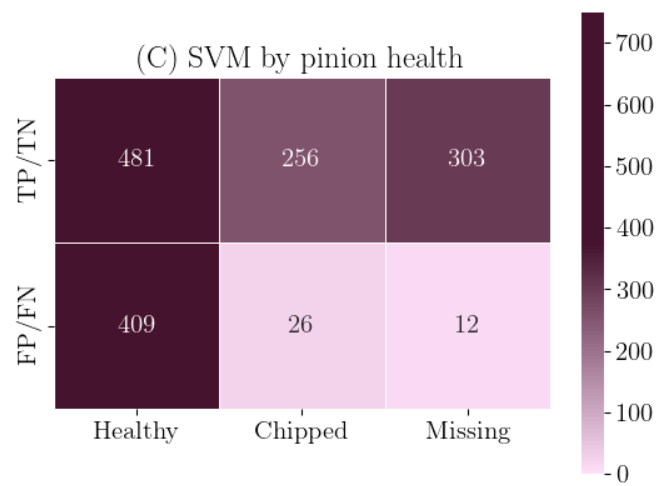


Figure C.8 – Confusion matrix for (C) division and SVM.

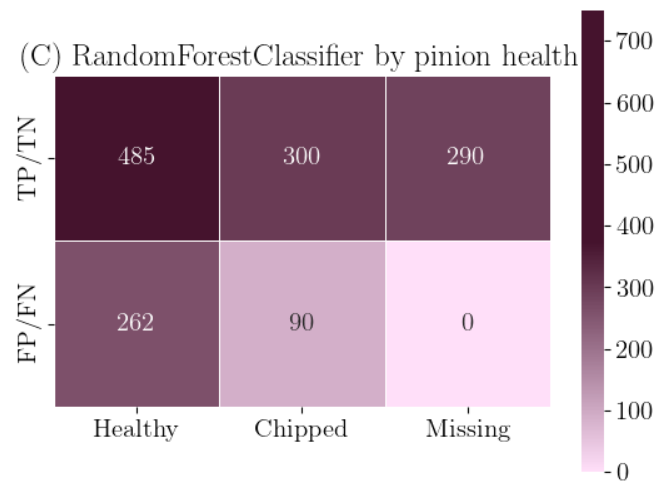


Figure C.9 – Confusion matrix for (C) division and Random Forest Classifier.

APPENDIX D – RESULTS FFT VS WELCH

Table D.1 – Comparison of FFT vs Welch’s method for Logistic Regression and SVM (A) division.

method	LR			SVM		
	welch	fft	none	welch	fft	none
AUC_{val}	0.93	0.93	0.93	0.95	0.94	0.95
ACC_{test}	0.82	0.82	0.81	0.87	0.86	0.84
BA_{test}	0.83	0.83	0.82	0.88	0.87	0.85
AUC_{test}	0.92	0.92	0.90	0.95	0.95	0.94
FPR	0.29	0.30	0.28	0.17	0.18	0.21
Time	1.13	1.10	0.99	26.30	26.33	20.19

Table D.2 – Comparison of FFT vs Welch’s method for Logistic Regression and SVM (B) division.

method	LR			SVM		
	welch	fft	none	welch	fft	none
AUC_{val}	0.91	0.90	0.90	0.93	0.92	0.91
ACC_{test}	0.64	0.66	0.67	0.63	0.64	0.65
BA_{test}	0.60	0.62	0.62	0.58	0.58	0.59
AUC_{test}	0.65	0.68	0.71	0.58	0.59	0.62
FPR	0.74	0.70	0.68	0.75	0.74	0.72
Time	0.83	0.84	0.78	8.43	8.52	6.85

Table D.3 – Comparison of FFT vs Welch’s method for Random Forest Classifier and XGBoost (B) division.

method	RFC			XGB		
	welch	fft	none	welch	fft	none
AUC_{val}	0.91	0.91	0.91	0.94	0.94	0.93
ACC_{test}	0.69	0.70	0.69	0.68	0.70	0.69
BA_{test}	0.64	0.65	0.64	0.62	0.64	0.64
AUC_{test}	0.79	0.82	0.84	0.80	0.83	0.82
FPR	0.60	0.50	0.58	0.64	0.57	0.62
Time	174.97	174.36	81.52	82.22	80.87	32.61

Table D.4 – Comparison of FFT vs Welch’s method for Logistic Regression and SVM (C) division.

method	LR			SVM		
	welch	fft	none	welch	fft	none
AUC_{val}	0.99	0.99	0.98	0.99	0.99	0.98
ACC_{test}	0.62	0.62	0.71	0.63	0.62	0.71
BA_{test}	0.57	0.57	0.69	0.58	0.57	0.70
AUC_{test}	0.71	0.69	0.75	0.70	0.67	0.74
FPR	0.77	0.71	0.86	0.71	0.68	0.82
Time	1.08	1.11	0.90	23.86	23.93	18.31

Table D.5 – Comparison of FFT vs Welch’s method for Random Forest Classifier and XGBoost (C) division.

method	RFC			XGB		
	welch	fft	none	welch	fft	none
AUC_{val}	0.98	0.98	0.98	1.00	1.00	0.99
ACC_{test}	0.70	0.71	0.70	0.68	0.65	0.70
BA_{test}	0.67	0.68	0.67	0.65	0.61	0.68
AUC_{test}	0.77	0.78	0.78	0.77	0.77	0.79
FPR	0.58	0.56	0.55	0.76	0.75	0.64
Time	299.03	300.64	135.80	77.96	77.94	31.37