IN SITU RAILWAY TRACK FAULT DETECTION USING RAILCAR VIBRATION

by

Jeffrey L. Pagnutti

A thesis submitted in partial fulfillment
of the requirements for the degree of
Masters of Applied Science (MASc) in
Natural Resources Engineering

The School of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE

Laurentian Université/Université Laurentienne
School of Graduate Studies/École des études supérieures

Title of Thesis Titre de la thèse	*IN SITU* RAILWAY TRACK FAULT DETECTION USING RAILCAR VIBRATION
Name of Candidate Nom du candidat	Pagnutti, Jeffrey L.
Degree Diplôme	Master of Applied Science

Department/Program Département/Programme	Natural Resources Engineering	Date of Defence Date de la soutenance	January 24, 2014

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Markus Timusk
(Supervisor/Directeur de thèse)

Dr. Brent Lievers
(Committee member/Membre du comité)

Dr. Krishna Challagulla
(Committee member/Membre du comité)

Dr. Lorrie L. Fava
(External Examiner/Examinatrice externe)

Approved for the School of Graduate Studies
Approuvé pour l'École des études supérieures
Dr. David Lesbarrères
M. David Lesbarrères
Director, School of Graduate Studies
Directeur, École des études supérieures

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

Abstract

This thesis investigates the development of an automated fault detection system developed for a novel lightweight railway material haulage system; in particular, the study aims to detect railway track faults at the incipient stage to determine the feasibility of maintenance decision support, ultimately with the function of preventing catastrophic failure. The proposed approach is an extension of the current state of the art in fault detection of unsteady machinery.

The most common railway track faults associated with train derailment were considered; namely, horizontal and transverse crack propagation, mechanical looseness, and railbed washout were the faults of interest. A series of field experiments were conducted to build a database of vibration, speed, and localization data in healthy and faulted states. These data were used to develop, investigate, and validate the effectiveness of various approaches for fault detection.

A variety of feature sets and classification approaches were investigated to determine the best overall configuration for the fault detector. The feature sets were used to condense data segments and extract characteristics that were sensitive to damage, but insensitive to healthy variations due to unsteady operation. The pattern recognition classifiers were used to categorize new data members as belonging to the healthy class or faulted class.

The fault detection results from the proposed approach were promising. The feasibility of an automated online fault detection system for the lightweight material haulage system examined in this study was confirmed. The conclusions of this research outline the major potential for an

effective fault detection system and address future work for the practical implementation of this system.

Acknowledgements

I would like to thank my supervisor, Dr. Markus Timusk, for his confidence, expertise, support, and inspiration during the course of this work; his experience in this field and advice has been an invaluable tool throughout the process of this study. I would also like to extend thanks to the engineering faculty at Laurentian University for their support and instruction.

I am greatly thankful to the Centre for Excellence in Mining Innovation for their financial support and to Allan Akerman for his continuous encouragement.

I would also like to thank Rail-Veyor® Technologies for their industrial partnership and technical support. I would like to thank Patrick Fantin and John McCall for their technical support and for handling the logistics on site. In addition, I am grateful for the technical assistance from Alexander Douglas, David Creasey, Esko Hellberg, and Eric Zanetti.

Finally, I would like to thank my family for their tremendous moral support, encouragement, and love.

To Paule.

Table of Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

Machine maintenance is an integral part of prolonged equipment life, in any case. The requirement for maintenance begins at component damage, regardless of the cause. Component damage is an inevitable aspect of operation that is corrected by component replacement. In many cases, neglect of component replacement increases potential for more severe machine failure in the future. For this reason, maintenance is vital for healthy operation. Immediate benefits of an effective maintenance strategy are increased reliability, additional production, and reduced safety hazards.

With evermore complex, large-scale operations around the world, maintenance has become more complex. The demand for more robust maintenance systems introduced condition-based maintenance. Condition-based maintenance strategies have adopted the philosophy to make repairs based on the actual condition of the equipment. This strategy offers several advantages over the conventional scheduled maintenance strategy: direct benefits are that critical equipment are only repaired if they need to be, avoiding unneeded down time due to premature maintenance; economic savings emerge since healthy components are not replaced; and, ideally, catastrophic failures due to undetected damage are avoided.

The principle of condition-based maintenance seems trivial in theory, but in practice, it requires a continuous knowledge of machine condition, which cannot be measured directly; information regarding machine condition must be inferred from measurements. These inferences of machine condition are based on indirect measurements of machine parameters such as temperature,

pressure, vibration, shaft position, acoustic emissions, and etcetera. This work focuses on developing computerized approaches to define machine condition based on measurement data.

1.1 Background

This section provides a brief background of maintenance engineering, condition monitoring, and railway systems. The chapters that follow will borrow terms and concepts from the background section, while expanding on concepts and the tasks involved for the experimental aspects of this work.

1.1.1 A History of Maintenance

In the years following World War II, technological advancements and high-volume production provided the means to put relatively complex products into the hands of consumers worldwide [1]. The war sparked innovation and demanded high production rates to supply military with munitions and equipment. During these times, maintenance of manufacturing equipment was required to reduce work stoppages due to equipment failure. It was not until the 1980s when maintenance became a priority due to safety and environmental standards [1]. Industries quickly learned that machine maintenance was essential to reduce production stoppages and increase product quality. The evolution of maintenance introduced three maintenance schemes: reactive, scheduled, and condition-based maintenance.

Reactive maintenance is a simple strategy to manage non-critical equipment. The principal is to perform maintenance once the equipment fails to accomplish its task correctly. This ensures a maximum life cycle of the component itself. One might consider a light bulb to be a component best suited for a reactive maintenance strategy. Once again, reactive maintenance strategies are

usually imposed on simple machines whose failures do not constitute a serious negative consequence.

Scheduled maintenance was developed to reduce the frequency of catastrophic failure due to a critical component failure. Components that are critical to machine operation will generally have a periodic maintenance schedule. Complex machinery with several components will have sacrificial parts that wear over time and require replacement. Replacement schedules are determined by failure histories, i.e., the average life of a part with the conservation of some specified safety factor. For example, if a part fails on average after six months of use, a periodic maintenance scheme might schedule part replacement after five months to ensure the part is unlikely to fail during operation. This maintenance scheme reduces downtime by reducing component failure; it also promotes a safer working environment and reduces risks of environmental hazard. However, the drawback of the probabilistic approach to scheduled maintenance is inherent in its design; there is always a chance that a component will prematurely fail, or is replaced in good health since the schedules are based on statistical estimates of component life cycle. Condition-based maintenance provides a solution to these drawbacks.

Condition-based maintenance strategies are being developed for industries where machine failure results in production stoppage, significant economic loss, catastrophic environmental damage, loss of life, or a combination of these events. It may also be more economical to practice condition-based maintenance in the case of production bottlenecks or high-capital machinery to increase equipment availability. In condition-based maintenance systems, sensors are used to measure physical parameters that are analyzed to determine the machine health. Condition monitoring is used as a means of diagnostic data collection to provide insight regarding the

current condition of the machinery. As the condition of the machinery deviates from the healthy state, a computer program provides maintenance decision support by alerting the operator that the system is responding abnormally, and if possible, estimates the severity of damage and remaining life. Many condition monitoring systems are application-specific, and in turn expensive. As such, candidature of a condition-based maintenance strategy is generally reserved for industrial, healthcare, or military equipment. The challenge of condition-based maintenance is to measure damage-sensitive, noise-resistant parameters and relate them to a state of health; namely, healthy or damaged. The difficulties arise in the presence of environmental noise, variations in operational speed and duty, and practical limitations. The resulting inferences regarding machine health are typically the product of high-level mathematical models, pattern recognition algorithms, and many other interdisciplinary tools.

While each of the maintenance strategies has its place, condition-based maintenance is the underlying subject of discussion for this work. Furthermore, many commercially available condition monitoring systems have become available to industry across a wide range of applications. Common applications of condition monitoring that are well-established predominantly include rotating machinery such as pumping and ventilation systems, gearbox monitoring in the aircraft industry, oil pipeline monitoring, electromechanical excavation units for mining, and bridge structural health monitoring.

1.1.2 Maintenance in the Mining Industry

The current mining industry focusses on maintenance as a tool to reduce downtime, increasing availability and production rates. The costs associated with maintenance in mining can amount

to 30% to 50% of total cost of production [2]. Given the significant cost associated with maintenance, it is a prime candidate for optimization.

1.2 The Rail-Veyor® Material Transport System

1.2.1 System Overview

For the experimental part of the present study, a series of tests were performed on a novel material haulage system called Rail-Veyor®. As an alternative to haulage trucking and traditional railway or conveyor-based methods, the system uses a passive train propelled by truck tires that mesh with its side-plates; the train is shown passing through a drive station in Figure 1-1.

Figure 1-1: The Rail-Veyor® train passing through a drive station (taken from [3]).

The prime mover consists of two rubber tires, each coupled to a gearbox, driven by an electric motor. Each drive station consists of two drives that are identical in construction, but oriented in opposing configurations; the drive station is further illustrated in Figure 1-2.

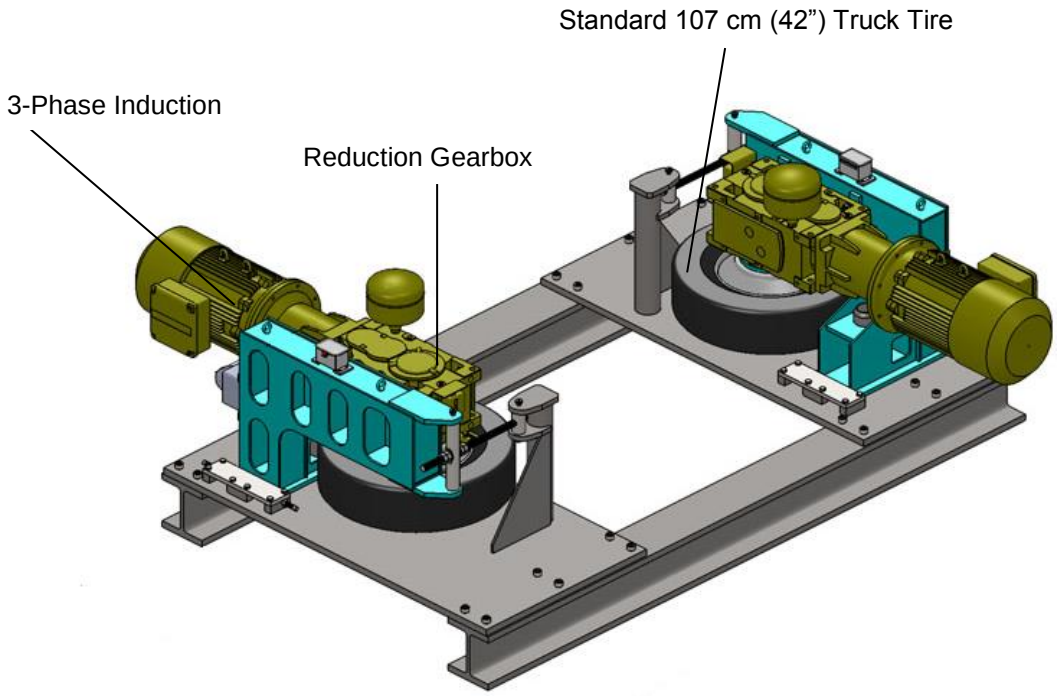


Figure 1-2: The Rail-Veyor® drive station configuration (taken from [3]).

The train engages with two rubber tires and is propelled along the railway track to the next drive station. The system uses standard 18-27 kg/m (40-60 lb/yd) lightweight rail. Since the construction of the railcars is lightweight when compared to traditional railcars, the Rail-Veyor® not only uses lightweight track, but also does not require a large railbed seen in freight railway systems; the railway track is a floating construction that uses steel tie plates located approximately 2 m (2 yd.) apart. The general layout of a simple Rail-Veyor® system is illustrated in Figure 1-3.

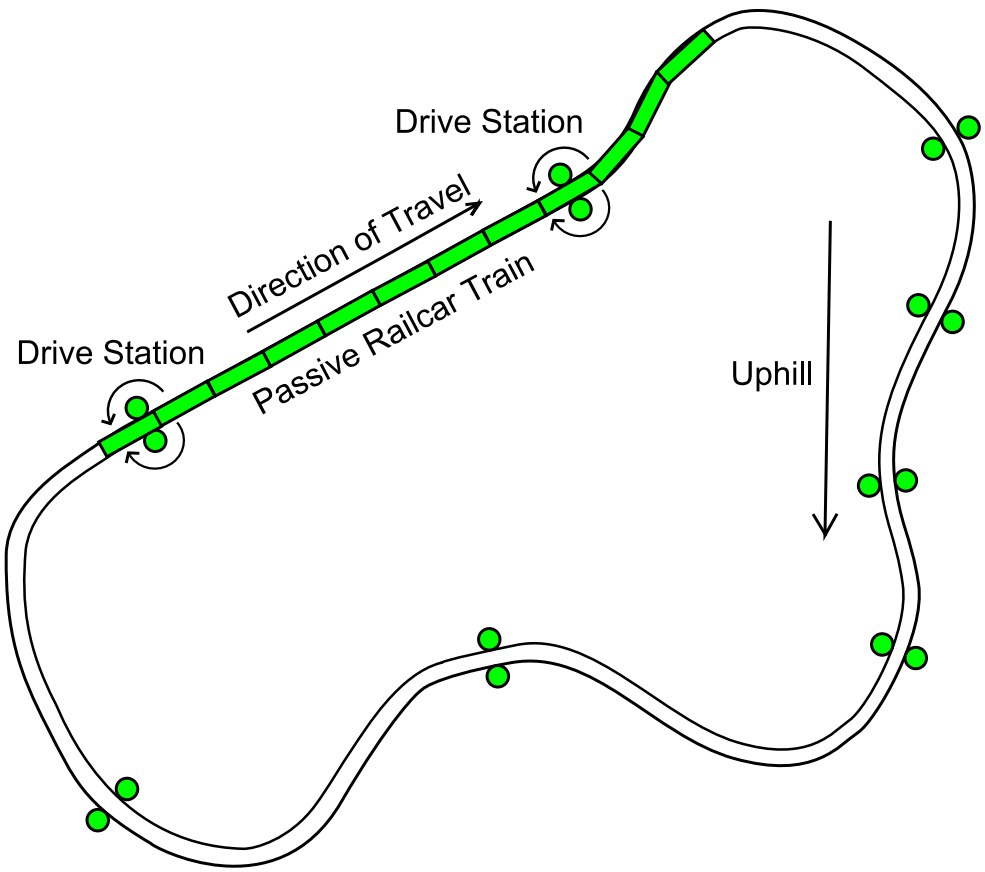


Figure 1-3: The general layout of a Rail-Veyor® material haulage system.

One element of the system at risk of failure is the railway track itself. Since the track has no redundancy, its preservation is important. Some of the maintenance challenges associated with such a system are closely linked with application-specific aspects. Considering that the Rail-Veyor® is geared towards mining haulage applications, many of the environments in which it is used are either harsh, remote, or high-throughput. As an example, in an underground mine, train derailment in a drift would result in a harsh, close-quartered working environment that makes re-railing the train more difficult. Similarly, in long haulage applications where the train travels through remote locations, a derailment would result in significant loss in production time. In

effort to develop a solution to these maintenance challenges, rail-wheel interaction over extended time and duty causing rail failure was examined in the study. The Rail-Veyor® test site is shown in Figure 1-4.

Figure 1-4: The Rail-Veyor® demonstration site used for *in-situ* data collection is situated in Sudbury, Ontario, Canada.

The inherent benefits of the Rail-Veyor® system in a mining operation compared to conventional rail systems are the smaller physical footprint which allows for reduced drift sizes and, in turn more stable ground conditions. In addition, the electrical propulsion system produces minimal emissions compared to diesel-powered trucking methods, which reduces the ventilation requirements, in turn reducing a significant portion of operational cost. Furthermore, one of the most attractive features of the Rail-Veyor® system is its simplicity in construction, functionality, and maintainability. Due to the simple construction, minimal training is required for repair and maintenance procedures.

During the period of data collection for the present research, the Rail-Veyor® system was still in the prototype stages; for this reason, comprehensive fault data and maintenance history specific to the Rail-Veyor® system was not available for this study. Therefore, the experimental design was adopted from similar equipment in the lightweight rail industry; this will be outlined in sections 3.4 and 3.5.

1.2.2 History of Rail-Veyor®

The Rail-Veyor® concept was invented by Mike Dibble in 1999. After two short years, a demonstration plant was running in Florida, United States. A feasibility assessment revealed potential cost savings (at haulage distances greater than 4 km) and environmental benefits as compared to conventional haulage methods. In search of a market for the Rail-Veyor®, Mike Dibble encountered Canadian entrepreneur Risto Laamanen. Risto then played a crucial role in securing a partnership with Vale to commission a demonstration site in Sudbury, Ontario, Canada.

In 2011, commissioning of the Rail-Veyor® system at Vale's Creighton mine 114 ore body began; the system was completed by April, 2012. The typical development advance rates at Vale's Creighton mine were roughly 60 m (200 ft.) per week before the Rail-Veyor® installation, which increased to 120 m (400 ft.) per week with the use of the Rail-Veyor® system. Production rates also increased from roughly 1,250 tonnes to 2,500 tonnes per day, and have the potential to increase to over 4,000 tonnes per day with an expansion of the current Rail-Veyor® system.

1.3 Railway Incidents and Derailment

Despite being one of the most efficient and safe modes of transport, railway systems are known to fail, with train derailment as a principal mode of failure. Train derailment can lead to loss of capital, damage to the environment, and loss of life. Traditional maintenance strategies for railway infrastructure are run-to-failure, scheduled maintenance, or condition-based maintenance. Run-to-failure philosophies can result in catastrophic failure and are not well-suited for the railway industry. Periodic maintenance strategies have proven to perform well for railcar maintenance, but may not be the most efficient strategy for railway infrastructure such as railway track. This study aims to develop an automated condition-based maintenance system for railway infrastructure.

In addition to potential environmental damage and possibility of death or injury, the costs associated with railway track damage can be significant. A report from the European Railway Research Institute (2000) estimated nearly $100 million (€70 million) per year is required for ultrasonic inspections alone; this does not include costs associated with repairs. In addition, Cannon *et al.* (2003) estimated that rail defects cost the European Union nearly $3.2 billion (€2 billion) per year, not including pre-emptive corrections or costs resulting from derailment.

While the above figures are rough cost estimates, they highlight the severity of the problem. The following section outlines the condition monitoring system breakdown used in this study.

1.4 An Overview of Condition Monitoring for Fault Detection

The key to performing maintenance based on condition is timely and accurate information on machine health. The challenge arises in the fact that machine health cannot be directly measured. One of the main tasks in developing a successful automated fault detection system involves taking indirect measurements such as temperature, to identify machine health. Machinery conditions can be monitored to provide insight regarding component health. The design of a condition monitoring system can be compartmentalized into modules or tasks that define the condition monitoring system architecture.

At the forefront, the machinery operates as intended, and is controlled via operator or predetermined autonomous routine. Feedback to the operator or control system is achieved through various sensors. Different phenomena such as oil temperature or pressure, shaft speed, vibration, or acoustic emissions can be measured to describe the operating conditions and machinery health. These parameters on their own generally do not provide enough feedback to show indications of component damage; this creates a multivariate problem that involves multiple parameters, dynamic external forces, and varying duty cycles. Therefore, automated fault detection systems generally require computational aid to automate the inspection and analysis process of the sensory data. To simplify the general problem, intelligent parameter selection (or sensor selection) is the first step in condition monitoring.

Data acquisition is used to gather historical information of the sensor signals. The historical signals are stored to develop a basis for comparison of new data elements. The basic principal is that, initially, the machinery is healthy and exhibits a certain signal, whereas, damage exhibits a dissimilar signal that can be measured.

Unsegmented time series data sets are generally too large to be useful for online detection of faults, so they are segmented or cropped about points of interest. These segments of data that contain sought-after information from the monitoring process are then further reduced by means of feature extraction.

In order to describe a large time series data sample in a compact form, one can extract statistical parameters forming what is referred to as a feature vector. Examples of statistical features of the data segment may include its root-mean-square value, population standard deviation, kurtosis, or arithmetic mean; however, any scalar or vector metric can be used to characterize the data segment. By defining well-suited feature vectors, it is intrinsic that the next stages in fault detection become simplified since well-suited features provide a better representation of machine condition than poorly selected features.

Classification is the next stage in the condition monitoring process where feature vectors corresponding to specific data segments are categorized into the healthy group or the unhealthy group. For the task of classification, the concept of training is used to develop a parametric model that classifies new measurements as healthy or unhealthy. Training data for healthy operating conditions are usually available since systems generally operate in an undamaged state; however, fault data is more difficult to acquire since production machinery cannot be induced with faults for the purpose of experimentation for practical reasons. Therefore, classification models can be established using the abundant healthy data. The role of the classifier is to identify if new measurements conform to the well-defined class of training data, otherwise, they are rejected as outliers to the healthy class, i.e., they are unhealthy. Whether faulted entities are

diagnosed or not defines the system as a fault diagnostic system or simply a fault detection system.

An overview of the general classification-based condition monitoring architecture is outlined in Figure 1-5 below.

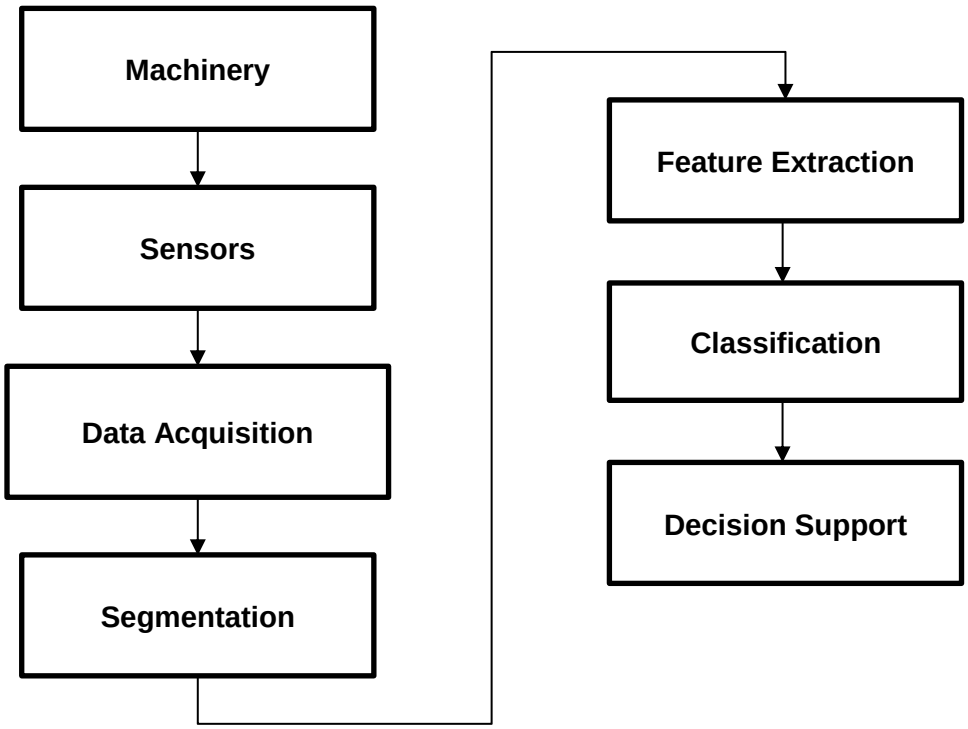


Figure 1-5: The general classification-based condition monitoring architecture.

1.5 Research Goals

Current works in railway track inspection systems use a variety of sensing techniques to accomplish fault detection; these techniques include ultrasonic inspection, magnetic induction inspection, pulsed eddy current, image recognition systems, and radiography. These systems

perform acceptably for their purpose; however, they are not suitable for the high-traffic nature of a mining production line.

The goal of this research is to develop a vibration-based railway track inspection system that integrates with existing machinery to perform the task of maintenance decision support. Intrinsic to the development of such a system is the analysis and refinement of signal processing techniques associated with the decision support system. The findings of this research aim to determine if vibration-based condition monitoring can be used for fault detection in railway systems. Furthermore, it is promising that the findings of this research will draw conclusions of interest for other domains of condition monitoring. The following section presents the task flow of the research process.

1.5.1 Work Flow

This section defines the research plan and task flow associated with the development of a vibration-based condition monitoring system for the Rail-Veyor® system. Figure 1-6 outlines the approach taken for the development of the condition monitoring system; a detailed description of the major tasks for this work follows.

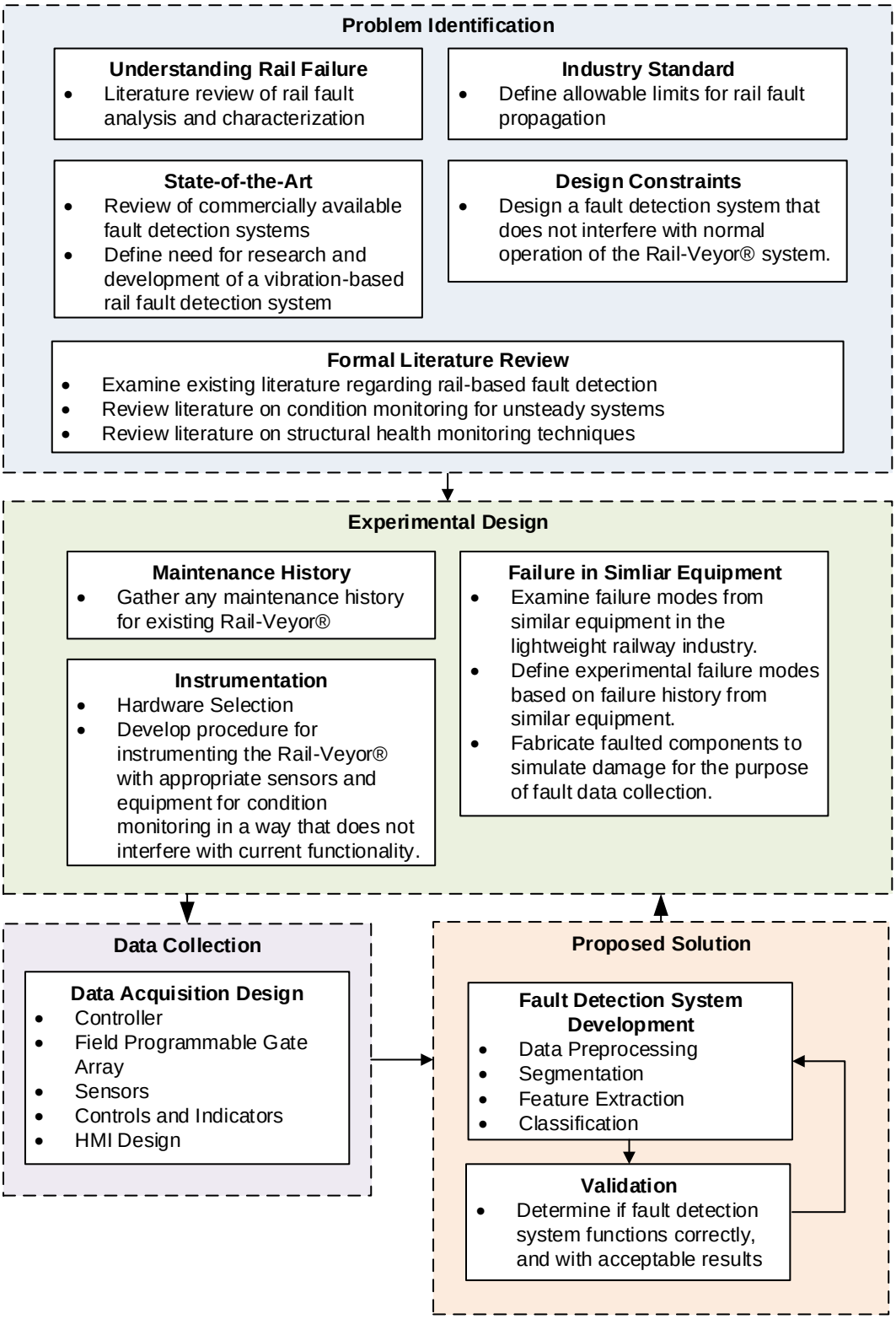


Figure 1-6: Task flow for development of a vibration-based condition monitoring system for Rail-Veyor®.

The first stage of the research was problem identification. Within this task, the author investigated the general nature of rail failure as a broad topic, wherein the general background of railway anatomy and failure modes were established. From this, the industry standards were reviewed to determine the allowable limits for various damage severity. Subsequently, the existing state-of-the-art in automated and manual railway inspection systems was examined. Following this survey of the state of the art, the shortcomings of the existing systems were highlighted. This introduced the design challenges of a novel condition monitoring system that aimed to overcome these shortcomings and could be applied to the Rail-Veyor® system. The design constraints were then analyzed and a formal literature review of rail-based fault detection and condition monitoring of unsteady systems was conducted to develop an experimental approach and preliminary design for further development of the fault detection system.

The experimental design within this research was based on the availability of a live demonstration site of the Rail-Veyor® system for experimental testing. This is a rare occurrence in industry where a full-scale, operational facility or machine is available for investigational purposes. As such, the absence of comprehensive maintenance history for the Rail-Veyor® system required the experimental design to be based on failure histories of similar equipment. The lightweight railway industry was used as a standard for failure modes since it shared many similarities; experimental design was based on these failure modes. Specific failure modes were then selected and simulated by inducing damage into otherwise-healthy rail specimens. Following the definition of experimental failure modes, suitable sensors for condition monitoring of the railway track were selected. Additionally, a data acquisition platform was specified and the human-machine interface (HMI) for experimental data logging was designed and fabricated.

The data collection stage consisted of designing and programming a suitable data acquisition application for the task of data logging. Sensor integration, as well as controller and sensor interfacing were the focus of this stage in the design. Following the design and implementation of the data acquisition system, the actual data logging commenced at the Rail-Veyor® demonstration site in Sudbury, Ontario, Canada.

The development of the proposed solution was organized based on knowledge gained from a formal literature review. Data preprocessing was used to normalize data and condition it into a more usable state. Segmentation was used to define a comparable dataset to reduce the overall variability during the development stage. A basis of feature sets and classification algorithms were selected for analysis and investigation as final steps of the fault detection system development. The validation stage was then used as a means of performance evaluation to determine if the fault detection system was successful. Following the validation stage, if the overall system performed poorly, elements of the fault detection design were adjusted in an attempt to produce improved results; however, the focus of this study was to evaluate the feasibility and compare different fault detection methods rather than to produce an optimized fault detection system.

1.6 Structure of this Thesis

This thesis describes the development of a condition monitoring system and outlines valuable conclusions drawn from the experiments. The structure is as follows:

Chapter 1 – Introduction

- Defines landmarks of maintenance engineering history, along with a background of the case study of this research, namely, Rail-Veyor®;
- Discusses railway system failure and a brief overview of condition monitoring for fault detection;
- Outlines research goals of this study.

Chapter 2 – Literature Review

- Reviews machinery condition monitoring;
- Outlines railway anatomy and the state of the art in rail flaw detection.

Chapter 3 – Field Experiments and Data Collection

- Describes the development of experimental design and includes specifications of the sensors and instrumentation used for this study.
- Investigates preliminary response signals.

Chapter 4 – Signal Processing and Fault Detection

- Proposes signal processing techniques for fault detection and compares the classification performance of the different approaches.

Chapter 5 – Conclusions and Future Work

- Draws conclusions from the study and encourages future work associated with this study.

This concludes the first chapter, in which the relevant background information has been presented. The next section provides a critical literature review of relevant material that formed the foundation for this research.

Chapter 2

2 Literature Review

A general background regarding the foundations of this study is presented in this chapter. This section begins with a review of machinery condition monitoring. A background of railway anatomy and failure modes follows. The current state of the art in industry relating to rail fault detection is then reviewed, which is followed by recent developments in railway condition monitoring.

2.1 Introduction

This chapter describes the state of the art in condition monitoring from an academic perspective and that in place in the railway industry. The goal is to investigate and report on principal works in the literature pertaining to condition monitoring and fault detection of rail-based systems.

2.2 Machinery Condition Monitoring

Machinery condition monitoring plays a central role in condition-based maintenance. The role of condition monitoring in condition-based maintenance is to establish a history of baseline data that represents healthy operation of the machinery, and to compare current measurements to that history to infer machine condition. The principle is deceptively simple: if new data elements exhibit an operational response similar to the historical elements, they also represent the healthy machine condition; otherwise, they represent an unknown condition, typically damaged. The aforementioned concept describes the most trivial operational situation. In many applications, systems operate under variable duty cycles, loading conditions, and environmental conditions that complicate the task of condition monitoring. Due to this operational variability, it becomes

increasingly difficult to compare new data to historical data since the operating conditions may differ. In practice, condition monitoring systems must be designed to be robust and insensitive to operational changes, but remain sensitive to monitored parameter changes due to equipment damage.

2.2.1 A Review of Condition Monitoring Techniques

As a maintenance practice, condition monitoring involves performing maintenance based on the actual condition of the machine. The central challenge to this approach is that it is rarely possible to measure the condition of a machine directly. For example, a cracked ball in a failing rolling element bearing cannot be observed with the naked eye. However, this fault can be observed indirectly through the change in vibration signature or temperature of the bearing. Therefore, the general idea behind fault detection in condition monitoring of machinery involves the use of transducer signals as an indirect means to determine if machinery is operating in a fault free condition, or shows signs of incipient damage that requires maintenance. As such, it is common practice in this field to monitor changes in transducer response as an indication that a fault is present. For components whose failure modes are well-known, characteristic frequencies from the component's vibratory response might shift, signifying a change in the physical connectedness or integrity of the system. The system response should be similar to the response that is observed under healthy operation; therefore, any significant change may be classified as characteristic of failure [4].

A range of different transducers can be used to characterize failure modes in machinery. Component failure can present itself through various measureable physical phenomena such as vibration, temperature, sound, oil particle content, acoustic emissions, torque, wave propagation, and many others. Depending on the type of failure, certain transducers capture the response with a more observable signal change [5]. In the case of condition monitoring, it is important to select the sensors that are most likely to reveal signal changes for expected modes of failure. Figure 2-1 illustrates examples of fault conditions and the corresponding measurement parameters that are typically used to characterize them. In the cases of rotating machinery faults as presented in Figure 2-1 it is evident that vibration is a good general choice for most expected fault conditions. Furthermore, vibration transducers present many practical advantages over other types of sensors; this is discussed in detail in section 2.5.

Parameter / Faults in Rotating Machinery	Vibration	Temperature	Pressure	Torque	Position	Flow	Lubrication Analysis
Unbalanced Shaft (Static)	X				X		
Unbalanced Shaft (Dynamic)	X				X		
Damaged Bearing	X	X		X			X
Damaged Gear	X						X
Lubrication Breakdown	X	X					X
Mechanical Looseness	X						

Figure 2-1: Machine Fault versus Parameter (Figure adapted from [5])

2.2.2 Vibration Signals for Fault Detection

While vibration signals contain useful characteristics of the system response, further processing is required to extract fault signatures from the overall signal. As an example, rolling element bearings are excellent candidates for vibration-based fault detection. Figure 2-2 depicts typical vibratory signals from healthy and damaged rolling element bearings. Comparing the healthy bearing signal to the faulted bearing signal in Figure 2-2, it is immediately evident that the faulted bearing is well-characterized by the vibratory response.

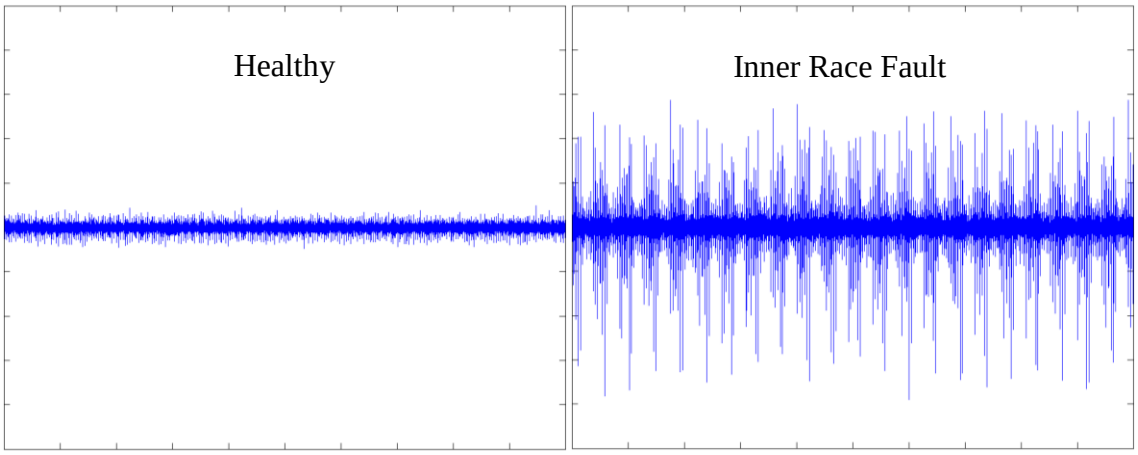


Figure 2-2: Vertical Accelerometer Signal of a Healthy Rolling-Element Bearing (left) and an Inner Race Fault in a Rolling-Element Bearing (right) at 13 Hz (800rpm).

Characteristic bearing frequencies that are a function of the operating speed are well-established indicators of condition for vibration-based monitoring of rotating equipment. From these characteristic frequencies, a frequency domain spectrum analysis of vibration transducer signals can effectively illustrate sideband frequencies associated with various bearing failure modes [6]. This type of analysis is particularly effective in signals that change from linear to nonlinear as in the case of a loose bearing. These aforementioned frequency domain techniques are highly

effective in signals that exhibit a periodic response since they provide insight into components of the frequency spectrum.

Time domain analysis can be very useful for transient system response since impulsive events would be dwarfed if the signal were transformed into the frequency domain [7]. In the case of time domain analysis, statistical features are a common starting point for feature selection since deviations from healthy operation are often well-characterized by descriptors such as root-mean-square, kurtosis, or variance of the sample. Similarly, regression models can be fit to time-series data and the model parameters can then be used as features that are indicative of machine condition [8].

Transient characteristics such as rise time, transient duration, and settling time can be derived from transducer signals for characterization of non-stationary events. This analysis is typically applied to acoustic emission signals, which sense micro strain energy bursts on materials at very high frequencies [9].

Time-frequency domain analyses have also been proven effective for structural health monitoring in many applications [10-15]. Wavelet analysis is a common time-frequency technique in which a basis function called the mother wavelet is scaled and shifted across a signal to extract time-frequency components of the signal. This technique has been used for fault detection in various machinery with successful results. The advantage of using a time-frequency technique such as wavelet analysis as opposed to a frequency domain technique such as Fourier analysis is that the temporal information is retained; this is especially valuable in the field of fault detection in unsteadily operating systems. Time-frequency techniques propose a balance in

time and frequency resolution that cannot be attained using conventional Fourier analysis (including the short-time Fourier transform).

In situations with steadily operating machinery with constant duty cycles, the mechanical response is constrained and the task of monitoring for faults is relatively straightforward. In comparison, machinery operating under variable speed and duty propagate these variations in transducer response, creating additional variability in the signals. The additional variability observed in unsteady operating machinery further complicates the task of machine learning and detection of anomalous events. Therefore, when techniques such as Fourier analysis are applied to an unsteady signal, the frequency content is revealed, but can appear as a misrepresentation of the original signal. Figure 2-3 illustrates the potential dissimilarities in a stationary and nonstationary signal with the same frequency content.

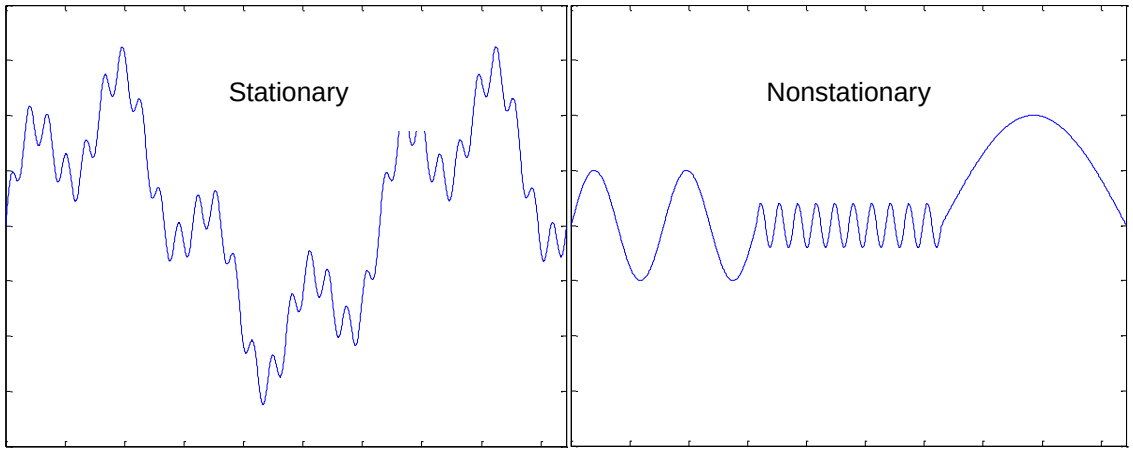


Figure 2-3: A comparison of stationary and nonstationary signals with the same frequency content.

Figure 2-4 reveals the ambiguity when using Fourier analysis on both stationary and nonstationary signals; in the frequency domain, it is not possible to maintain the temporal data contained in nonstationary signals.

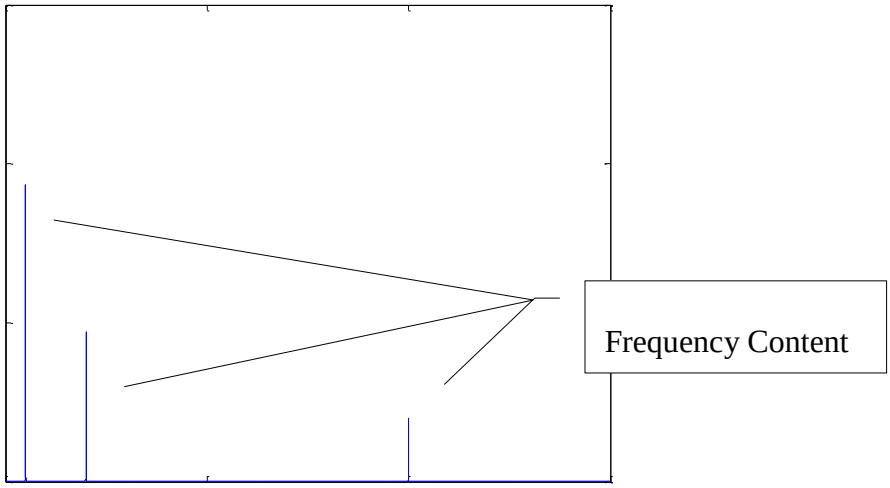


Figure 2-4: The fast Fourier transform (FFT) of the stationary and nonstationary signals of Figure 2-3 both have the same shape of frequency spectrum.

2.2.3 Condition Monitoring of Unsteadily Operating Machinery

Many systems operate at variable speed and load. Mobile equipment is a prime example where steady state is rarely an operational mode. The variation in operational speed and load is what defines unsteadily operating machinery. Research investigating the fault detection of variable speed and duty machinery is of great interest in industry and academia. The difficulties in developing a fault detection system for unsteadily operating equipment are associated with the operational variability; since under healthy conditions, the system response is constantly in flux due to speed and load changes. Development of condition monitoring systems for machinery operating in unsteady conditions is a popular topic of interest because steady state machinery is

finally becoming well understood; the more complex problem of fault detection in unsteadily operating machinery can now be solved using knowledge gained from fault detection in stationary machinery. Furthermore, these unsteadily operating machines are often large, complex and expensive pieces of equipment.

There are various techniques in the literature that attempt to mitigate the difficulties of extracting fault signatures from time varying vibration measurements from unsteadily operating machinery. Toliyat *et al.* [16] investigated the use of wavelet packet decomposition of nonstationary signals used for rail defect diagnosis. The study examined magnetic induction response signals from healthy, vertically split, piped, horizontally split, and transverse split rail. Raw transducer signals were decomposed twelve levels using wavelet packet decomposition. A damage index was defined as the summation of the discrete wavelet transform coefficients, which represented apparent energy. The resulting energy distributions across the 12 levels of decomposition show separation when comparing healthy and faulted damage indices.

Firlik *et al.* [17] conducted a study involving a light rail vehicle, equipped with 36 sensors, for the task of condition monitoring for fault detection. The goals of the study were to detect and localize rail vehicle and track faults. The system was designed for passenger tram systems and was put into operation during the investigation to determine the feasibility of on-line fault detection for main tramway vehicles and track. The system proved to be a promising solution for health condition monitoring of the tramway infrastructure.

Tsunashima *et al.* [18] also conducted research in condition monitoring of railway track, but investigated conventional and high-speed railway infrastructure. The first part of the study

investigated a fault detection system for rail corrugation using railcar vibration through wavelet-based multi-resolution analysis (MRA). Since significant differences between the healthy and fault vibration signals were not observable, the wavelet technique was chosen to extract time-frequency information using a multi-level decomposition. The wavelet decomposition resulted in the detailed decomposition at recursively halving frequency ranges; this allows the observer to view the frequency content of various frequency bands without losing localization in time, as in the fast Fourier transform. Localization in time was important since the fault occurrence was transient. Corrugation response signals were observable within the 125-250Hz range, but expected fault signals in the 500-1000Hz range (a crack) and in the 62.5-125Hz range (track irregularity) were not present. Subsequently, cabin noise was analyzed, concluding that its spectral peak could also be used for corrugation detection, but cabin noise alone did not reveal a distinct difference between healthy and corrugated sections of track. Extensive field testing concluded that railcar noise could be used as a means to detect corrugation faults. In addition, a comparison of track health before and after maintenance revealed a distinct difference in vertical acceleration root-mean-square (RMS) measurements; the overall RMS decreased in sections that were repaired. The study highlighted the effectiveness of different detection techniques and how they can be used to effectively detect railway track faults that are physically different: corrugation was detected in the 125-250Hz range where spectral noise was effectively used to detect it; impulsive cracks in the 500-1000Hz range were detected more effectively in the measurement signal; and, track irregularities were observed in the low frequency level of the wavelet decomposition as well as in the gyroscope measurements. The aforementioned techniques were investigated in the development of Rail-Veyor® fault detection system.

Yang *et al.* [19] proposed a method for fault detection based on flow ripple measurements of hydraulic vane pumps. The study compared the flow ripple from healthy pumps and pumps with artificially induced wear characteristics of common fault conditions, and varied the head pressure of the pumps to simulate variable duty. The goal of the study was to identify inconsistencies in the flow ripple shape, attributed to faulted components.

Figure 2-5 shows the difference in flow ripple due to impeller damage at varying degrees and at different head pressures.

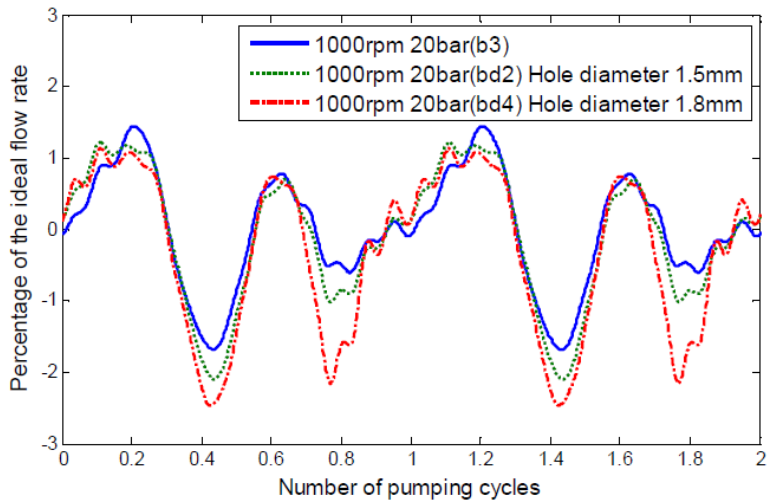


Figure 2-5: The flow ripple measurements with simulated fault progression (taken from [19]).

The results point out that flow ripple measurements are sensitive to changes in mechanical health of a vane pump.

Each of the aforementioned techniques uses a different signal processing method, yet the approach for fault detection relies on separation of some feature metric representing the machine

state from an otherwise noisy signal. The current research aims to develop a low cost hardware solution that benefits from powerful signal processing for fault detection.

2.2.4 Fault Detection and Decision Support

Fault detection is a key area in machine condition monitoring; it is the backbone to the decision support system. In order to detect faults, the system needs to be trained to recognize faults. Similar to the way people discriminate between acquaintances and strangers, the fault detection system is improved by observing the same characteristics in high volume (seeing the same face every day) or by observing outstanding features that describe the data (seeing a very unique individual compared to others). That is, if the training data set is small, or the 'outliers' are not distinctly different in some observable way, the classifier cannot perform effectively. Therefore, the feature sets should be rich with descriptive characteristics that highlight differences between target and outlier classes and the data set has to be of sufficient size.

Decision support is a post-processing task used as a triage system to weight classification results and determine whether the decision support should insist on full system shutdown, or warn the operator that a potential hazard exists [20].

2.3 Railway Anatomy and Failure Modes

This section briefly introduces relevant railway anatomy and some terminology; this is followed by a brief overview of typical failure modes encountered in the railway industry.

2.3.1 Railway Anatomy

Figure 2-6 shows some conventional rail terminology; references to different sections of the rail will correspond to this naming convention.

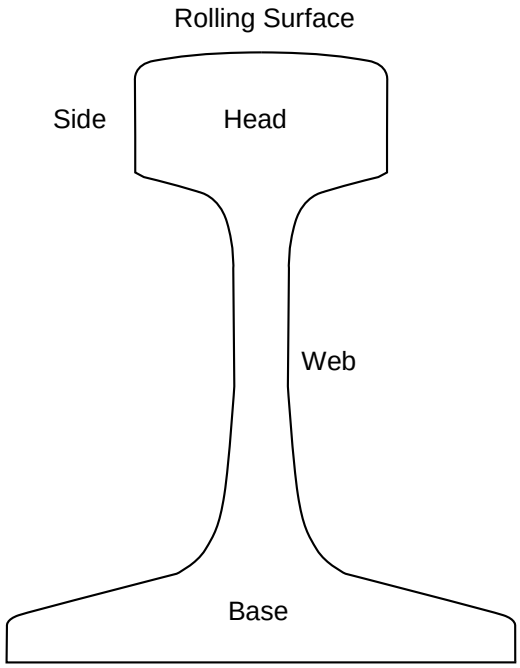


Figure 2-6: Rail Terminology

Figure 2-7 shows reference plane conventions for rail sections; these orientations will be referenced throughout this document.

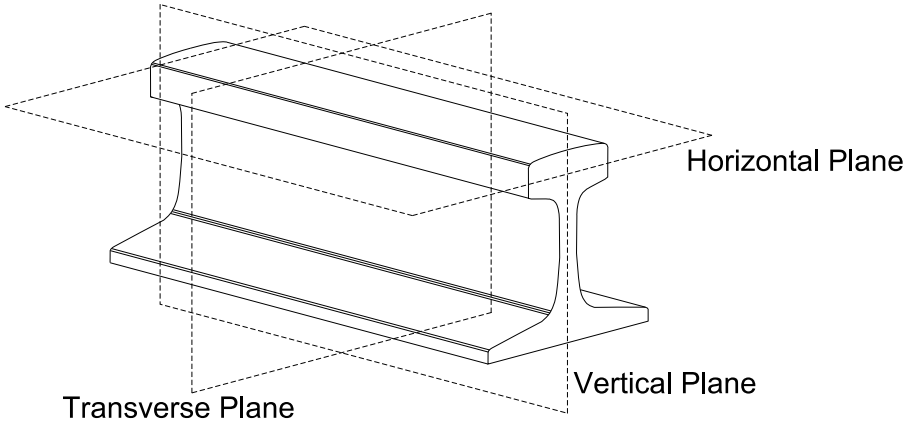


Figure 2-7: Reference Planes

Sections of rail are joined together using fishplates (or joint bars). Fishplates are bolted through the web to tie two sections of rail together. Figure 2-8 shows an installed fishplate.

Figure 2-8: Fishplates Join Rail Sections (taken from [21])

This concludes the section on relevant railway anatomy. The next section deals with railway failure modes.

2.3.2 Railway Failure Modes

Table 2-1 shows railway failure modes of federally regulated main track railway in Canada from 2003 to 2012. During that period, equipment and track related faults accounted for over 80 percent of derailments.

Table 2-1: Transportation Safety Board of Canada Derailment Statistics (taken from [22])

	2003	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
Total number of assigned factors	172	188	227	170	178	151	74	84	108	63
Environmental	5	7	10	8	17	12	6	1	6	2
Equipment	61	70	83	54	60	42	23	26	40	19
Track	67	71	87	64	56	61	29	31	37	25
Actions	26	23	28	21	20	19	8	20	17	14
Other assigned factors	13	17	19	23	25	17	8	6	8	3
Derailments by number of assigned factors	156	160	198	139	159	128	67	80	101	63
One factor assigned	136	140	173	122	148	117	65	77	96	63
More than one factor assigned	13	18	22	15	9	11	2	2	4	0
No factor assigned	7	2	3	2	2	0	0	1	1	0

In addition to this data, in 2001, the Federal Railroad Administration reported 290 derailments due to broken rails, a commonly occurring track fault.

Based on the high occurrence of equipment and track related faults causing derailment, the failure mode and effects analysis has identified a critical area in need of condition monitoring.

The majority of rail failure in heavy haulage railway systems is due to propagation of internal defects from excessive wear and fatigue. Apart from this, the most predominant cause of failure can originate from manufacturing processes, improper operation, or regular wear. Defects at the manufacturing level are becoming less frequent with improved processes and inspection [23]. In terms of improper operation, fault prediction becomes a much more involved task since the operational variation is unknown [24]. A simple shift change to a different operator can significantly increase the duty imposed on the railway system [20].

Railway track is particularly resilient since its natural failure modes are limited to propagation of internal defects, but this resiliency leads to unexpected breakages due to its long life cycle [25]. Therefore, due to the stochastic nature of material metallurgy and varying degrees of manufacturing technology, it is very difficult to make a time-based prediction as to when the railway track will fail [23]. For this reason, automated inspection systems are particularly well-suited for railway condition monitoring. Machinery condition monitoring techniques can be used to develop a structural health monitoring system for railway track to detect faults at the incipient stage, before catastrophic failure occurs.

2.3.3 Railway Track Failure Modes

Common rail failure modes are described in the Railroad Track Maintenance and Safety Standard [26]. Typical faults that have been known to cause derailment are horizontally split heads, head-web separation, vertical split heads, loose or broken fishplates, bolt hole cracks propagating through the head, complete rail breakage in the transverse plane, and railway ballast washout. In heavy haul applications, Cannon *et al.* (2003) illustrated that these faults accounted for over 65% of rail faults; another 25% of rail faults were caused by welding issues or engine burn, neither of which are present in the current study[1] [25].

Horizontally split heads originate from an inclusion that propagates to form a crack under repeated heavy loading. As the defect progresses, rail degeneration causes improper operation of the track system. The location of the initial inclusion can become a constructive feedback zone where the surrounding material weakens due to the initial damage, in turn causing newly formed micro cracks. Crack propagation eventually breaks through the surface of the rail and significant loss of material is apparent. The general appearance of a horizontally split head is illustrated in Figure 2-9 (A). Similarly, head-web separation is due to propagation of internal defects, but originates and propagates between the head and web; however, it can also be the result of a stress concentration at the adjoining fillet. The appearance of head-web separation is shown in Figure 2-9 (B).

[1] The case study of this research uses fishplates to join rail sections, not welds, and there is no prime mover powering the axles of the train, eliminating the possibility of engine burn related faults.

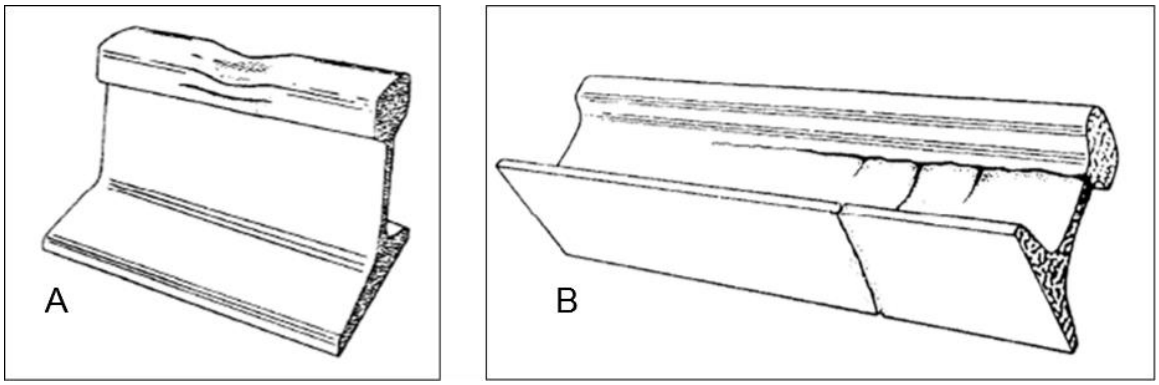


Figure 2-9: The general appearance of a horizontally split head (A), and head-web separation (B) (taken from [26]).

In addition, small surface cracks can shadow larger cracks in automated ultrasonic inspection, masking the larger defect deeper in the rail head.

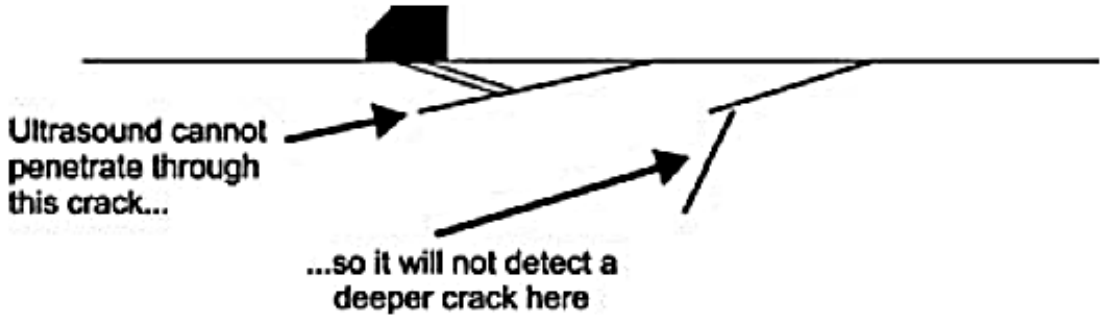


Figure 2-10: Crack shadowing hindering ultrasonic inspection (taken from [25]).

Broken rail is commonly the result of repeated heavy loading; its general appearance is illustrated in Figure 2-11 (A). Bolt hole cracks are shown in Figure 2-11 (B), which are a consequence of stress concentration and tensile loading at the joint due to thermal contraction or freight load, mainly [25].

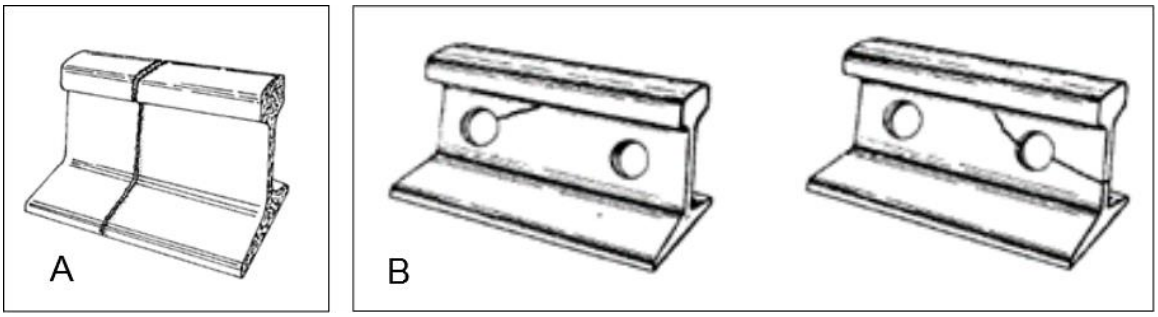


Figure 2-11: The general appearance of a broken rail (A), and bolt hole cracking (B) (taken from [26]).

Washouts and loose fishplates are largely produced from a lack of inspection and maintenance. In the case of the latter, mechanical looseness is to be expected and is easily remedied, but requires attention nonetheless. Washouts ensue where environmental conditions have eroded the ballast of the railway track, or the ballast was inadequate for the freight payloads. This concludes the review of railway failure modes relevant to this study. The following section investigates commercially available condition monitoring solutions for railway track inspection.

2.4 Commercially Available Railway Monitoring Systems

Structural health monitoring is growing increasingly popular as existing structures age. In many cases, the structural integrity of a system remains unknown. For example, railway systems have existed for well over 100 years, with hundreds of thousands of kilometers of track, thus it has not been possible to account for the condition of all railway track. Additionally, the track condition is susceptible to environmental stresses, operational wear, and other factors that diminish the structural integrity of track. For this reason, manual inspection of infrastructure, bridges, railways, and other structures is being undertaken to assess remaining useful life. In some instances, the sheer quantity of components requiring inspection can be overwhelming; in other

instances, it may be hazardous to perform an inspection. Condition monitoring is put in place to provide measurement data throughout the working life of components or systems. This data creates a history of operation such that degradation and breakdown can be extracted from the measurements. There are a variety of commercially available condition monitoring systems that provide decision support for railway systems; a few of these systems are listed below.

Campbell® Scientific is a multinational corporation that provides customized monitoring systems. The system measures rail-bed moisture, vibration, strain and load to provide the user with a strip chart of the sensor signals. The software provides analysis tools as well, which raise flags and alarms. Case studies are available for structural health monitoring, vehicle testing and performance, mining, bridge monitoring, and many other fields, but no literature is available for railway monitoring systems. The uniqueness of the Rail-Veyor® system would require a customized monitoring system in this case anyway. [27]

ESG Solutions monitor micro seismic events at a geotechnical level to ensure critical landmasses are stable. This is a localized monitoring scheme where sensors are located at a potential fault location such as a bridge where large boats pass under. The high potential for rail-bed failure at this location makes it a candidate for this type of monitoring and analysis. For continuous monitoring of railway systems, the shortcomings of this system are evident. [28]

Innowattech is an energy harvesting research and development group who use piezoelectric pads for railway monitoring. They provide data such as speed, train payload, wheel defect presence, and rail track health monitoring. Again, the disadvantage of this system is that it is localized in

space. For global track condition monitoring, transducers would be required along the entire track. [29]

Strukton Rail uses ultrasonic and eddy current sensors to inspect railheads. This inspection method detects incipient faults within the railhead. This method of global fault detection is particularly attractive since it introduces a moving inspection unit. This facilitates global inspection of railway systems without the use of many stationary transducers [30].

Sperry® is a pioneer of non-destructive evaluation of railway systems. Sperry equipment uses mainly ultrasonic and magnetic induction transducers as a means of detection. Digital images are recorded to allow the operator to view the railhead to make a final decision if an alarm is raised.

In addition to commercially available monitoring systems, current research and development is paving the way for improved fault detection systems.

2.5 Research and Developments in Railway Inspection Systems

This section reviews different methods of railway condition monitoring. Each method is briefly defined, then, some advantages and shortcomings of the methods are summarized.

2.5.1 Ultrasonic Inspection

Ultrasonic inspection is a technique that uses a beam of ultrasonic energy emitted into the railhead that is reflected and measured by ultrasonic transducers. Changes in amplitude and angle of the reflection provide information regarding the integrity of the rail section and can

indicate sub-surface cracks. In practice, multiple beams are emitted at varied angles to evaluate the entire rail cross section [24].

Ultrasonic inspection machines are either manual or integrated into an inspection car. Figure 2-12 and Figure 2-13 show a manual inspection walking stick and a high-speed ultrasonic inspection probe as a reliable means of subsurface crack detection.

Figure 2-12: A Manual Ultrasonic Inspection Unit (taken from [23])

Figure 2-13: A High-speed Inspection Unit (taken from [23])

In practice, a great deal of fine-tuning is necessary to produce reliable results. Threshold values, sampling window length, point of measurement, and other factors reduce the efficiency of this technique. Currently, false alarms generally outnumber correct fault classifications [24] making the technique impractical for high-throughput operations such as material haulage.

2.5.2 Pulsed Eddy Current Techniques

With pulsed eddy current techniques, a current is induced to the railway track and the eddy current-induced magnetic field is measured. Variances in the eddy current impedance reflect variances in the rail cross section, which signifies an anomaly [24]. This approach works very well in practice, performing more reliably for near-surface defect detection than ultrasonic techniques [31]. However, the sensing probe is extremely sensitive, so the gap between the sensor and the rail must be held constant for consistent results [24].

2.5.3 Magnetic Induction Inspection

Magnetic induction inspection is used for surface crack detection of in-service rails. A magnetic field is induced into the rail specimen using an electromagnet, whereby search coils measure the resultant magnetic field. Fluctuations in the measured magnetic field are the result of inconsistencies in the rail that denote a potential defect [23].

Magnetic induction inspection methods are particularly good at detecting transverse defects; however, faults in the vertical and horizontal plane often go undetected. Due to the nature of the induced magnetic field, low-speed inspection is preferred. This field strength constraint limits the inspection speed to a maximum of approximately 35 km/h [23].

2.5.4 Image Recognition for Rail Inspection

Image recognition and vision systems perform extremely well for detection of missing bolts, railhead wear, and other surface geometry inconsistencies [24]. These inspection systems can detect a wide variety of surface defects, although the inspection speeds are governed by the severity of the fault, i.e., when inspecting for large cracks or corrugation, speeds can be higher than when inspecting for small cracks or slightly uneven surface geometries. In addition, vision techniques cannot detect sub-surface faults, so the technique cannot replace ultrasonic inspection systems [24].

2.5.5 Vibration-based Condition Monitoring for Rail Inspection

Vibration is a typical phenomenon measured for fault detection systems as it is tied to the mechanical connectedness of the machinery. In the healthy state, machines operate with a set of natural frequencies that describe the system connectedness and response to external excitation

[4]. The propagation of inclusions or defects leading to cracks and unhealthy operation alter the underlying mechanical connectedness of the system. This change in system connectedness causes an intrinsic variation in the set of natural frequencies of that system. This affects the vibration response, which is indicative of the fault propagation. For these reasons, vibration signals were monitored for the development of a fault detection system.

Vibration-based condition monitoring has been used to detect faults in rotating machinery such as helicopter gearboxes, motor shaft bearings, industrial ventilation systems, and pumping systems. The frequency spectrum of a vibration signal can reveal energy variations around the rotating frequency that are representative of faulted components. That is, a healthy bearing might show an energy spike at the rotating frequency, while a bearing with an inner race fault might show energy spikes at the rotating frequency and the ball pass frequency of the inner race. This has been examined in detail and proven to be an effective method of detecting bearing faults in steadily operating machinery. Vibration-based fault detection for unsteadily operating machinery has more recently been under investigation for use with variable speed and load conditions [32]. In the case of rail-based systems, the periodicity of events is unique to the track infrastructure. This leads to the development of a condition monitoring system using a small data set, and the focus of this study.

Table 2-2 shows a comparison of the aforementioned fault detection techniques with respect to some advantages and disadvantages in the railway industry.

Table 2-2: A comparison of fault detection techniques for railway track.

	Disadvantages		Advantages		
Technique	Impedes Production	Time Consuming	Global Detection	Surface Detection	Sub-surface Detection
Ultrasonic - Handheld		X	X		X
Ultrasonic - Automated	X		X		X
Magnetic Induction		X	X	X	
Pulsed Eddy Current		X	X	X	
Image Recognition	X		X	X	
Vibration			X	X	

The comparison of fault detection techniques in Table 2-2 emphasizes the main drawbacks in advanced inspection techniques for the railway industry. Although these techniques are suitable for some railway infrastructure, the production-oriented nature of the mining industry cannot allow for impeded production. This constraint has introduced the opportunity for development of a novel approach to the task of railway condition monitoring.

Chapter 3

3 Field Experiments and Data Collection

In the previous chapters, the difficulties associated with railway condition monitoring of production-oriented systems was presented. This chapter defines the experimental approach taken to develop the railway maintenance decision support system and the reasoning for such an approach.

The data acquisition (DAQ) platform used to gather field data is discussed in section 3.2, followed by a short overview of the data acquisition application in section 3.3. Two sets of experiments were designed in the development of the fault detection system. The preliminary experiments (described in section 3.4) were performed to study the general system response and provide a basis for subsequent experimental design. The subsequent experiments (described in section 3.5) were then conducted to gather condition monitoring data for the signal processing portion of the fault detection system.

3.1 Overview

In order to conduct research in the field of condition monitoring, data representing the system response are required. Data collected during healthy operation and failed states are used for the training and testing of fault detection algorithms. It is often difficult and impractical to gather fault data for systems that are already in production since it requires the presence of damage; since the presence of damage reduces performance and production rates, and also can be hazardous to workers or the environment, it is very rare that an industrial partner is willing to operate machinery in the damaged state for the purpose of investigation.

Fault detection systems are generally applied in situations where downtime is extremely costly and catastrophic failure is unacceptable. For example, a passenger train in a high-traffic city could benefit from a fault detection system by reducing downtime. Equivalently, a fault detection system could be used to monitor the mechanical health of a helicopter ambulance's gearbox, which is needed for safeguarding life. In both cases, removing the machinery from operation during use or production to induce faults for the purposes of conducting experiments would be impractical. The Rail-Veyor® system used in this study provided a unique opportunity to instrument and seed faults in an industrially representative system for the purpose of condition monitoring research.

The first task of the data collection stage was to instrument the system. Once the instrumentation was mounted, the train was run across the same section of track to gather baseline data representing fault free, normal operation of the system. Subsequently, faulted components were introduced into the track system and additional data was gathered. The data would then be used for development, testing, and validation of the fault detection system described in this research.

Two rounds of data collection trials were conducted; the first round was a preliminary analysis of different failure modes at various operational conditions. After the first round of experiments, the field data were manually inspected to provide insight regarding the general system response (i.e., were the accelerometers operating within their desirable operating limits; were the train speed measurements in agreement with speed set points; was there an observable difference in system response due to failure modes versus healthy modes, and etcetera). The second set of tests were designed based on conclusions from the analysis of the preliminary experiments, in which fewer failure modes were investigated at fixed operating conditions.

3.2 Sensors and Instrumentation

In order to collect data representing the vibratory response of the Rail-Veyor® system as it passed over the tracks, it was necessary to instrument the train with a high-resolution vibration data acquisition system. Digital inputs and outputs were also used to capture train speed measurements, dead reckoning reference signals, and as indicators of program status. The system also needed to be battery powered since the train had no onboard power supply, and needed to be able to tolerate shock and vibration due to the operating conditions. The following section describes the various hardware used for instrumentation of the Rail-Veyor® system.

3.2.1 The DAQ Enclosure and Electrical Connectivity

An enclosure was necessary to protect the electronics from dust and impact due to the environmental conditions. The enclosure housed two lead-acid batteries to power the data acquisition platform and sensors. The enclosure also contained the data acquisition platform and external storage. Sensor wiring was routed from the DAQ platform to panel-mount BNC connectors for external connection. A master switch inside the enclosure was used to turn the system on, and another switch external to the case was used to start and stop the data logging process. The DAQ setup is shown in Figure 3-1 below.

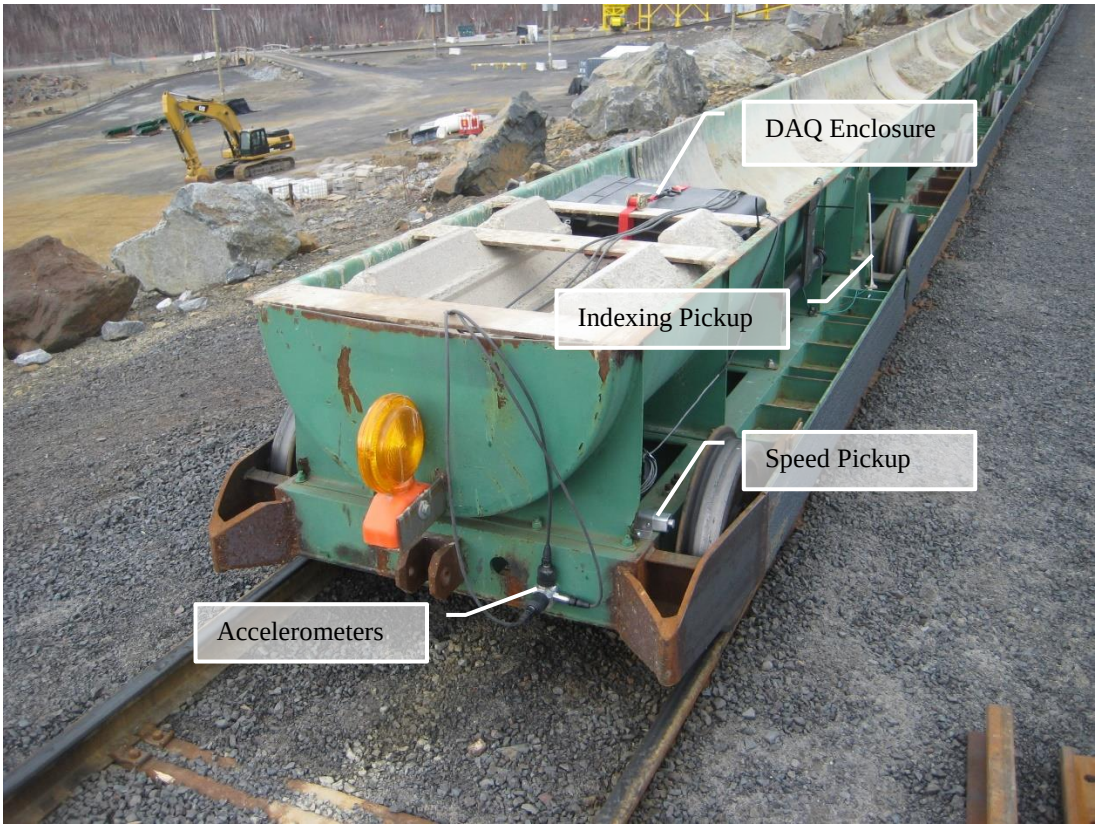


Figure 3-1: The front car of the train instrumented with the DAQ.

3.2.2 Data Acquisition Platform

In order to meet the aforementioned requirements for a high-resolution robust data acquisition system, a solid-state embedded platform was selected. The data acquisition system consisted of a National Instruments (NI) CompactRIO (cRIO) equipped with a real-time controller (NI cRIO 9022) and field programmable gate array (FPGA) chassis (NI cRIO 9113). The cRIO uses NI LabVIEW for software development for the real-time controller and FPGA chassis. The CompactRIO is a modular device that can accommodate an extensive list of instrumentation and is configurable for different timing, synchronization, signal processing, and antialiasing. This system is designed to tolerate harsh industrial environments, including shock and impact, as well

as high temperature. Furthermore, its FPGA chassis offers flexibility as well as high throughput, ideal for condition monitoring research. The NI cRIO is shown in Figure 3-2 below.

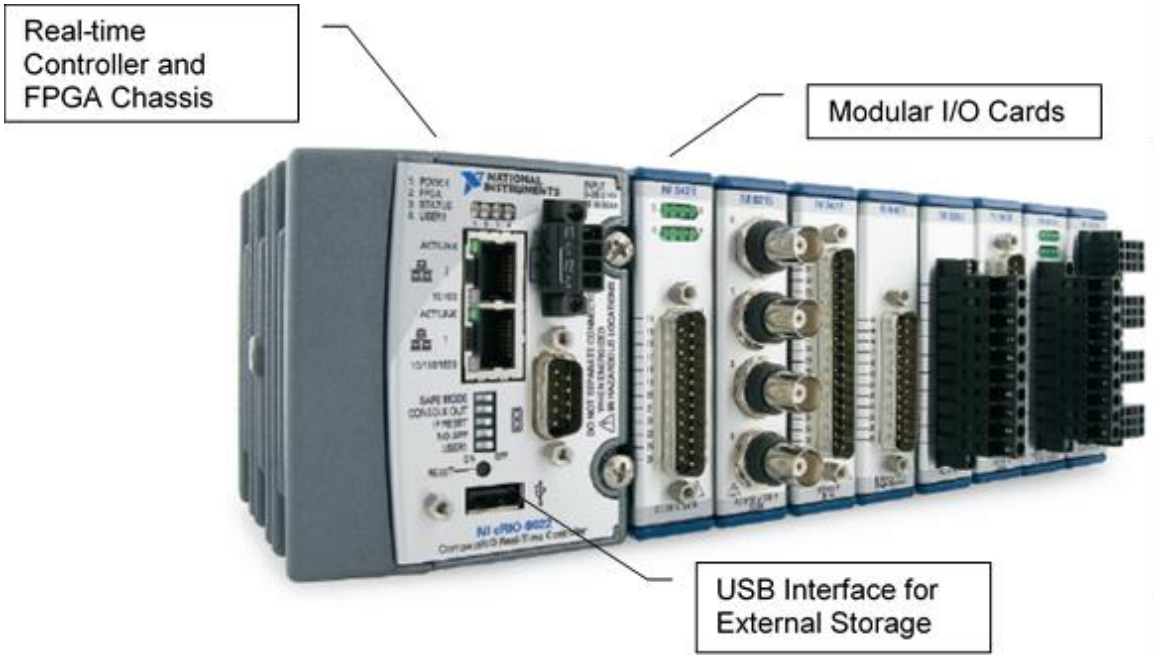


Figure 3-2: The NI CompactRIO Embedded Control and Acquisition Platform (adapted from [33])

3.2.3 Platform Modules

For accelerometer interfacing, a 4-channel data acquisition module (NI 9324) was selected, equipped with integrated electronic piezoelectric (IEPE) enabled selectable current excitation, and built-in anti-aliasing filtering. The 9234 module is capable of sampling at rates up to 51.2 kS/s at 24-bit, fixed resolution, per channel, simultaneously. It has a bandwidth of 23.04 kHz and operates between -5 V to 5 V. The module is stable between -40 °C to 70 °C during operation.

The TTL pulse signals for speed and indexing, as well as control to start and stop data logging required digital I/O channels; an NI 9402 digital I/O module was used to accomplish this task. The module has four bidirectional digital channels, capable of maximum sampling rates of 20 MHz.

3.2.4 Accelerometers

IEPE accelerometers were used for sensing vibration in this study. Their function can be thought of as a simple mass-spring system where the spring is a piezoelectric crystal. The movement of the transducer transmits through the mass, causing deflection, and strain of the piezo-crystal. The induced voltage from the piezoelectric force is then amplified through the integrated circuitry contained within the accelerometer, and a two-wire conductor transmits this signal to the platform interface modules. Other configurations of accelerometer exist, such as the ceramic shear accelerometers used in this study, yet the basic functions are the same.

Two models of accelerometers were used to measure the vibration response of the train. The PCB accelerometer model 603C01 has a sensitivity of 10.2 $mV \cdot s^2/m$ (100 mV/g), and measurement range of ±490 m/s^2 (± 50 g); the PCB accelerometer model 626B03 has a sensitivity of 102 $mV \cdot s^2/m$ (1000 mV/g), and measurement range of ±49 m/s^2 (± 5 g).

The accelerometers were oriented in a tri-axial configuration illustrated in Figure 3-3. The advantage of using single-axis accelerometers fastened to a cubic block is the modular capability of swapping out one accelerometer for a different measurement range if necessary.

Figure 3-3: Orthogonally oriented accelerometers mounted to the front railcar.

During the design stage of the experimental apparatus, definitive system response was unknown, so the modular capabilities of this setup allowed for sensor exchange during the mode classification experiments to accommodate system responses that fell outside the sensor measurement ranges. In fact, the setup shown in Figure 3-3 depicts three 10.2 mV·s^2/m (100 mV/g) accelerometers, which were subsequently changed to three 102 mV·s^2/m (1000 mV/g) accelerometers and one 10.2 mV·s^2/m (100 mV/g) accelerometer in the transverse direction. The accelerometers were mounted to the chassis of the railcar, which has no suspension system; this provided a strong vibration signal transmitted from the front axle of the railcar.

The adaptation of the new configuration allowed the signal to operate across more of the measurement range, without signs of transducer saturation[2]. Comparing Figure 3-4 and Figure 3-5, it is evident that the general shape of the signal is similar; however, the peak-to-peak amplitude is roughly 0.4 V for the 10.2 mV·s^2/m (100 mV/g) accelerometer, compared to nearly 6 V for the 102 mV·s^2/m (1000 mV/g) accelerometers. In the first case, operating at only 4% of the full-scale range reduces resolution to roughly 1-bit, while in the second case, operating at nearly 60% full-scale range, nearly 15-bit resolution is achieved across the same signal.

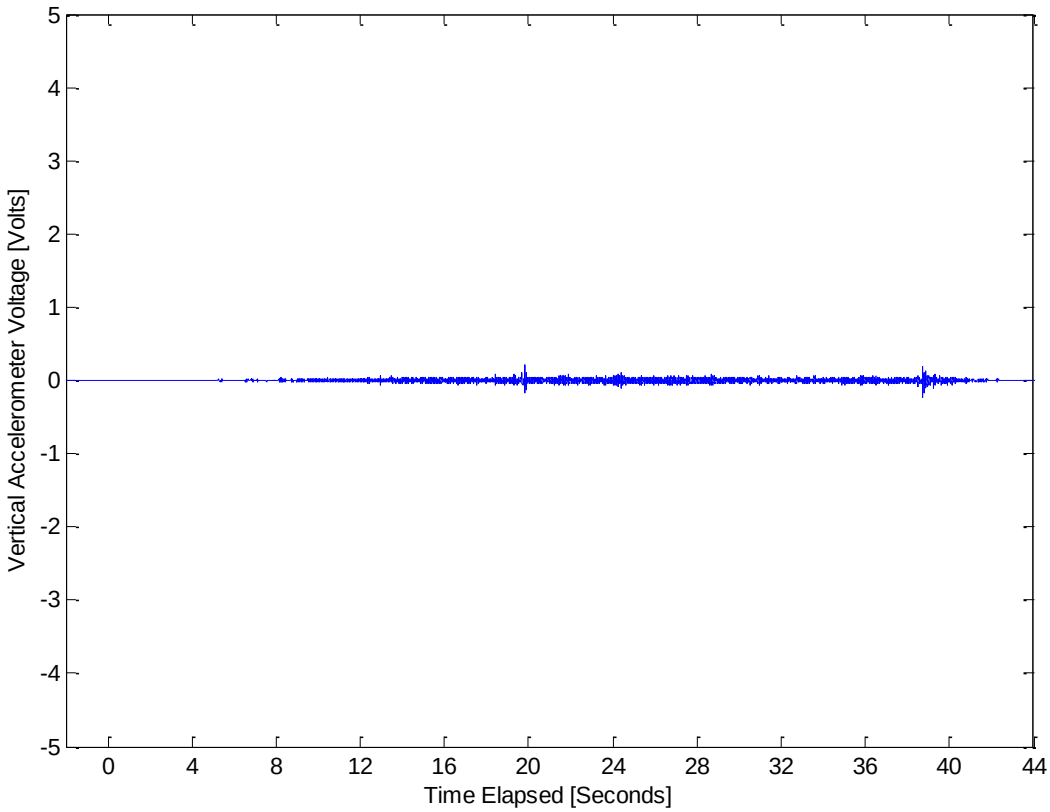


Figure 3-4: Vertical 10.2 mV·s^2/m (100 mV/g) accelerometer voltage output during healthy operation.

[2] Since the quantization resolution was fixed, having signals span more of the measurement range of the sensors makes better use of the available resolution.

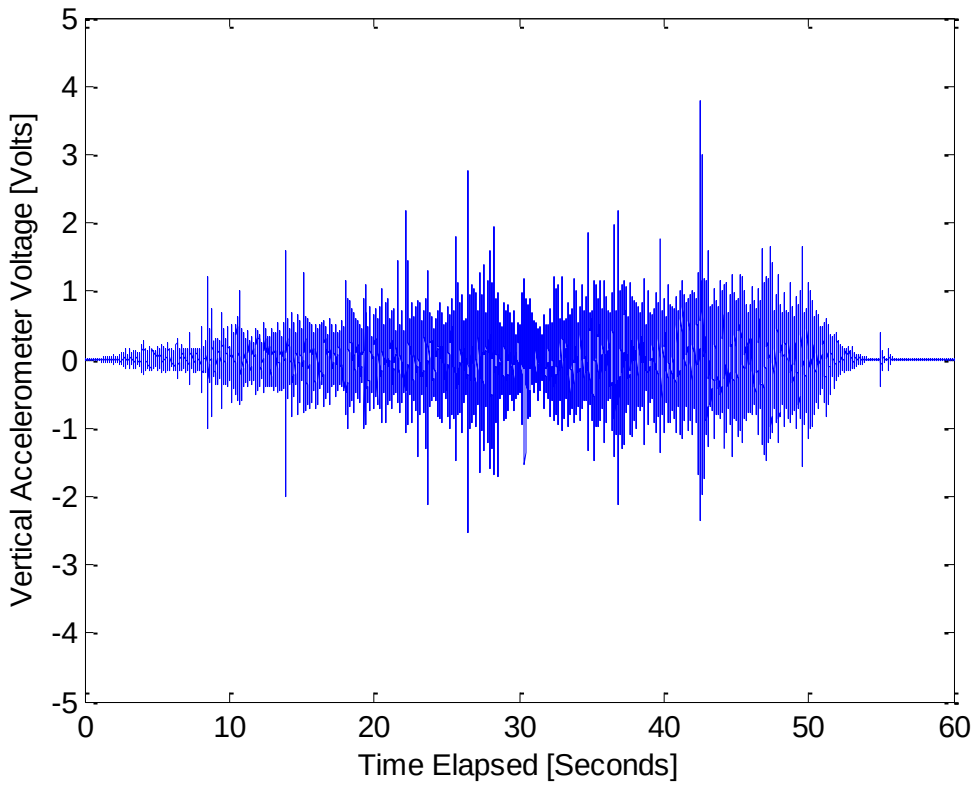


Figure 3-5: Vertical 102 mV·s^2/m (1000 mV/g) accelerometer voltage output during healthy operation.

3.2.5 Speed and Indexing Measurements

For the sake of simplicity, the same transistor-transistor logic (TTL) magnetic pickup was used for both wheel speed measurement and for indexing A, B, and fault locations. The logic magnetic pickup is a powered pickup that outputs a 0 V or 5 V TTL signal. While a magnet is present in the sensing range, the signal reads 5 V, and in the absence of a magnetic field, the signal reads 0V.

Figure 3-6 shows the magnet placement on the wheel flange for speed measurement; the irregular gap spacing is used to identify a complete wheel rotation.

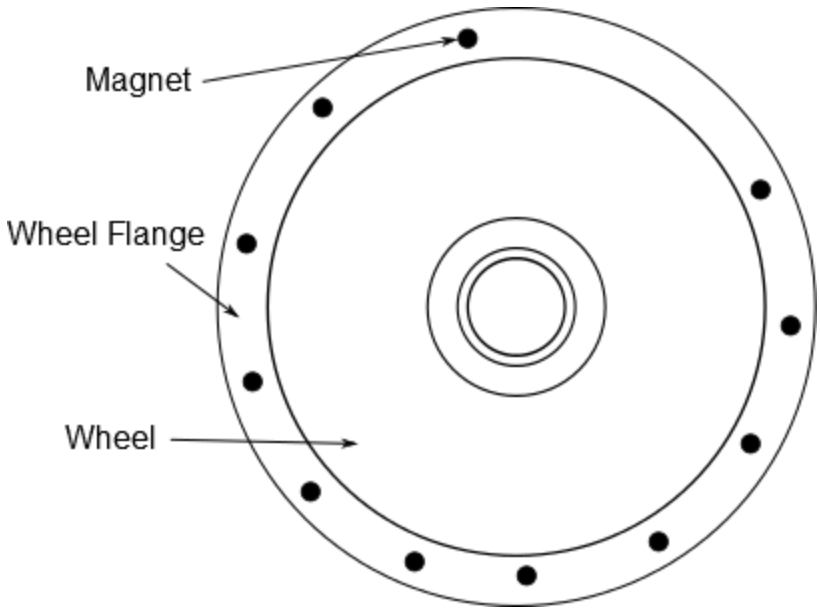


Figure 3-6: Magnet configuration on train wheel for speed measurement.

The magnetic pickup was mounted rigidly to the railcar chassis such that as the wheel rotates, the magnets pass by the sensor, creating a pulse train that could be transformed into a speed signal. The physical setup is shown in Figure 3-7 below.

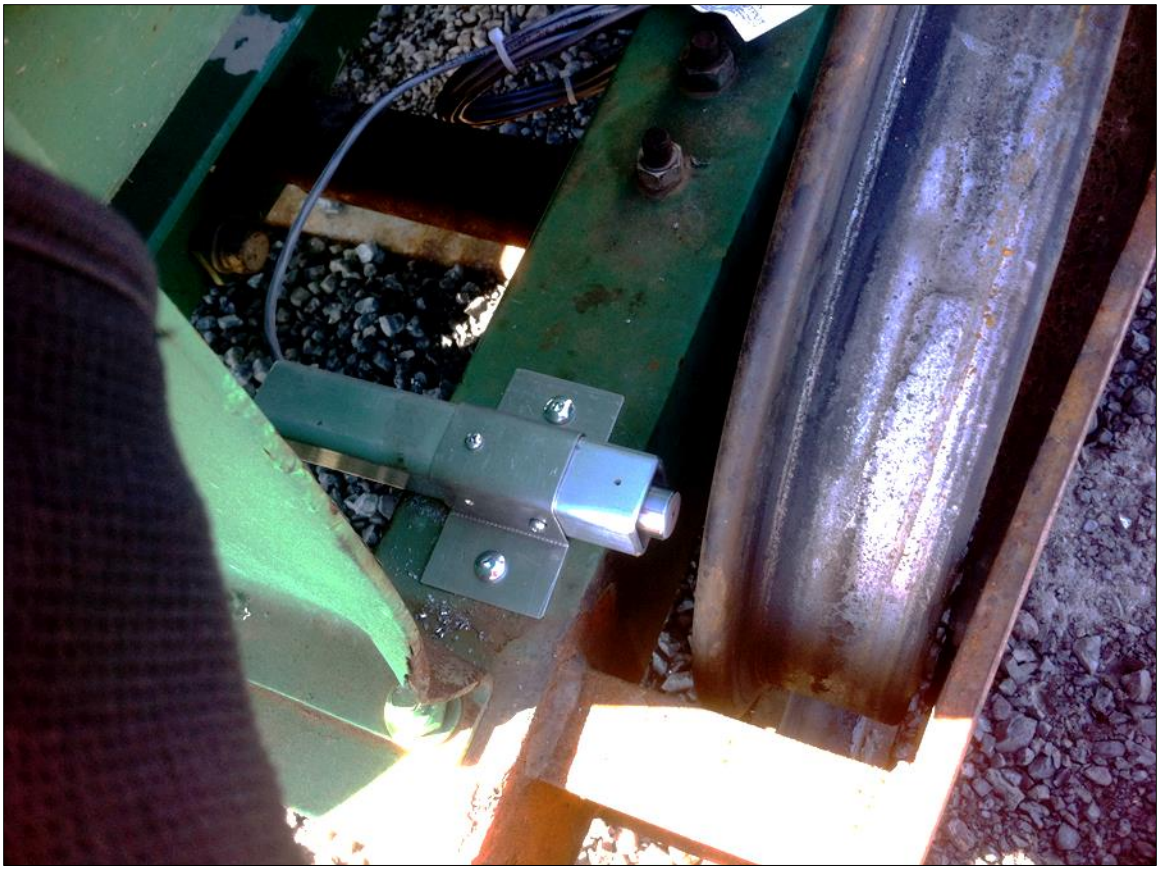


Figure 3-7: The speed transducer setup is shown above; the magnetic pickup is aimed at the wheel flange, which has magnets around its perimeter.

The indexing setup was designed in the same manner as the wheel speed sensing method. That is, a magnetic pickup was fixed to the railcar chassis, and magnets were mounted on indexing arms located roughly 20 m before the fault, at the fault location, and 20 m after the fault. With this configuration, three locations within the signal were known, while other locations could be determined using the speed signal and wheel perimeter. For the purpose of this study, direction was recorded manually since it was unnecessary to integrate it into the experimental apparatus for the study.

3.3 Data Acquisition Application

The control program for the DAQ system was written in NI LabVIEW, a graphical programming language. An external toggle was used to start and stop data logging and store the measurement data to an external hard drive. The program automatically labelled files by the controller clock timestamp. Four accelerometer channels and two digital input channels were sampled at 5 kHz and written to technical data management streaming (TDMS) files with associated metadata. The build-in anti-aliasing feature of the acquisition card ensures that frequencies above 5 kHz did not affect the measurement data.

Since the cRIO has no graphical interface, an indicator light was connected to the remaining digital I/O channel for visual feedback regarding the program state. Upon system power-up, the program was set to run; when the program was ready for data logging (typically about 10 seconds), the indicator would illuminate. While data logging, the indicator would remain off, and following termination of the data logging, the indicator would remain off until the data file was properly defragmented and saved before illuminating the indicator again, signifying that the program was ready to begin data logging again.

3.4 Preliminary Field Experiments

Relative to conventionally-established railway systems, the Rail-Veyor® system concept is a new design with little service history. Therefore, the overall behaviour of the system was examined as the focus of the preliminary experiments. In addition, the preliminary experiments investigated mode classification for the subsequent experimental design.

Preliminary testing was conducted at the Rail-Veyor® demonstration site over a one-week period in which 106 test runs were performed.

3.4.1 Methodology

To develop an understanding of the system response to variable duty cycles, the experiments were conducted over a range of operating conditions. The experiments were run under no payload in healthy conditions at 1, 2, 3, 4, and 5 m/s; these situations would arise following the unloading stage and with the train en-route to the loading station. Next, the train was loaded with approximately 10 tonnes of crushed rock distributed evenly over 20 railcars, and the system response was recorded again at the various speeds in the healthy state; this simulates the situation where the train is fully loaded and is en-route to the unloading station. An indexing apparatus was used for dead reckoning of the proposed fault location. The proposed fault location was situated along a relatively level, straight section of track approximately 20 m (66 ft.) from the next drive station. Figure 3-8 shows the general arrangement of the experiment. The aim was to develop a dataset representing healthy baseline data for comparison to fault data. To be consistent, the healthy baseline data was always collected prior to any disturbances to the system; this ensured that the healthy baseline data did not contain anomalies due to the track disturbances caused by removal and replacement of faulted track sections.

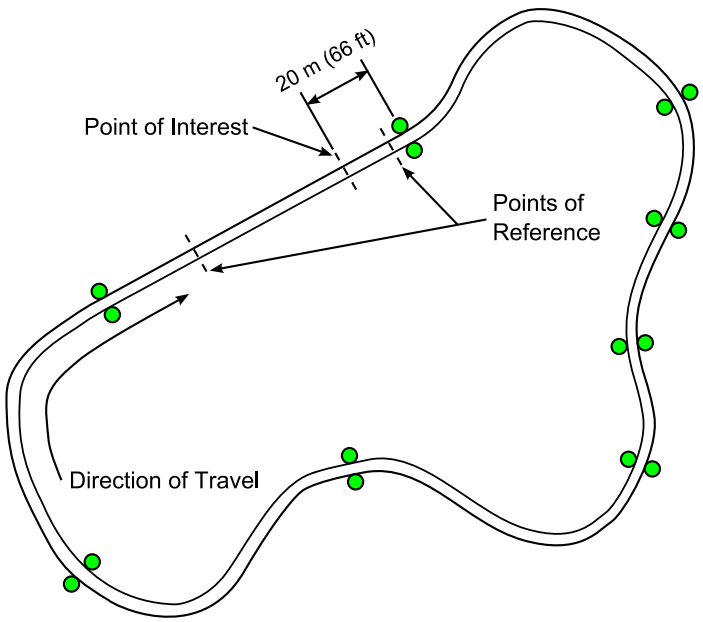


Figure 3-8: The General Arrangement of Field Experiments at the Rail-Veyor® Test Site.

Once a representative set of healthy operating data was collected, data was recorded in the various faulted states. An eight-foot section of healthy rail, centred at the proposed fault location, was removed and replaced with rail that was induced with one of the intentionally seeded faults outlined in Table 3-1 below. Care was taken to disrupt the section as little as possible to ensure that the only changes in the system response were due to the fault and not the experimental process. The process is depicted in Figure 3-9 below.

Figure 3-9: A healthy section of rail being removed to introduce a faulted section (shown on the right).

The fault conditions under investigation for the preliminary experiments are listed in Table 3-1 below.

Table 3-1: List of faults investigated for preliminary tests.

Fault Name	Probable Cause of Fault	Length of Crack
Bolt Hole Crack	Stress concentration	25 mm
Washout	Unstable ballast	N/A
Washout with Crack in Base	Unstable ballast	50 mm
Horizontally Split Head	Inclusion propagation	300 mm
No Fishplate/ Broken Rail	Mechanical loosening/ Inclusion propagation	100%

Since there was very limited historical data of past failures of the Rail-Veyor® track system, the faults under investigation in this study were based on a review of the Railroad Track Maintenance and Safety Standards [26] (as described in section 2.3.2). The defect severity was selected based on the maximum allowable defect size for active passing track, yard, or holding track with speeds less than 4.5 m/s. Furthermore, defects caused by the train prime mover, such as engine burn, were omitted since the Rail-Veyor® train propulsion is passive.

3.4.2 Observations

Figure 3-10 describes a time series vibration response of the train as it ran through one test run. The figure shows the acceleration increasing relatively linearly as the train ramps up to steady speed. This is followed by the middle section of the plot where the train is at constant speed, in which impulsive events related to the track geometry are present. Two significant events occur near sample 4.75×10^5 and sample 5.5×10^5; these are associated with the train meshing with the drive station and passing over a vehicle crossing area, respectively. This is in agreement with

dead reckoning results established from the speed and index signals. Next, the time series vibration response decays as the train slows to a stop.

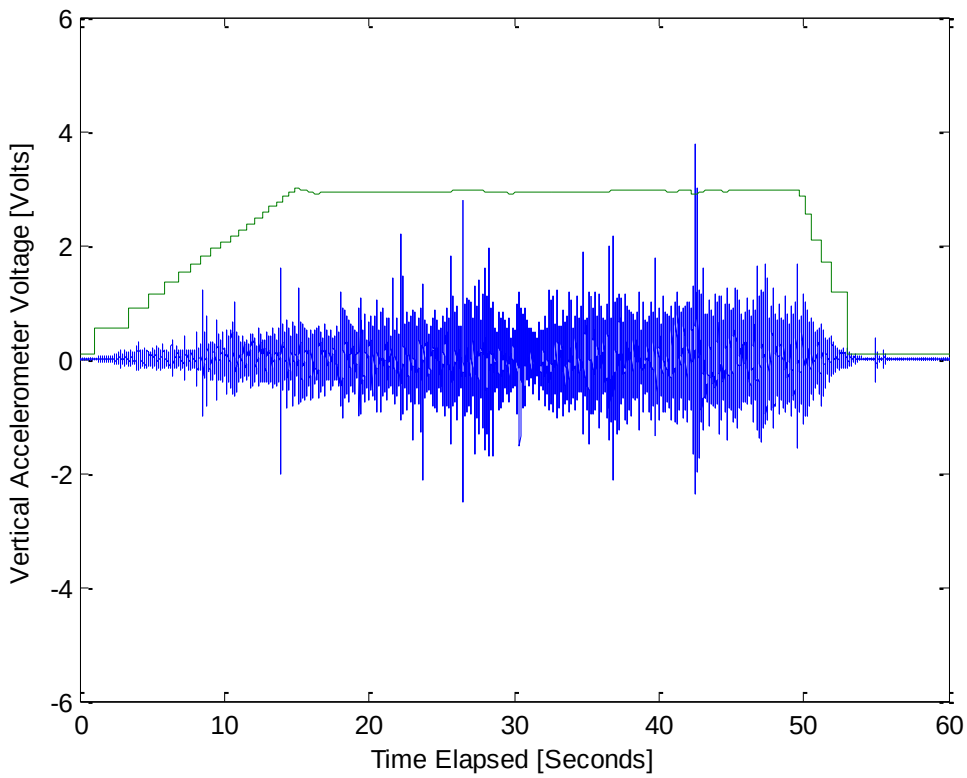


Figure 3-10: A time series vibration plot of vertical acceleration for one test run; train speed is marked by the green line following the same vertical scale in m/s. The steady state speed is 3 m/s.

Initial observations of the transducer signals presented a series of implications that shaped the design of subsequent experiments. Because there was no preliminary vibration data on the rail system, the suitability of the accelerometers needed to be verified. To begin, accelerometer output levels were roughly only 4% of the full-scale measurement range. After an initial test, the 100 mV/g accelerometers were exchanged with higher-sensitivity (1 V/g) accelerometers that better suited the application. Next, subtle faults such as bolt hole cracks were overwhelmed by operational and environmental noise. Examining Figure 3-11, it appears as though impulsive

behaviour is exhibited near A and C; evidence of anomalous activity at the fault location (B) is absent, or at least not obvious.

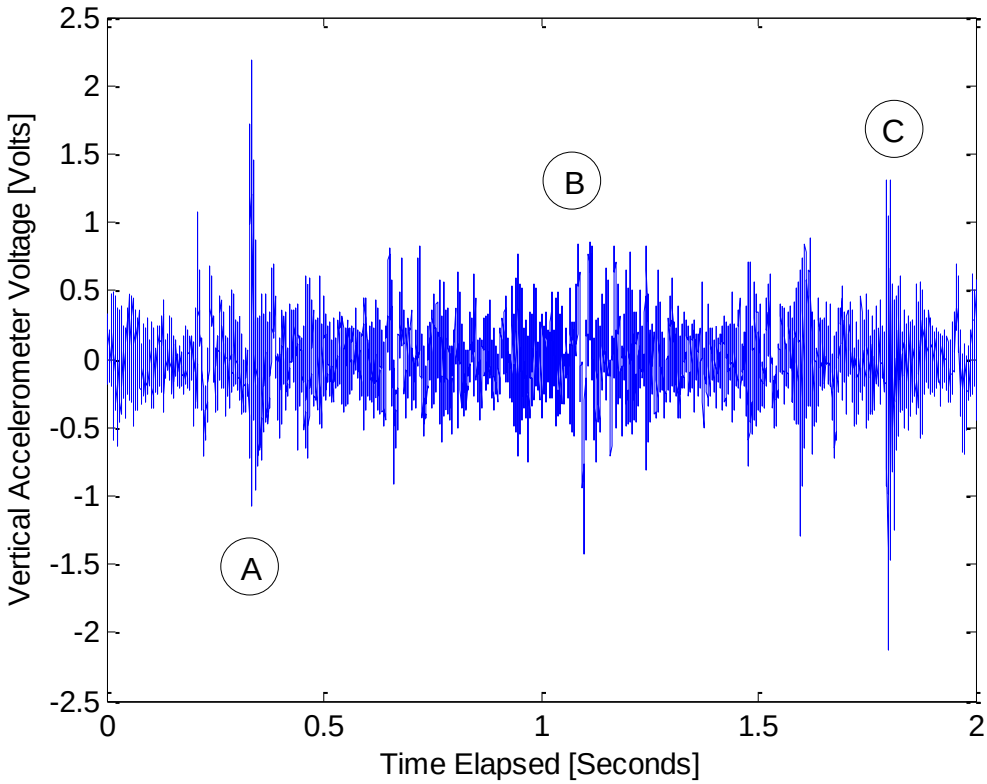


Figure 3-11: Vertical accelerometer signal centered at bolt hole fault location.

Comparing the time series vibration plot in Figure 3-11 to dead reckoning results from the speed and indexing signals, the location of the impulsive events along the track was determined. Using the train speed measurement and sensor sampling rate, the location on the track corresponding to points A and C were calculated using Equation 3-1 below (the location of B is defined by the indexing sensor).

$$\text{Number of samples per metre} = \frac{Sampling\ Rate}{Train\ Speed} = 5000\,\frac{samples}{s}\cdot\frac{s}{3\,m} = \frac{5000}{3}\,\frac{samples}{m} \qquad \text{Equation 3-1}$$

The relation between number of samples and distance travelled was simplified since the train speed was constant.

Figure 3-12 shows a plan view of the preliminary experimental setup where (B) denotes the expected fault location and (A) and (C) are joints in the railway track. From Figure 3-11, using approximate train speed and time elapsed between A and B, the displacement was approximately 1.9 metres; this distance coincides with the joint in the rail section marked in Figure 3-12 (A). Similarly, the distance from B to C in Figure 3-11 corresponded to 2.4 metres; this distance aligns closely to the rail joint labelled as (C) in Figure 3-12.

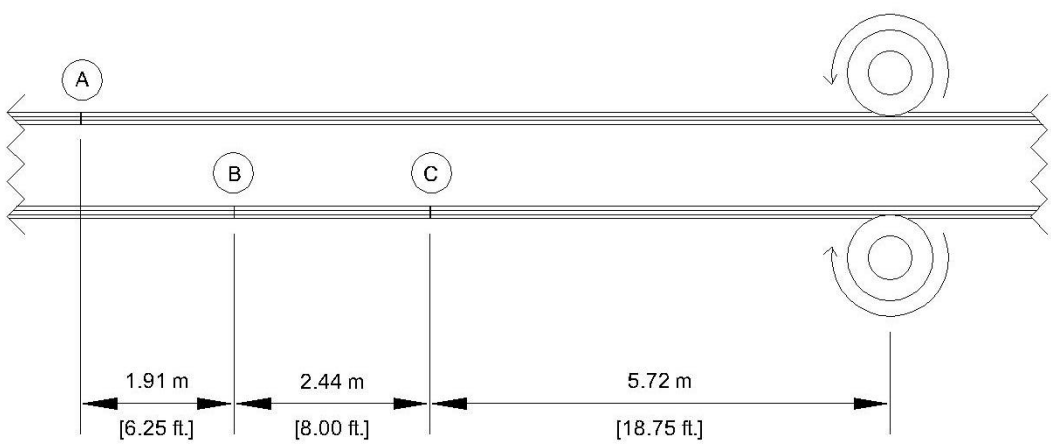


Figure 3-12: A Plan View of the Track Geometry

The fact that joint (A) and (C) appeared in the vibration response, but joint (B) did not, initiated further investigation into the track setup. After investigating the track, it was observed that joints (A) and (C) were separated by a gap of nearly 5 mm; the two sections of rail joined at (B) were adjacent to each other. Therefore, the initial field experiments have exemplified some of the difficulties associated with variation in track geometry.

In contrast, comparatively severe faults such as the no fishplate or broken rail condition introduced highly impulsive shock loading to the system, which saturated the transducers. These effects can be seen in Figure 3-13 at the fault location (B). It is also noteworthy to mention the transient event near (C); dead reckoning based on speed and indexing signals determined the event to be the train passing over the same significant joint gap as observed in Figure 3-11; however, shock resonating from subsequent cars travelling over the broken rail also aligns with (C) since the railcars are also 2.44 m from axle to axle. Due to the high shock loading conditions in this failure mode, it is more convincing that the acceleration transients are due to shock resonance from subsequent cars rather than the 5 mm gap also presents at (C). The subsequent experiments used a 3 m fault specimen to avoid having axle lengths coinciding with rail section lengths.

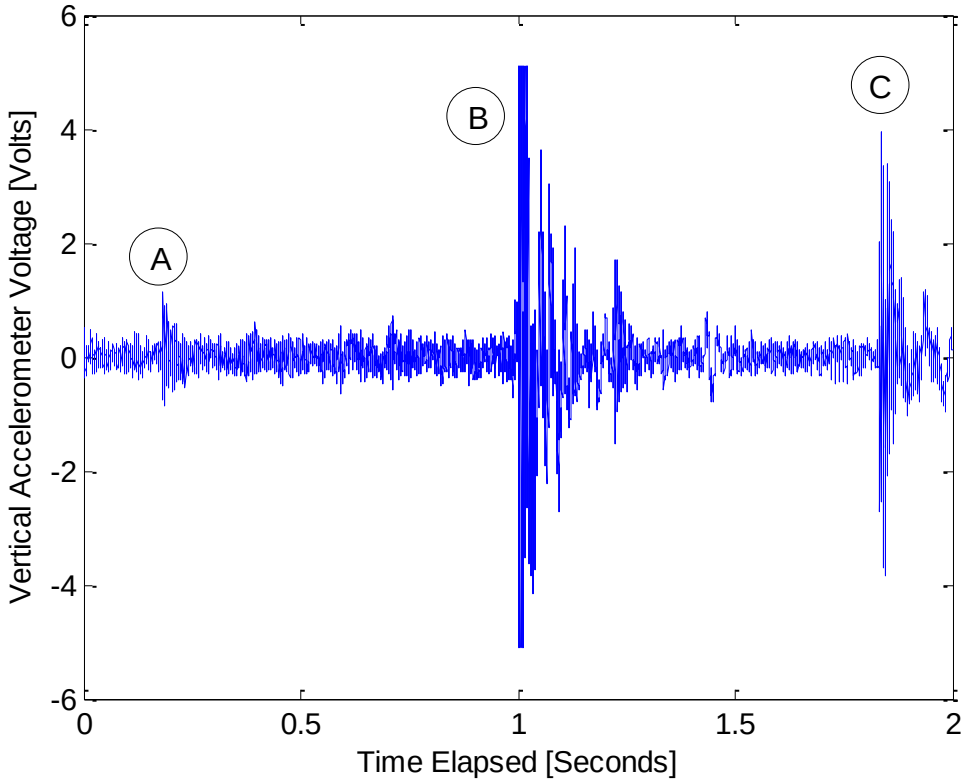


Figure 3-13: Vertical accelerometer signal centered at no fishplate/broken rail fault location.

Since the vibration response from the bolt hole crack (low severity) could not be observed by visual inspection, and the response from the broken rail (high severity) exceeded the measurement range of the transducer, failure modes with a medium level of severity were selected for the second round of experiments. In addition, a 10.2 mV·s^2/m (100 mV/g) and 102 mV·s^2/m (1 V/g) were installed in the vertical orientation to measure high and low magnitude vibrations, respectively; a 102 mV·s^2/m (1 V/g) accelerometer was also oriented along the direction of motion of the train and in the transverse direction to provide tri-axial measurement.

Due to availability and time constraints of the Rail-Veyor® demonstration system, one speed of operation was selected in order to record enough data to train and validate the fault detection

system. A standard operational speed of 3 m/s was selected for subsequent testing. Variations in the train payload had little observable effect on the vibration response, so various loading conditions were not considered for future experiments.

3.5 Field Experiment Data Collection

Following the preliminary analysis of the dataset collected from of the first set of experiments, a revised experimental test plan was developed to construct a larger data set for the development of the fault detection system. In the scope of work, the train would be run at 3 m/s for 50 repetitions in each state of track health. The reasoning was primarily due to time required to perform a test run; in order to have enough data to make the study statistically relevant, more data in one configuration was collected, rather than small data sets in various speed and loading conditions. The subsequent experiments were performed over a one-week period at the Rail-Veyor® demonstration site in Sudbury, Ontario, Canada.

3.5.1 Methodology

The methodology for the subsequent set of experiments was similar to those of the opening experiments. A new section of track was selected upon the conditions that it was undisturbed, in healthy condition, and along a straight path. Prior to any railbed disturbances, data was recorded in the healthy state. Following the baseline data collection, track sections were replaced with rail sections induced with seeded faults. Modifications to the experimental design were attributed to knowledge gained from the preliminary experiments. The new set of faults under investigation for subsequent testing is shown in Table 3-2 below.

Table 3-2: List of faults investigated for subsequent tests.

Fault Name	Probable Cause of Fault	Length of Crack
Washout with Crack in Base	Unstable ballast	30 mm
Washout with Crack in Base	Unstable ballast	60 mm
Loose Fishplate	Improper Fastening/Vibration	N/A
Horizontally Split Head	Inclusion propagation	300 mm

Figure 3-14 shows the experimental setup of the slight washout with a 30 mm crack in the base. The washout was approximately 75 mm deep and spanned approximately 1.25 m.

Figure 3-15 shows the washout section spanning approximately 2.5 m; the washout depth was approximately 75 mm and the simulated crack in the base of the rail was 60 mm.

Figure 3-16 illustrates the size of the simulated crack induced to create the horizontally split head. The simulated crack had a length and thickness of approximately 400 mm and 5 mm, respectively.

Figure 3-14: The experimental setup of the slight washout with 30 mm crack in the base.

Figure 3-15: The expermental setup with the drastic washout and 60 mm crack in the base.

Figure 3-16: The horizontally split head fault used in this study. The simulated crack had a length of approximately 400 mm.

3.5.2 Preliminary Observations from Collected Data

The second set of data was significantly improved in comparison to the first set. In particular, signals occupied the measurement range without saturation, and in the case of more severe faults such as the horizontally split head, the response over the fault can be easily differentiated when compared to the healthy region immediately before the fault, as illustrated in Figure 3-17.

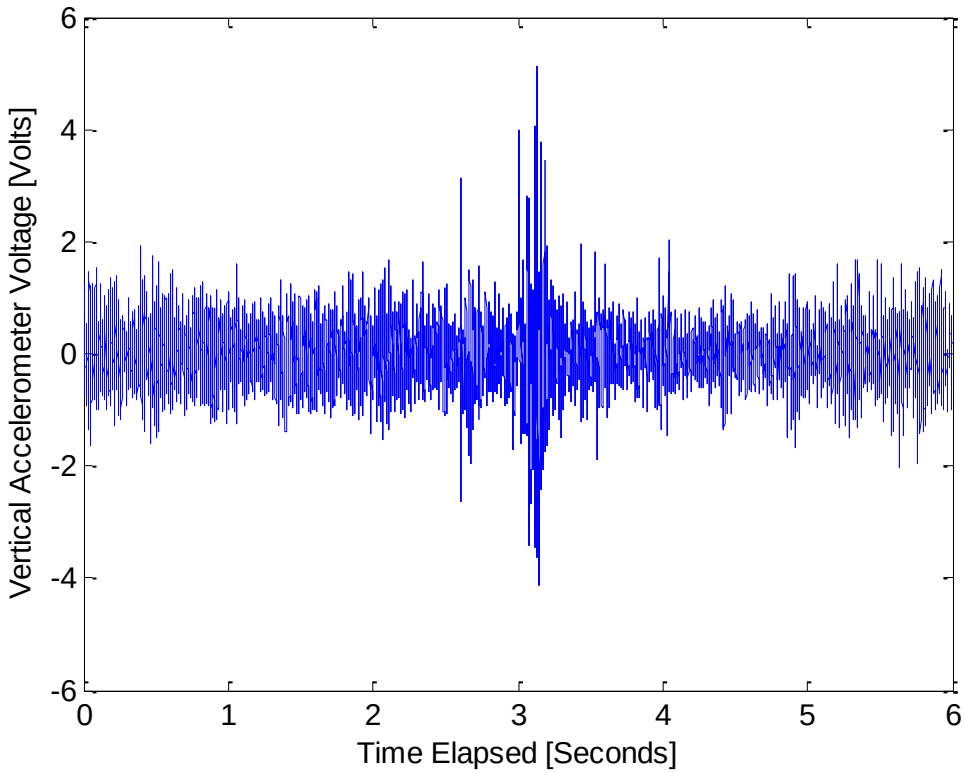


Figure 3-17: The vibration signature exposes the horizontally split head at sample 1.5×10^4.

Although the fault signal can be easily identified in the horizontally split head fault condition, less drastic faults such as a loose fishplate suggest that fault detection of incipient faults is non-trivial. A snapshot of the transducer response for the loose fishplate condition is shown in Figure 3-18.

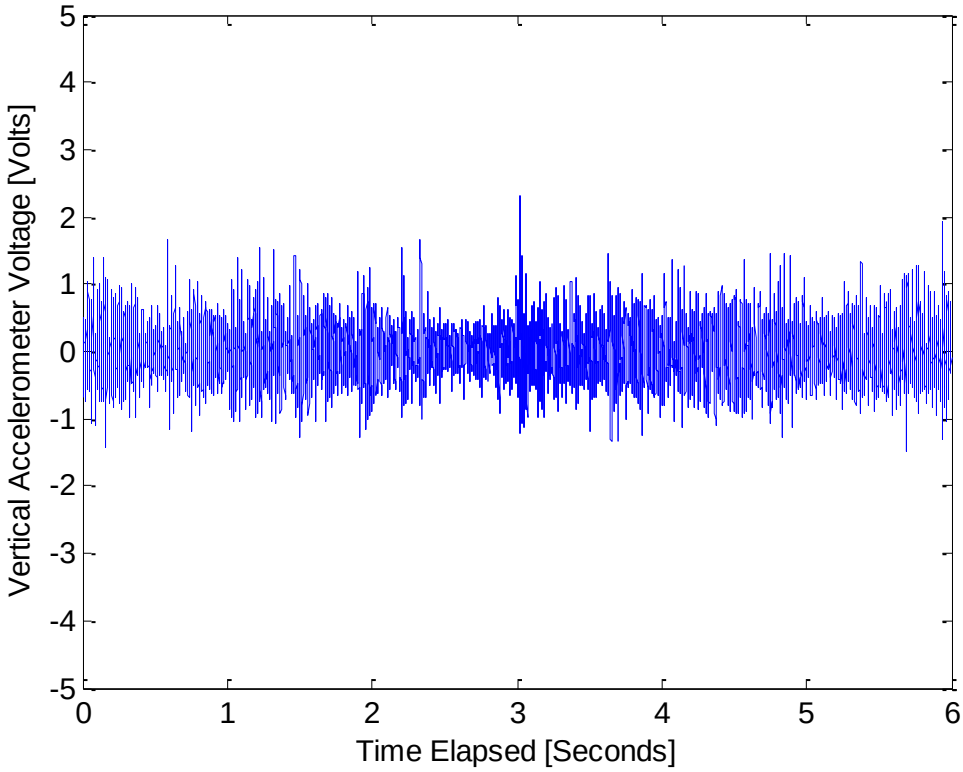


Figure 3-18: The transducer response from the loose fishplate condition does not show any clear signs of anomalous events and appears to be random.

Zooming in closer to the fault location in the loose fishplate condition, evidence of a disturbance becomes more prominent. Figure 3-19 illustrates the difficulty in identifying novelties within the signal even when the location of the fault is known. Therefore, the decision support system will attempt to detect such anomalous signals and warn the operator of potential damage.

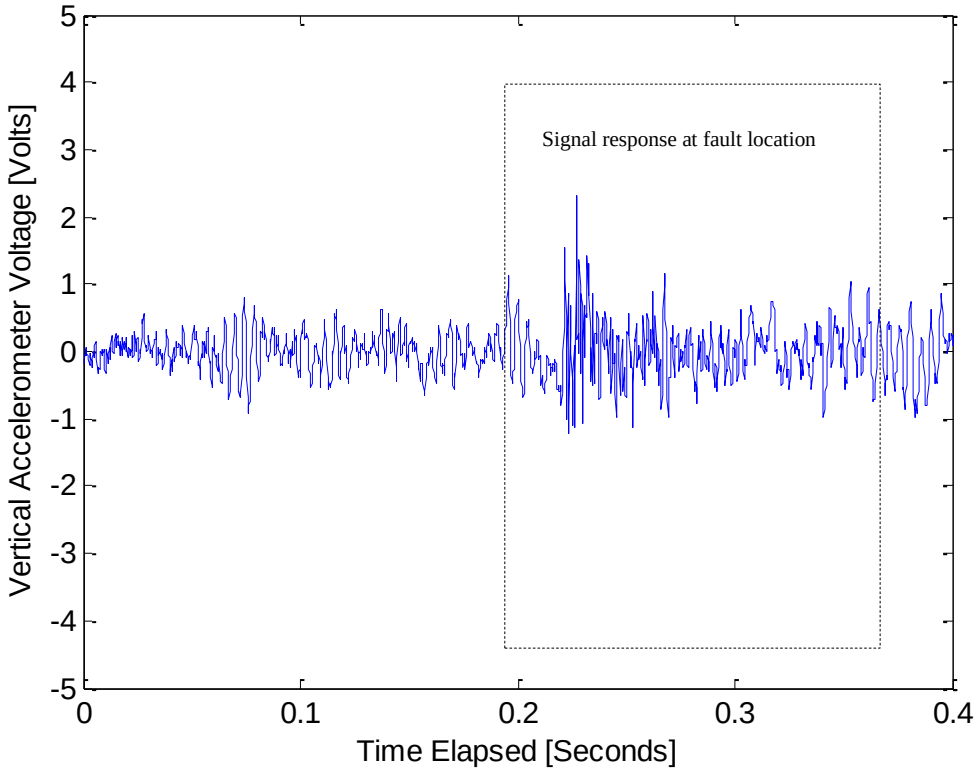


Figure 3-19: A close-up of the loose fishplate response located at sample 1.5×10^4 (1000 samples is roughly 0.5 m in terms of distance, at roughly 3 m/s).

Either case of washout did not reveal itself in the transducer response. At this point, the author found it useful to construct a series of plots, one for each failure mode, superimposing 30 healthy signals onto 30 faulted signals to see if the general shape of the signals would conform under each condition. The procedure is depicted in Figure 3-20, Figure 3-21, Figure 3-22, and Figure 3-23. The figures illustrate that subtle faults such as the washout conditions are very difficult to discern from healthy conditions, and more severe faults can be identified in the vibration signature when compared to healthy signatures.

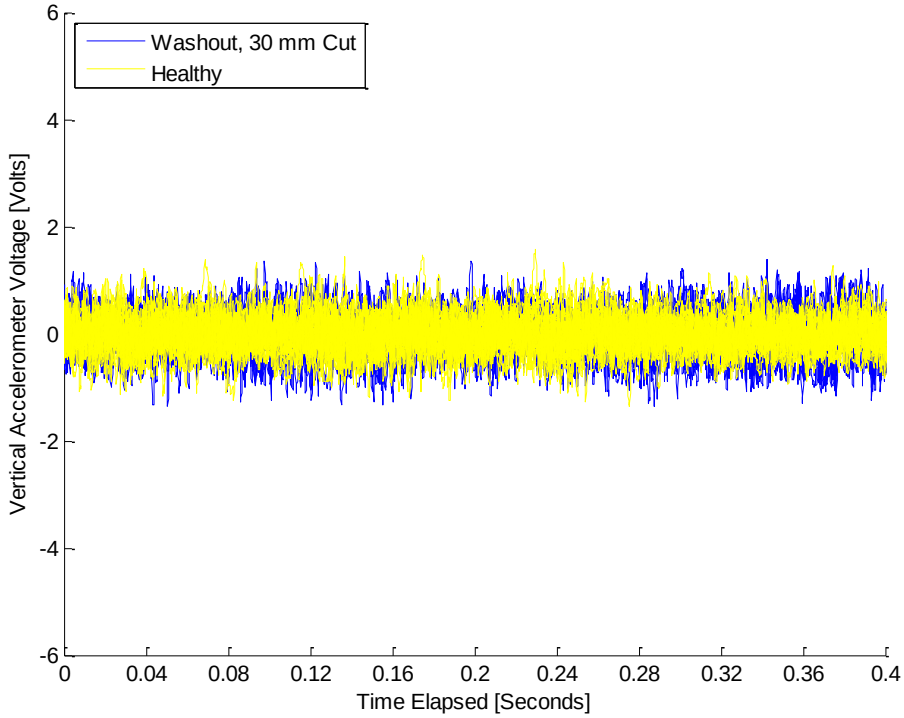


Figure 3-20: Superimposed signals showing shape differences of washout (30 mm crack) and healthy data.

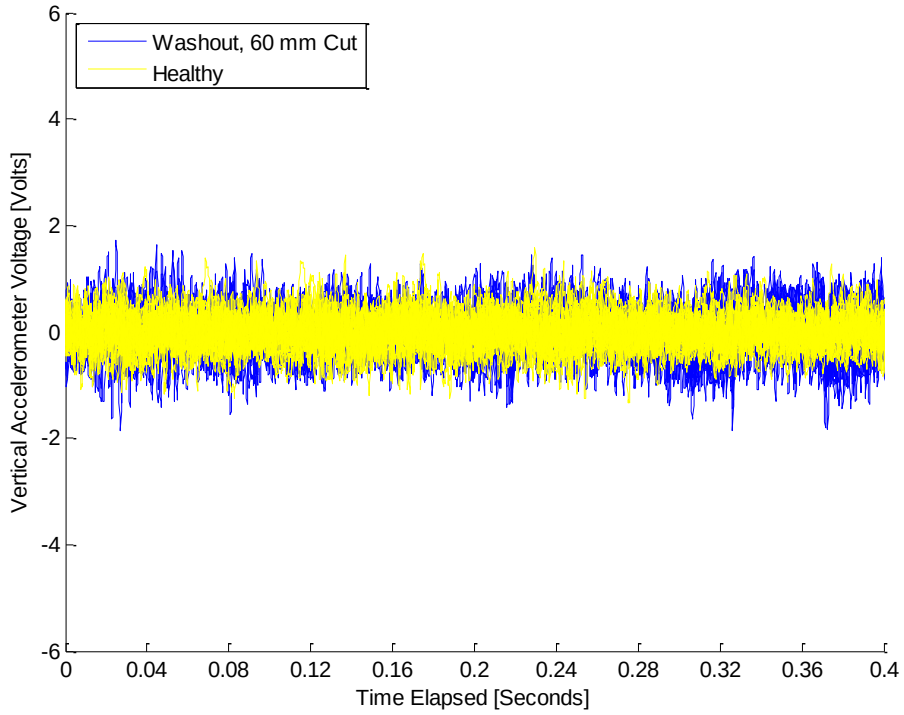


Figure 3-21: Superimposed signals showing shape differences of washout (60 mm crack) and healthy data.

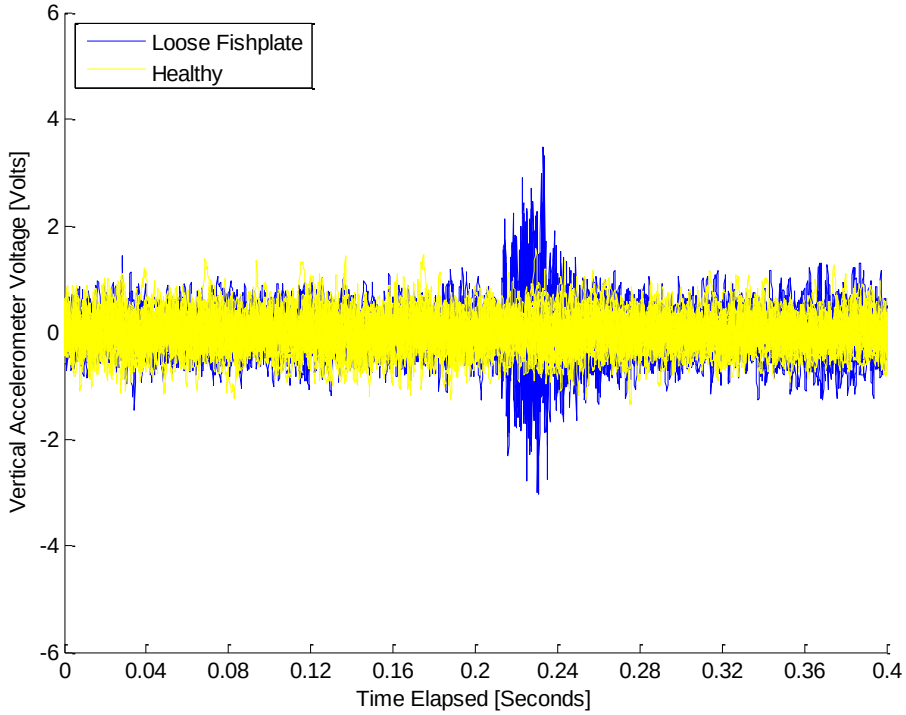


Figure 3-22: Superimposed signals showing shape differences of loose fishplate and healthy data.

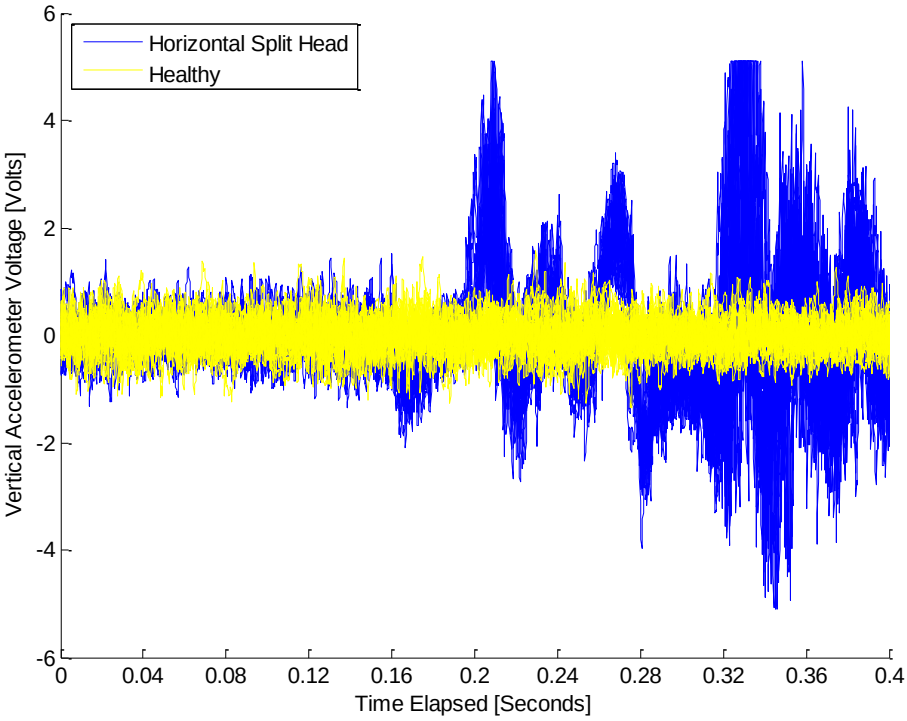


Figure 3-23: Superimposed signals showing shape differences of horizontally split head and healthy data.

From the preliminary observations, the aspirations of a successful fault detection system are promising since preliminary visual inspection of the data revealed structural similarities in different groups of data; however, multivariate feature extraction and pattern recognition techniques are considered vastly more effective in the case of fault detection of machinery in variable states of operation and duty. Signal processing and further development of the fault detection system are outlined in Chapter 4.

Chapter 4

4 Signal Processing and Fault Detection

4.1 Overview

At this stage in the development of the decision support system, the experiments have been conducted and the data is in its raw form. Several steps are required to transform a group of continuous machine signals (the raw data) to useful decision support. These steps include segmentation, preprocessing, windowing, feature extraction, dimension reduction, and classification.

Segmentation is the first stage in refining the data to a more useable form. Once a working segment is defined, the segment is windowed. Windowing is the formation of subsets of data, within the segment, that are used to produce descriptive feature vectors. The windows are further reduced by means of feature extraction. Feature vectors are then reduced by means of dimension reduction techniques. Once a reduced feature vector is formed, it is passed to the classifier. Figure 4-1 shows the major processing steps of the fault detection process.

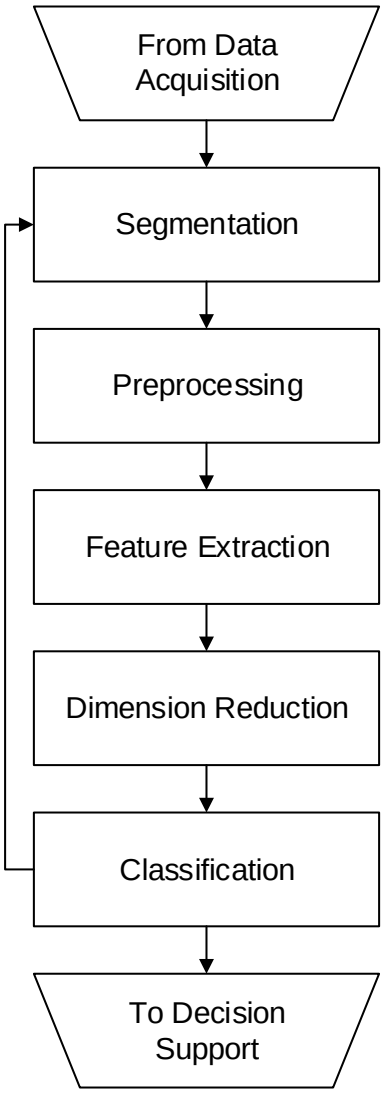


Figure 4-1: Work-flow through signal processing stage

4.2 Segmentation

Segmentation is the process of partitioning a discrete time-varying signal into distinct sets of samples that can be compared with one another in the classification stage. The goal of segmentation is to reduce variations in the signal and to reduce the signal size itself to focus on a segment of interest for fault detection.

By segmenting the data for sections with constant speed, the signal variation due to variable speed is eliminated; start-up transients are omitted in the segmentation process. Thus, segmentation is a form of experimental control. Similarly, start-up transients might contain valuable information for the task of fault detection, so they might be the segment of interest.

Consider the colour spectrum of visible light to be components of a signal; in the broadest sense, light content from two images can be compared, searching for minute differences across the range of colours or light components. Furthermore, by focussing more attention on a narrower range of wavelengths, better resolution is gained to detect anomalies within each range. Further, one may choose to zoom in on very narrow spectral widths, investigating each width across the spectrum for novel activity. The trade-off of each approach is that by investigating smaller segments of a signal, it becomes easier to notice changes in the segment, but it requires more segments to be analyzed.

There are several options for segmenting the data as it is collected from a railcar travelling along the railway track. In particular, one can consider the entire loop of track or individual sections of the track as a single segment. In the configuration shown in Figure 4-2, the train travels in one direction around the loop. As the train completes a voyage around the track loop, the condition

monitoring system records the system response for one complete cycle. The following sections discuss some possible strategies for segmenting the data with attention paid to efficiency and practicality. The three approaches are summarized in Table 4-1.

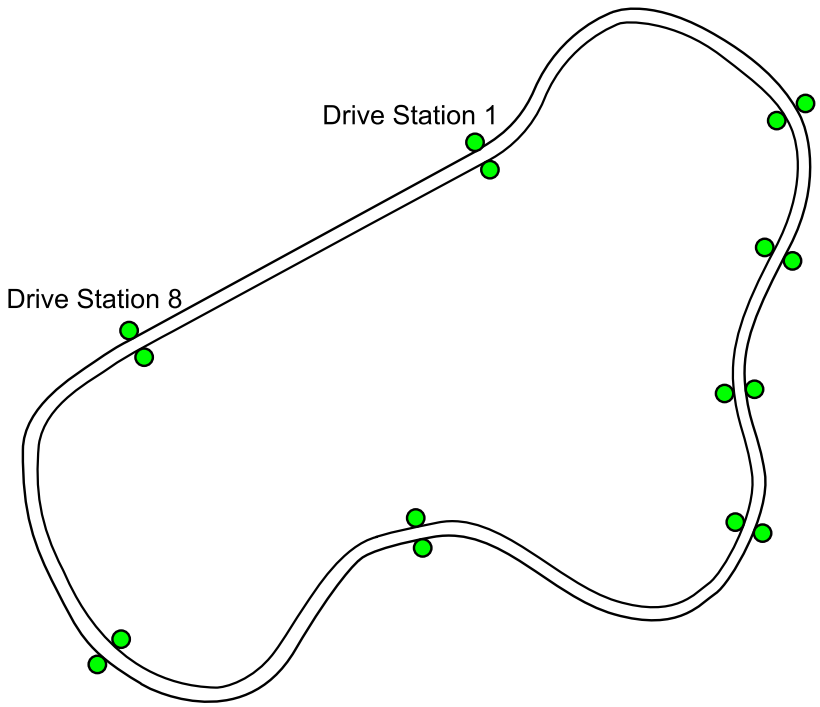


Figure 4-2: A plan view of the track layout with drive stations

4.2.1 Generalized Approach: One Complete Track Loop as a Segment

The most generalized segmentation approach is to crop the data such that an entire track cycle is defined as a segment. In this scenario, the start of the segment is also the termination point as the train loops back to its origin. There are both advantages and disadvantages associated with this approach to segmenting the data.

The advantage of this approach is a generalized data segment that captures the response throughout an entire cycle. This leads to a large data set, rich with information gathered along the complete track length; it is a snapshot of the entire loop.

The drawback of this approach is that it requires a much more complex model to accommodate signal variation due to changes in track geometry, ground conditions, fluctuations in speed, and operational variations such as stopping and starting at different locations along the track.

4.2.2 Quasi-Generalized Approach: A priori-based Segment Definition

This approach is based on the assumption that geometrically similar sections of track will generate similar vibration responses under the same operating conditions. For example, straight sections of track would be grouped together for comparison since they are geometrically similar. Similarly, curved sections of track would be only be compared to other curved sections of track for fault detection.

To implement this type of segmentation, discrete segments of the track would need to be selected by the designer of the fault detection system to define comparable segments. For example, the designer would analyze the existing track and define where all straight sections begin and end, then use only data from those segments to train and develop the fault detection system. Once the fault detection system is in place, it would require input or the ability to detect when it has entered a straight section or curved section.

The overarching benefit of this approach is that it results in a series of simpler models since the signal variation across the segment is reduced. However, it would be a challenge to precisely locate these sections physically along the track network. Furthermore, there may be differences in (seemingly similar) sections that may not have been accounted for (e.g., track condition, ground condition supporting the track, track grade, and etcetera).

4.2.3 Specialized Approach: Direct Comparison

This approach is the most specialized; it entails segmenting the track into small sections corresponding to, one metre, for example. This small section of track is only compared to itself each time the train passes over it.

The advantage of this scheme is the reduction in response variability, corresponding to reduced variability in track geometry; however, this provides only one comparable segment each time the train travels around the entire track loop, since the one-metre section of track at metre x is only compared to that same one-metre section the next time the train passes over it. From a condition monitoring standpoint, this approach would yield the best results since it is the least variable.

Table 4-1: Summary of Segmentation Approaches

Approach	Definition of Segment	Benefit	Cost
Generalized	Entire track	- Results in most data available for training of system - No a priori knowledge required to pre-define segments	- Inherent variation in segments could result in over generalized model - May be difficult for system to differentiate between normal variances in track and mechanical faults
Quasi-Generalized	Geometrically similar sections of track	- Improved reduction in variability from drastically different track conditions - Could result in simpler classification	- Requires a priori knowledge and significant intervention in setting up the system to locate similar segments
Specialized	Specific location on track	- Most direct comparison for determining track condition - Lowest signal variation due to track setup	- Least data available for training of system

4.2.4 Segment Selection

The author has chosen to simplify the problem by segmenting the signal using the specialized approach (direct comparison). It was chosen since it captures the least amount of signal variation, making the overall fault detection process simpler. The fault detection process is simplified since the segment conforms to a more steady state signal.

Consider an inner race fault in a rolling-element bearing; as the bearing spins at a constant speed, the rolling elements interact with a flaw in the inner race. At a high rotational speed, the rolling elements interact with the flaw many times per second. This corresponds to a large number of revolutions, each characterizing the state of health of the bearing as it travels through one cycle. In the end, the signal from one full revolution can be compared to thousands of other revolutions. Fault detection techniques can effectively differentiate between one bearing state and another since so much data is available. That is, ten thousand healthy revolutions can be compared to ten thousand faulted revolutions, leading to a statistically valid classification.

In unsteadily operating machinery, the fault signal is not always repeated, eliminating this high frequency of observation, which produces smaller data sets. By controlling the segmentation, a transient signal can be decomposed into a series of steady-state signals. At this stage, the data is cropped to define a segment of the signal that will be used for the development of the decision support system. Consider Figure 4-3:

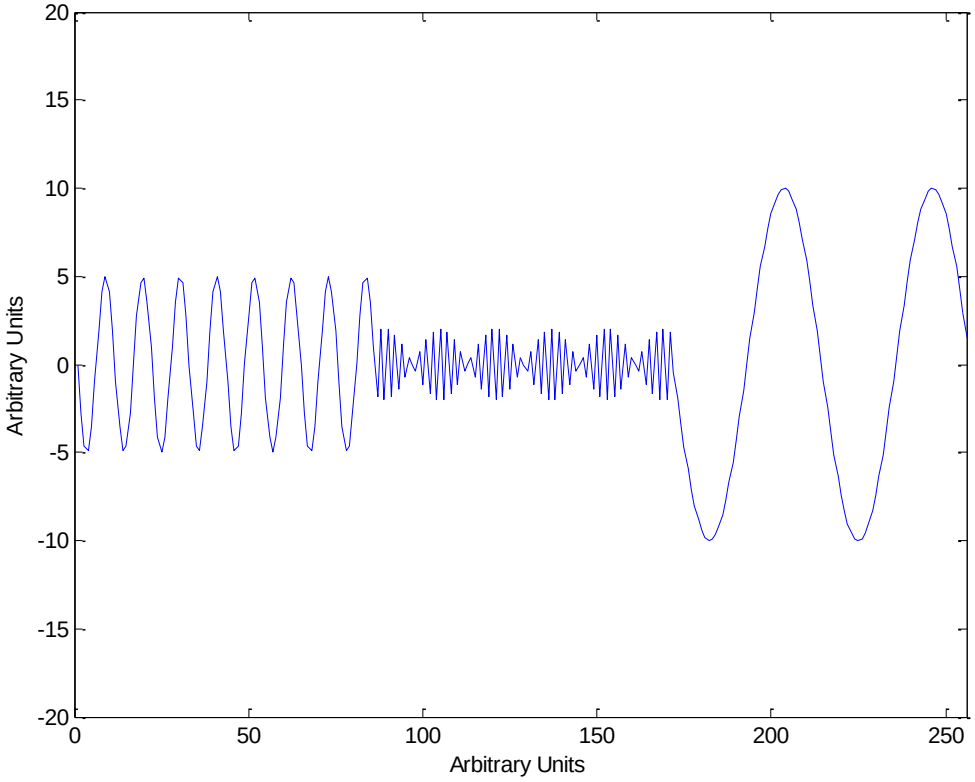


Figure 4-3: An Example of a Transient Signal

If the entire signal is used for feature extraction, the transient nature of the signal complicates the classification task. However, if the signal is segmented by windows with similar amplitudes, the task is simplified. Figure 4-4 illustrates the concept more clearly.

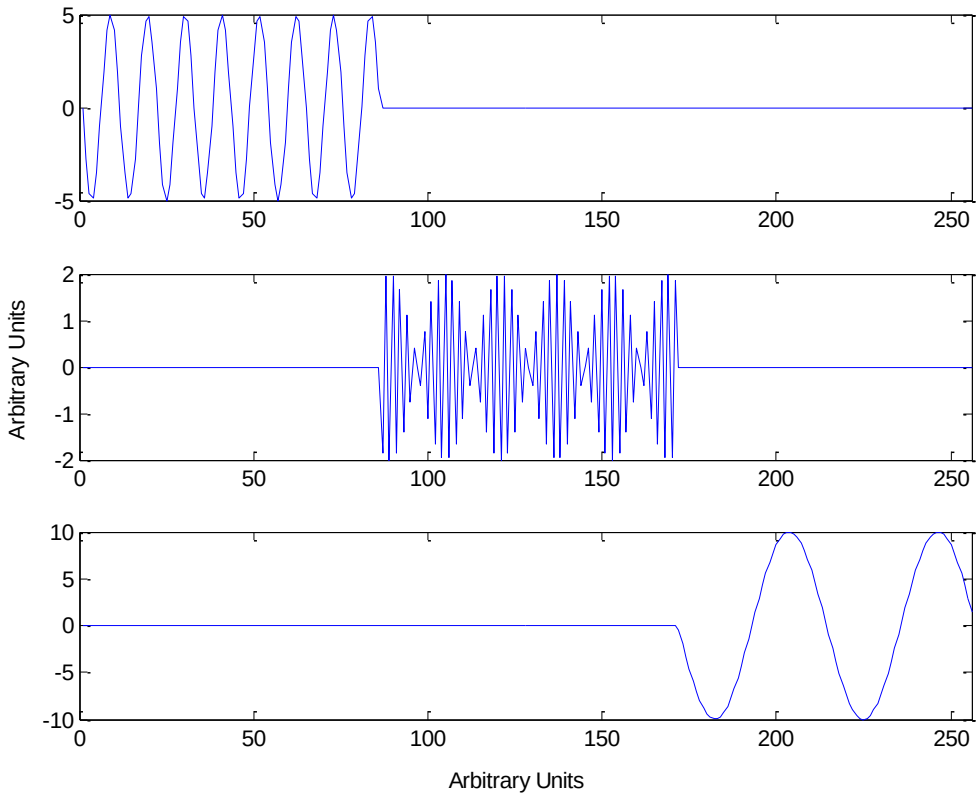


Figure 4-4: An example of segmenting a transient signal to create stationary segments.

It becomes evident that classifying a nonstationary signal can be much more difficult than classifying a stationary one. The overarching reasoning behind this is that the fault detection system is essentially used to detect variability, wherein, by reducing baseline variability in the signal (making it more like a steady-state signal), the problem is simplified.

Therefore, by using the specialized approach (direct comparison), the likelihood of developing a successful fault detection system increases since it is a simplified problem.

The following section examines the effects of segment size selection with a performance comparison.

4.2.5 Segment Size Selection

Up to this point, the direct segmentation approach has been selected, but the exact size of each segment has not been defined. Segment size can greatly affect the performance of this fault detection system in two ways: if the segment is too large, the transient signal (due to the fault) it is supposed to detect can be lost amongst the surrounding signal; and, if the segment is too small, the transient signal can be split across two segments.

Various segment sizes were considered during the development of the decision support system. Strict rules on window size or shape selection are not well established in the literature, so a survey of different window sizes was taken, while a rectangular window was used[3].

Preliminary fault detection results representing average false positive and false negative classifications were used to compare segment size performance. The results are shown in Figure 4-5.

[3] The simplest windowing function was chosen since its implications with the discrete-time Fourier transform would not be encountered.

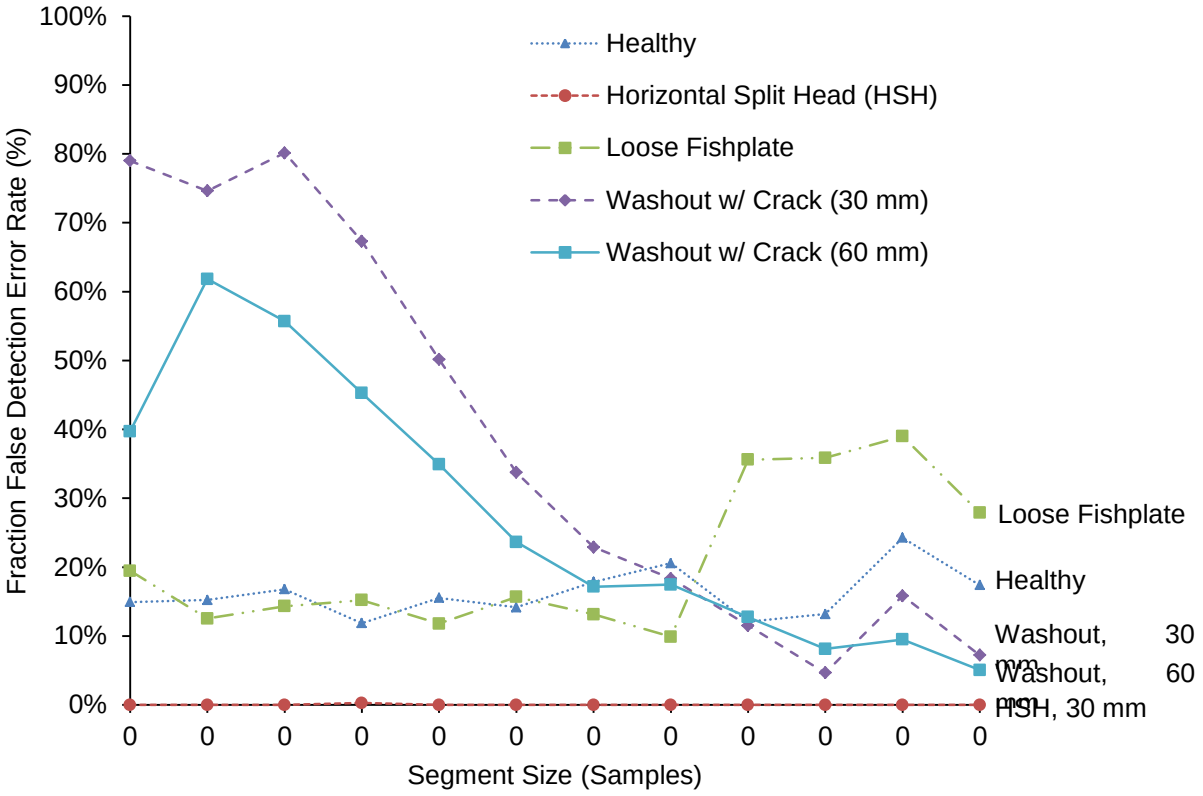


Figure 4-5: A survey of segment size performance for fault detection.

A number of conclusions can be drawn based on the results of this study. The results indicate that healthy signals are classified correctly irrespective of the window function size. That is, signals measured at the fault location do not vary across the spectrum of proposed window lengths. The author attributes this result to the fact that the signal is steady in this scenario (i.e., transients due to impact do not appear in the healthy track section). Similarly, the horizontally split head fault is insensitive to the various segment sizes investigated in this study. This is to be expected since the observations in Figure 3-23 of section 3.5.2 clearly illustrate the presence of the fault. The window size selection had a more sensitive effect on the loose fishplate and washout conditions.

Focusing attention on the behaviour of the loose fishplate condition across the range of window sizes, it can be seen that after 2,000 samples per window, the error rate rapidly increases. Referring back to section 3.5.2, the general shape of the loose fishplate vibration response appears as a short burst, which attenuates rapidly. It now becomes evident that window sizes greater than roughly 2,250 samples tend to smear the loose fishplate vibration response across the window.

The general shape taken by each washout condition mirrored the loose fishplate performance trend. In both washout cases, the error rates drop drastically as window lengths increase; this is attributed to the subtle vibration response in the washout conditions (i.e., no inconsistencies in the vibration signature are visually discernable in these cases, yet trends in Figure 4-5 suggest that the inconsistency seems to occur across a larger span).

A window size of 2,000 samples exhibited the best overall performance. Therefore, subsequent analyses used 2,000 samples as the standard window size.

4.3 Preprocessing

Preprocessing techniques are used as signal enhancement tools used to gain further insight into the nature of a signal. Filtering, frequency analysis, and time-frequency analysis were briefly investigated through the development of this study, but did not enhance the time series vibration signals. Appendix B discussed the results of some preprocessing investigations that were unsuccessful.

4.4 Feature Extraction

Feature selection based on the literature review established a series of five feature vectors for comparison. Following the preliminary results, a principal component analysis was applied for dimension reduction of the feature vectors.

Table 4-2 defines the different feature sets that were compared in this study. The performances of each individual feature set were compared to determine the most effective characterization of the data segments they represented.

Table 4-2: Feature set metrics.

Statistical Features	Maximum
	Minimum
	RMS
	Arithmetic Mean
	Absolute Mean
	Harmonic Mean
	Standard Deviation
	Kurtosis
	Clearance Factor
	Crest Factor
	Impulse Factor
	Shape Factor
	Interquartile Range
	Trapezoidal Integral

Transient Features	Counts
	MARSE
	Rise Time
	Duration
	Peak Amplitude

Statistical Features with PCA	Maximum
	Minimum
	RMS
	Arithmetic Mean
	Absolute Mean
	Harmonic Mean
	Standard Deviation
	Kurtosis
	Clearance Factor
	Crest Factor

AR Coefficients	Order 10
	Order 20
	Order 50
	Order 100
	Order 200

Statistical, AR100	Statistical features of the 100 AR coefficients

4.4.1 Statistical Features

Statistical features are often used as a first step for feature selection in condition monitoring applications. They represent the properties of the statistical distribution of the data set. As such, they can be well-suited to the characterization of condition monitoring signals.

The following is a list describing characteristic equations or pseudo code of statistical features:

Maximum:

```
max = x(1);
for i=1:length(x)
   if x(i)>max
      max=x(i);
   end
end
```

Equation 4-1

Minimum:

```
min = x(1);
for i=1:length(x)
   if x(i)<min
      min=x(i);
   end
end
```

Equation 4-2

RMS:

$$x_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$$

Equation 4-3

Mean:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Equation 4-4

Absolute Mean:

$$\bar{x}_{ABS} = \frac{1}{N}\sum_{i=1}^{N} |x_i|$$

Equation 4-5

Geometric Mean:

$$\left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}}$$

Equation 4-6

Harmonic Mean:	$\left(\frac{1}{N}\sum_{i=1}^{N} x_i^{-1}\right)^{-1}$	Equation 4-7
Standard Deviation:	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$	Equation 4-8
Kurtosis:	$\sum_{i=1}^{N}\frac{(x_i - \bar{x})^4}{N\sigma^4}$	Equation 4-9
Clearance Factor:	$CLF = \frac{x_{max}}{\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert x_i \rvert}\right)^2}$	Equation 4-10
Crest Factor:	$CF = \frac{x_{max}}{x_{RMS}}$	Equation 4-11
Impulse Factor:	$IF = \frac{x_{max}}{\bar{x}_{ABS}}$	Equation 4-12
Shape Factor:	$SF = \frac{x_{RMS}}{\bar{x}_{ABS}}$	Equation 4-13
Range:	$x_{range} = x_{max} - x_{min}$	Equation 4-14
Interquartile Range:	median(x(indexOf(median(x)):length(x))- median(x(1:indexOf(median(x))));	Equation 4-15
Trapezoidal Integral:	$\frac{1}{N}\sum_{i=1}^{N}(x_{i+1} - x_i)$	Equation 4-16

4.4.2 Statistical PCA Features

A principal component analysis performed on the statistical feature set determined that the first 10 features represented 99.9 percent of the variance in the vector. A reduced feature set was defined with only the first 10 of the 14 statistical features.

4.4.3 Autoregressive Features

Autoregressive (AR) models are linear prediction models that determine future values based on previous ones. The coefficients of the model define it, and can be determined using the autocorrelation function in Equation 4-17, then, by solving the Yule-Walker equation in the form of Equation 4-18:

$$\hat{r}_{xx}[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]x[n-k], \quad 0 \le k \le p-1, \qquad \text{Equation 4-17}$$

where $\hat{r}_{xx}$ is the autocorrelation function, k is the coefficient index, x is the signal, N is the number of elements in the signal, and p is the number of poles in the autoregressive model.

$$\begin{bmatrix} r_{xx}[0] & r_{xx}[-1] & \cdots & r_{xx}[-p+1] \\ r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[-p+2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[p-1] & r_{xx}[p-2] & \cdots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix} = -\begin{bmatrix} r_{xx}[1] \\ r_{xx}[2] \\ \vdots \\ r_{xx}[p] \end{bmatrix} \qquad \text{Equation 4-18}$$

Autoregressive (AR) coefficients were used as feature vectors in the analysis of the feature selection process. AR model coefficients of order 10, 20, 50, 100, and 200 were chosen as feature vectors for the classification task.

4.4.4 Statistical AR100

Statistical features of the AR100 model coefficients were used as a feature vector for the Statistical AR100 feature set.

4.4.5 Transient Signal Features

In acoustic emissions (AE) analysis, transient features shown in Figure 4-6 are used as model parameters to characterize micro strain energy bursts.

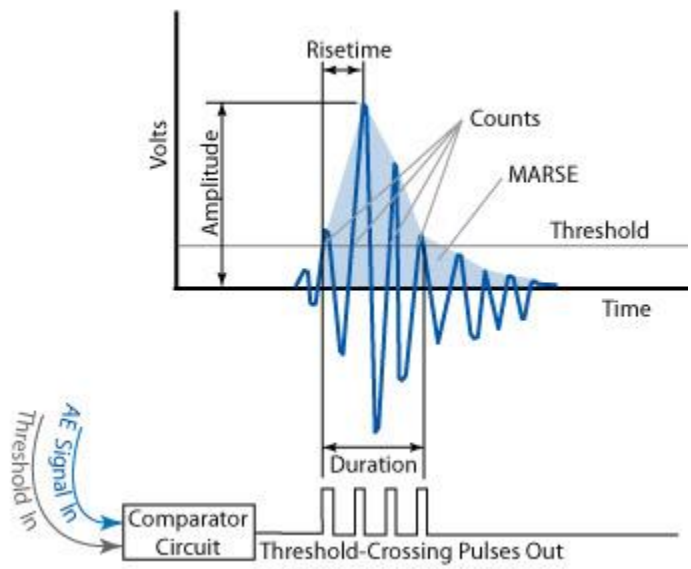


Figure 4-6: Acoustic emissions signal features adopted to this study (taken from [34]).

Equation 4-19 through Equation 4-22 show relevant code snippets used to determine the signal features.

Threshold: threshold = x_RMS; Equation 4-19

Counts: Equation 4-20

```
hits=0;
for i=1:numel(x)
    if x(i)>x_RMS
        hits=hits+1;
    end
end
```

Rise Time: Equation 4-21

```
crossing_index=0;
x_max=[x(1) 1];
for i=1:numel(x)-1
    while threshold_crossed==0
        if x(i)<x_RMS && x(i+1)>x_RMS
            threshold_crossed=1;
            crossing_index=i;
        end
    end
    if x(i+1)>x_max(1)
        x_max(1)=x(i+1);
        x_max(2)=i+1;
    end
end
rise_time=(x_max(2)-
crossing_index)*sampling_period;
```

Duration: Equation 4-22

```
pdxdt=[];
ndxdt=[];
for i=1:numel(x)-1
    if x(i)<threshold && x(i+1)>threshold
        pdxdt=[pdxdt i];
    end
    if x(i)>threshold && x(i+1)<threshold
        ndxdt=[ndxdt i];
    end
end
duration=(pdxdt(numel(pdxdt))-
pdxdt(1))*sampling_period;
```

The max() and trapz() functions in MATLAB® were also used to determine the peak amplitude of the signal and MARSE[4], respectively. The transient feature set selected for this study was based on commonly used feature set metrics in acoustic emission-based fault detection. The relationship between the current work and fault detection using acoustic emissions (AE) is the transient nature of the fault signal.

A performance comparison of the different feature sets is presented in section 0.

4.5 Feature Vector Dimension Reduction

The 'curse of dimensionality' was the motivation for feature set dimension reduction. It states that the required amount of training data increases exponentially with the addition of features. If a feature vector has n features, its covariance matrix contains n^2 parameters. Therefore, if the amount of training data is not large in comparison to n^2, the classifier boundary will over-fit the data set since training is based on the distribution of features. By reducing the feature set and its covariance, the classifier is trained in a lower-dimension feature space. Therefore, the required amount of training data can be reduced using dimension reduction.

Principal component analysis (PCA) was used for dimension reduction of feature sets. The linear dimension reduction technique maximizes the variance of each component (feature) with respect to its predecessor (the previous feature in the feature vector), on an orthogonal basis

[4] The signal MARSE is equal to the area under the envelope containing the half-wave rectified portion of the signal; it is graphically illustrated in Figure 4-6. The MARSE is a representation of the transient event energy burst and is commonly used as a signal feature in acoustic emission-based fault detection.

(such that they are linearly independent). PCA was used to identify the amount of variance in the data represented by each component of the feature vector.

For example, if the first three features in a feature vector represent the same variance that the full fifty components represent, the additional components are not linearly independent and do not contain any additional data contributing to signal variance. Therefore, PCA can be used to identify features that independently generate the most variance in the signal.

In the context of this study, a PCA was performed to examine apparent performance enhancement by means of dimension reduction of feature vectors. The assumption was that by reducing the dimensionality of the feature vector while maintaining the same general description of the signal, the classifier would produce a more generalized model; a more generalized model would effectively produce fewer false negative results, yet remain effective when identifying outliers (or faults).

4.6 Classification

4.6.1 Overview

The objective of classification is to identify the class to which a new observation belongs. For example, one might be attempting to distinguish between triangles and quadrilaterals; a distinguishing feature would be the number of vertices. If a new element is presented and has three vertices, it should be classified as a triangle. Therefore, the output of classification is the assignment of a new observation to a specific predefined class or group.

In the context of CBM, the role of classification is to distinguish between feature vectors belonging to groups representing different machine states (healthy versus unhealthy). To accomplish this, training data is used to construct a mathematical model that describes each class. The model defines a boundary that discriminates between membership and outlier; however, in a multidimensional feature space, the boundary is geometrically abstract and strictly mathematical.

The first step in the classification stage was to create two subsets of the available data: a training data set (85% of the total data set) and a quarantined data set (15% of the total data set). The original data set was randomly shuffled such that the subsets were representative samples of the original set, and to ensure that the classifier was not trained on sequential observations.

Classification was separated into two stages: training and operation. The training stage was used to generate the classifier using exemplar data. The classifier was fit to the exemplar data with an acceptance criterion, namely, fraction rejection. Fraction rejection was defined as the fraction of allowable exclusions that is acceptable[5]. The fraction rejection was the criterion used to stop the training process (i.e., when the model was in agreement with the fraction rejection, training was stopped). Figure 4-7 illustrates a simple classifier with two different fraction rejections with a simplified 2-dimensional data set. Once the classifier was trained, quarantined data was introduced to test performance of the classifier. Quarantined data represented new observations to be classified as either healthy or unhealthy data. Identities of the quarantined data were

[5] It is important to allow exclusions; otherwise, the classifier can become over fit to the exemplar data and perform poorly during operation. Similarly, allowing too many exclusions can result in an over-generalized classifier.

known by the author; therefore, based on classification results, the number of false classifications was used as a performance measure.

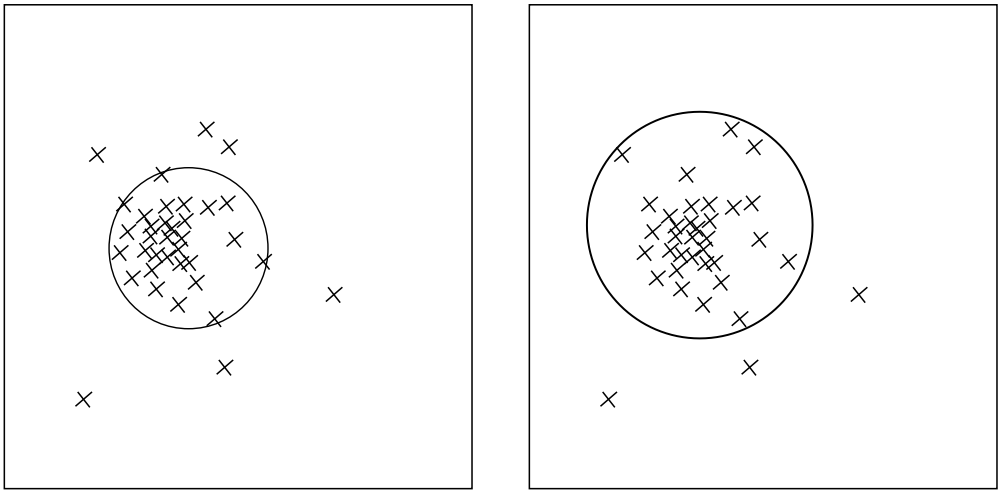


Figure 4-7: Classifiers with different allowable fraction rejection.

One-class classification is a specific type of classification that results in membership or non-membership (outlier). The objective of one-class classification is to determine if the new observation is a member or not (healthy or not healthy). This is useful in fault detection since damage can be present in many forms and varying degrees of severity, making it difficult to define a specific class for each type of damage. As a result, one-class classification can be used to identify if new observations are healthy or not; however, a drawback to this approach is the inability to diagnose faults.

As a proof of concept, one-class classification was used for fault detection. Performance was evaluated by considering false negatives and false positives; false negatives were defined as observations of healthy operation classified incorrectly as faults (false alarms); false positives

were defined as observations of damage classified incorrectly as healthy (missed faults). This may seem counterintuitive as compared to other disciplines; however, in the context of this study, positives are considered healthy datum and negatives are considered faulted datum. Table 4-3 illustrates the definition of false positive and false negative errors defined in this study.

.

Table 4-3: Types of error defined in this study (adapted from [35]).

		True Class Label	
		Target	Outlier
Assigned Label	Target	True Positive Target Accepted	False Positive Outlier Rejected
	Outlier	False Negative Target Rejected	True Negative Outlier Rejected

The following sections define the classifiers that were compared in this study. The MATLAB Data Description toolbox for pattern recognition [35] was used for the training and performance evaluation of the various classifiers. The classifiers were all trained to fit a fraction rejection threshold of 2.5 percent.

4.6.2 Gaussian Density Estimation

Gaussian density estimation is a probability density estimation using a Gaussian basis function. A Gaussian density estimator fits a normal curve over the data set. For multivariate systems, each dimension has an associated Gaussian density estimation. The distance of a new object to the probability density function determines if it is associated to the target (healthy) class, or the outlier (fault) class. This method is effective for general data sets with a Gaussian distribution. It is recommended for use as a starting point in one class classification, defined in [35].

4.6.3 Parzen Density Estimation

Parzen-window density estimation is a non-parametric form of kernel density estimation. The non-parametric property of the Parzen density estimator allows the kernel density function to be estimated without *a priori* knowledge of the distribution. In comparison to the Gaussian density estimation, where a Gaussian function is fit to the data set, a Parzen density function is the sum of Gaussians fit to each datum. This leads to the development of a unique kernel density function that is tailored to a specific data set. Figure 4-8 shows the behaviour of Parzen-window density estimation according to different data sets. The progression of data distribution from Figure 4-8 (a) to (d) illustrates how Parzen-window density estimation remains well-suited to even small data sets. Also, for data sets that do not follow a Gaussian distribution, as in a system with a discrete number of operating modes, the Parzen-window density estimator remains well-suited for the data description.

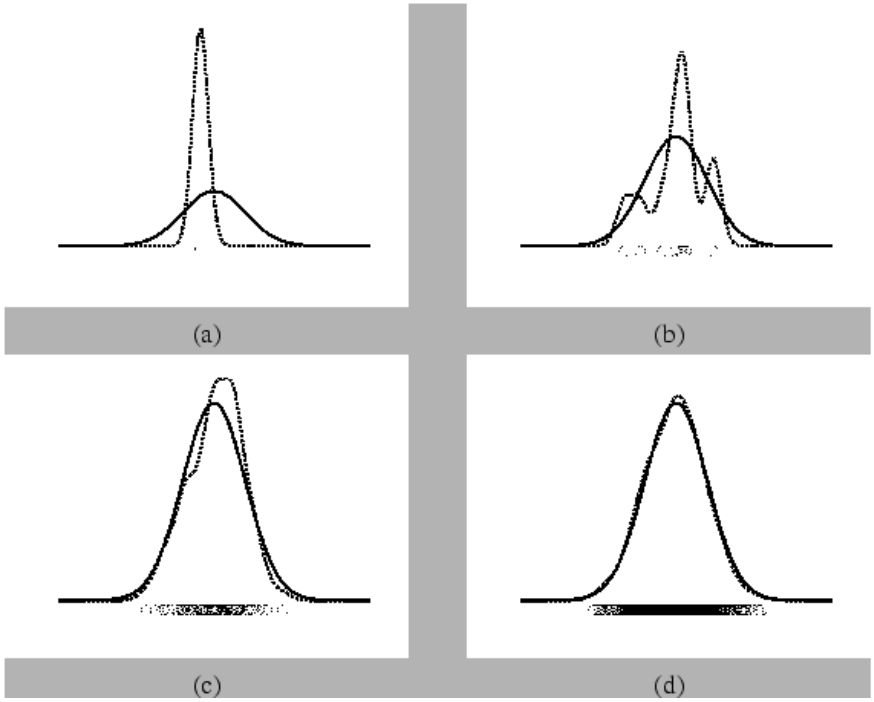


Figure 4-8: A comparison of Gaussian density estimation (solid) to Parzen-window density estimation (dotted) (taken from [36]).

For classification, new objects are fit with a Parzen window, then a nearest neighbour approach is taken, wherein the cumulative convolution of neighbouring window functions defines the likelihood that the new object is indeed a member or an outlier. The probability density function is modelled with a maximum likelihood condition, with a leave-one-out methodology to achieve the desired training error. The accordance threshold (fraction rejection) is set by the user prior to training. The width parameter is optimized by the maximum likelihood estimation function in the data description toolbox, parzenml.

4.6.4 K-Means Clustering

K-means clustering is a classification method that produces *k*-defined clusters to separate feature vectors. Simply put, the user must define the number of clusters, wherein the algorithm iterates

the definition of cluster centres. In the initial step, cluster centres are arbitrary. Then, data points are associated to the nearest cluster centre. After each data member is associated with a cluster, new centres are computed based on the centroid of data points belonging to each cluster. The association step is then repeated and the new centroid is computed again. These two processes are repeated until convergence is reached. Figure 4-9 exemplifies the *k*-means process. Arbitrary cluster centres are established in Figure 4-9 (A); data points are assigned to the arbitrary centres based on squared Euclidean distance in Figure 4-9 (B); new centres are defined based on cluster centroids in Figure 4-9 (C); and, convergence is eventually reached, wherein the final result is shown in Figure 4-9 (D).

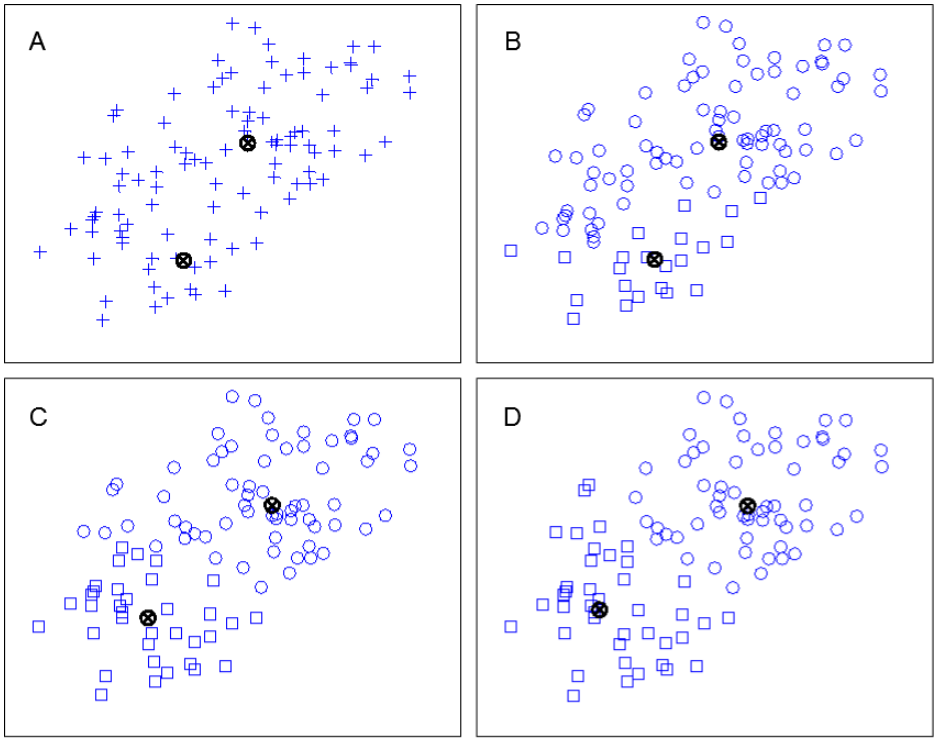


Figure 4-9: Stepping through the K-Means algorithm.

As a classification tool, the cluster centres act as a basis in which new members are placed in terms of their square distance (or some other distance measure) to a nearest neighbour. The K-means classifier was trained using five clusters, or prototypes based on the default parameters set within the data description toolbox.

4.6.5 Nearest Neighbour Clustering

The nearest neighbour data descriptor uses the nearest neighbour clustering algorithm. The algorithm begins with each data member as its own cluster. Next, the two points with the least separation distance (usually Euclidean) are bundled into a new cluster. The process terminates when all the data belongs to the same cluster. Then, based on the number of defined clusters, the latest *N*-linkages are broken, leaving the desired number of clusters. In the case of one class classification, the leave-one-out method applies.

4.6.6 Auto-encoder Classifier

An auto-encoder is a feed-forward backpropagation artificial neural network-based classifier that reconstructs the original data set by means of model-based mapping. The classifier is trained in the way of a traditional artificial neural network, where the mean-square-error is used as a training threshold. In terms of classification, outliers are separated when they surpass the threshold of mean-squared-error in the reconstruction phase.

The auto-encoder was trained with five hidden layers, a hyperbolic tangent sigmoid transfer function for the hidden layers, and a linear transfer function for the output layer. The training used Broyden, Fletcher, Goldfarb, and Shanno (BFGS) quasi-Newton backpropagation method

since it had lower computational memory requirements than the faster Levenberg-Marquardt backpropagation method.

4.6.7 PCA Classifier

The PCA data descriptor is another reconstruction-based classifier that uses residual error to categorize new objects. The mapping between the original feature space and the PCA subspace is used to determine the projection of the new object on the PCA subspace. The reconstruction error is defined by Tax [35] as

$$f(x) = \left\|\mathbf{x} - \mathbf{x}_{proj}\right\|^2. \qquad \text{Equation 4-23}$$

If $f(x)$ is greater than the user-defined threshold, the object is rejected as an outlier. A correlation coefficient of 0.9 was used for training based on the default parameters set within the data description toolbox.

4.6.8 Support Vector Classifier

A support vector machine is a mathematical model that represents a hyperplane (in multidimensional learning) that separates data classes, optimized for the greatest margin of separation. Figure 4-10 demonstrates the function of support vectors and illustrates the fact that the geometric representation is tangible, unlike the workings of abstract classifiers, such as neural networks [37].

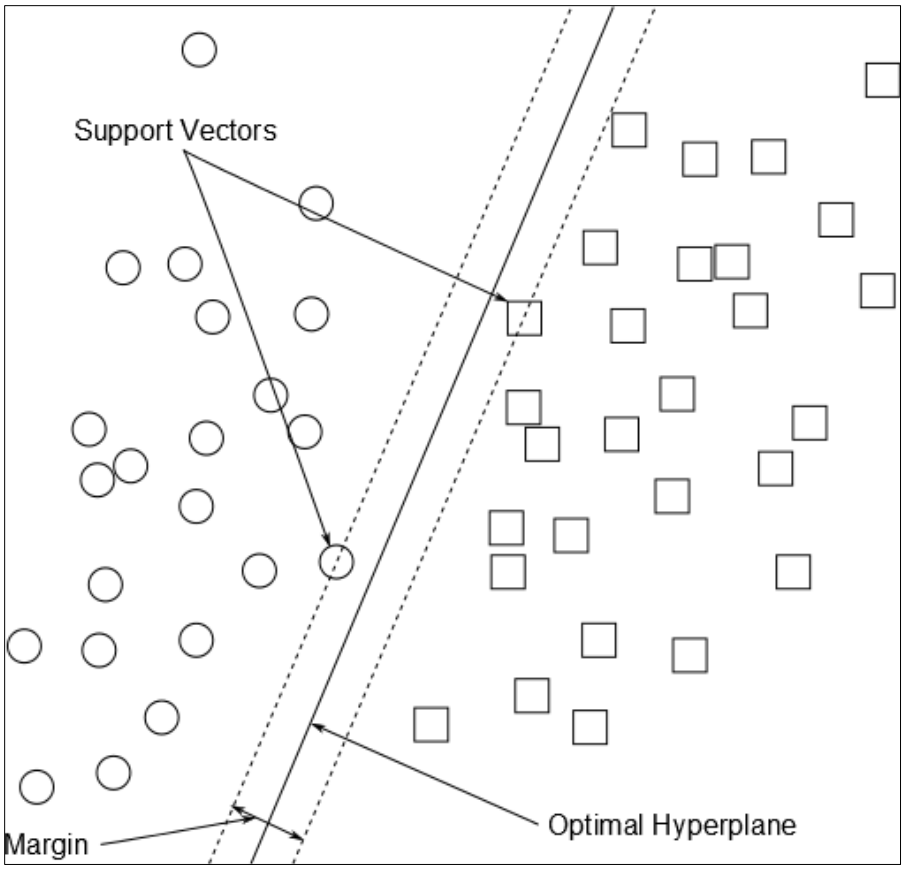


Figure 4-10: Data separation via support vectors.

Support vector data description uses the same tools as support vector machines with the exception that, instead of a hyperplane to separate pre-existing classes, it fits a hypersphere around the one-class set. The classification is a simple accordance test to determine whether the new object lies within the hypersphere or not. When applied to machine condition monitoring, data lying outside the hypersphere is considered novel (and potentially faulty). The support vector classifier used a Gaussian kernel and smoothing parameter of 5 based on the default parameters set within the data description toolbox.

Chapter 5

5 Fault Detection Results

This section evaluated the performance of the fault detection system when trained and tested on data collected at the Rail-Veyor® demonstration site. It investigated false positive and false negative error rates of the classifier validation during the testing stage. Performance was regarded as the lack of error throughout this section.

The objective of analyzing the fault detection results was to expose the best system configuration: the optimal combination of feature set and classifier for fault detection. Furthermore, the need to strike a balance between false positive and false negative errors was critical. The goal was to control and minimize the number of false alarms while maintaining the ability to detect faults. This goal was the primary justification to not over fit or over generalize the fault detection system.

This was not an exercise in classifier optimization for practical reasons. Optimization of the classifier configuration to the data set would have specialized it for that specific segment and would not necessarily have been the best optimization for all segments.

A critical parameter in selecting the best system configuration was a low false negative (false alarm) error rate. False alarms were considered detrimental to the fault detection system since they could cause the operator to become desensitized to the alarms and lose confidence in the system. For example, if the system warned the operator of damage and, upon inspection, the track was healthy, the operator might ignore the next warning.

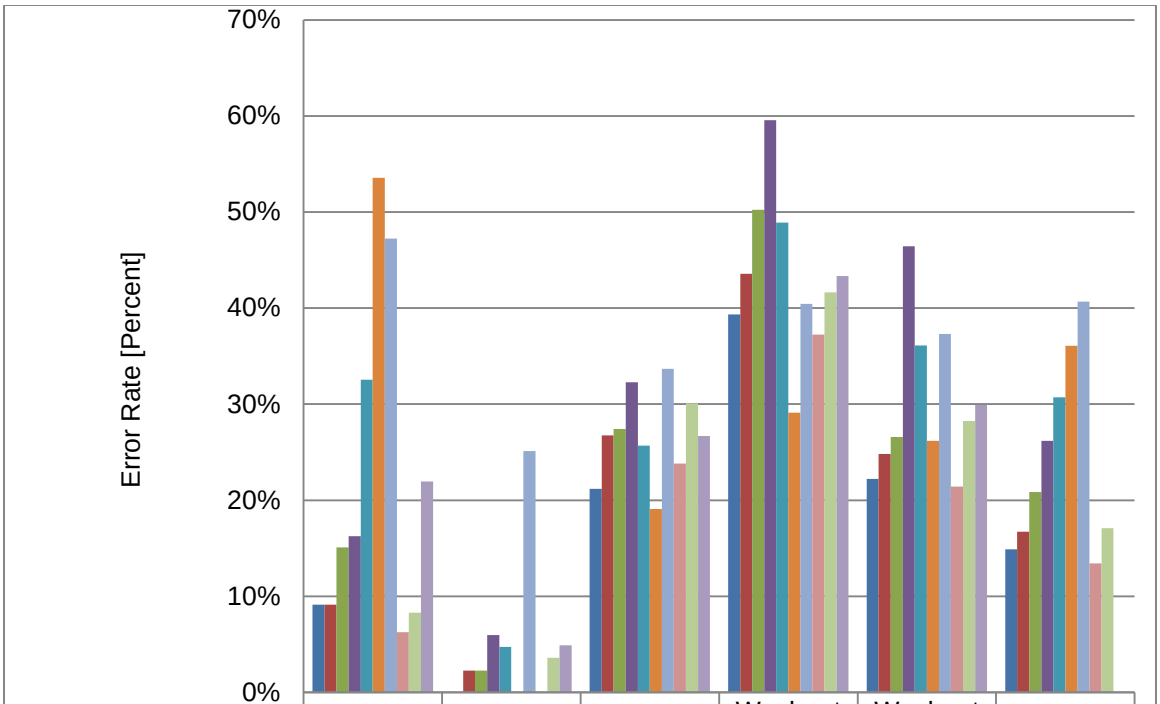
Each feature set-classifier combination was investigated to determine the best overall configuration. This resulted in 63 unique permutations of feature set-classifier combination.

The following sections report on the best configurations of feature set and classifier in a number of ways. Section 5.1 investigates the best overall feature set performance. Section 5.2 shows the fault detection results categorized by classifier to show best overall classifier performance. Section 5.3 explores the best overall configuration of feature set and classifier in an absolute lowest error sense. Section 5.4 examines the result of a more generalized segmentation approach, as discussed in section 4.2.2. Section 5.5 discusses the deviation of classification results if the training process is repeated.

5.1 Comparing Feature Set Performance

Figure 5-1 presents fault detection results for different feature sets graphically and in tabulated form. Classifier performance was averaged across each feature set to illustrate overall performance of individual feature sets. Categorized conditions present error rates based on incorrect classification and revealed if certain faults were more detectable than others.

The best performance was from the transient feature set. It exhibited the lowest total error rate of 13 percent; the full Statistical feature set also performed well with a total error rate of 15 percent. The Statistical with PCA and Statistical of AR100 feature sets also performed well, each with an error rate of 17 percent. These feature sets also had a low false negative error, which was critical in this investigation; however, all of these feature sets exhibited higher false positive rates for three of four fault conditions, meaning they failed to detect faults.



	False Negative	Horizontal Split Head	Loose Fishplate	Washout with Crack (30 mm)	Washout with Crack (60 mm)	Overall Average
Statistical	9%	0%	21%	39%	22%	15%
Statistical with PCA	9%	2%	27%	44%	25%	17%
AR10	15%	2%	27%	50%	27%	21%
AR20	16%	6%	32%	60%	46%	26%
AR50	33%	5%	26%	49%	36%	31%
AR100	54%	0%	19%	29%	26%	36%
AR200	47%	25%	34%	40%	37%	41%
Transient	6%	0%	24%	37%	21%	13%
Statistical of AR100	8%	4%	30%	42%	28%	17%
Average	22%	5%	27%	43%	30%	

Figure 5-1: Fault Detection Results – Comparing Feature Sets

Statistical and transient feature sets exhibited the lowest false negative error rates in this investigation. It was concluded that high false negative error rates exhibited by the autoregressive feature sets were due to the stochastic nature of the system response paired with environmental and operational noise; reconstruction of random fluctuations in the signal seemed

to hinder these feature sets. The statistical and transient feature sets were more tolerant to noise since the features were based on overall properties of the signal rather than reconstruction.

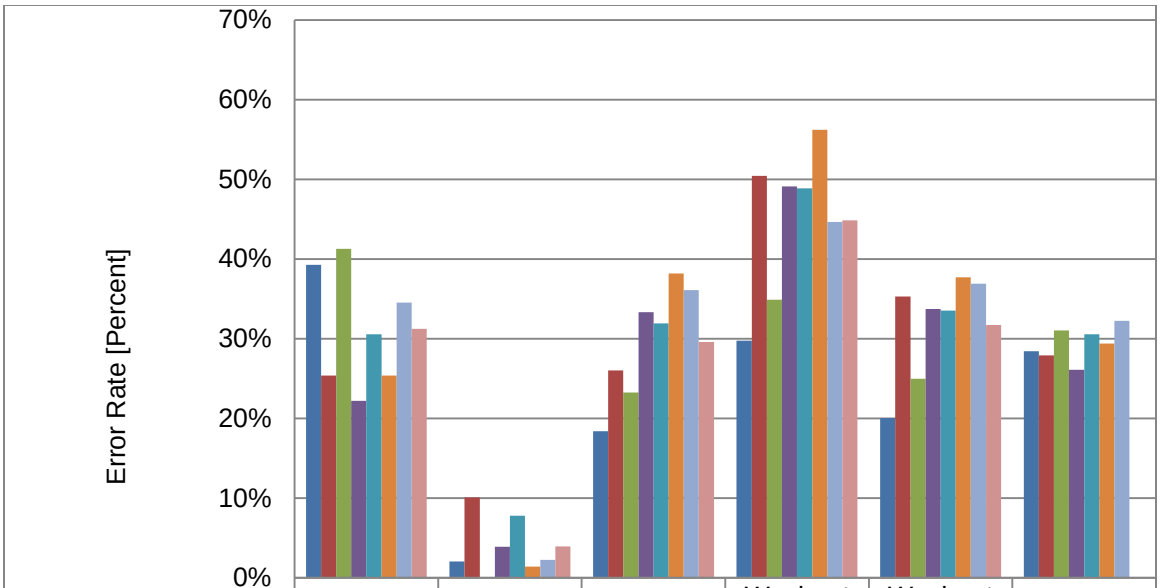
Certain faults were more detectable than others. Horizontally split head faults were often correctly identified; they were detected 95% of the time, on average. Figure 3-23 illustrated severity of the horizontally split head response in comparison to the healthy response and was attributed to the high likelihood of fault detection. Loose fishplates and the washout with 60 mm crack were detected nearly one in four times, which was also attributed to their damage severity relative to the healthy response. The less drastic washout (with 30 mm crack) was not consistently detectable, regardless of feature set selection.

5.2 Comparing Classifier Performance

It is generally accepted that the task of a classifier is rendered trivial in the case of intelligent and data-rich feature selection. This section presents a comparison of classification results from different classification approaches.

The role of classifiers is the same, irrespective of the choice of classifier; they are used to identify membership of new observations based on a model. Therefore, classification results across the range of fault conditions should be similar among the different classifiers. Figure 5-2 illustrates this concept by demonstrating similar error margins for each condition.

Fault detection results of overall classifier performance indicated that no single classifier could successfully identify faults significantly better than any other in this application; however, the results support that different faults were more detectable than others were.



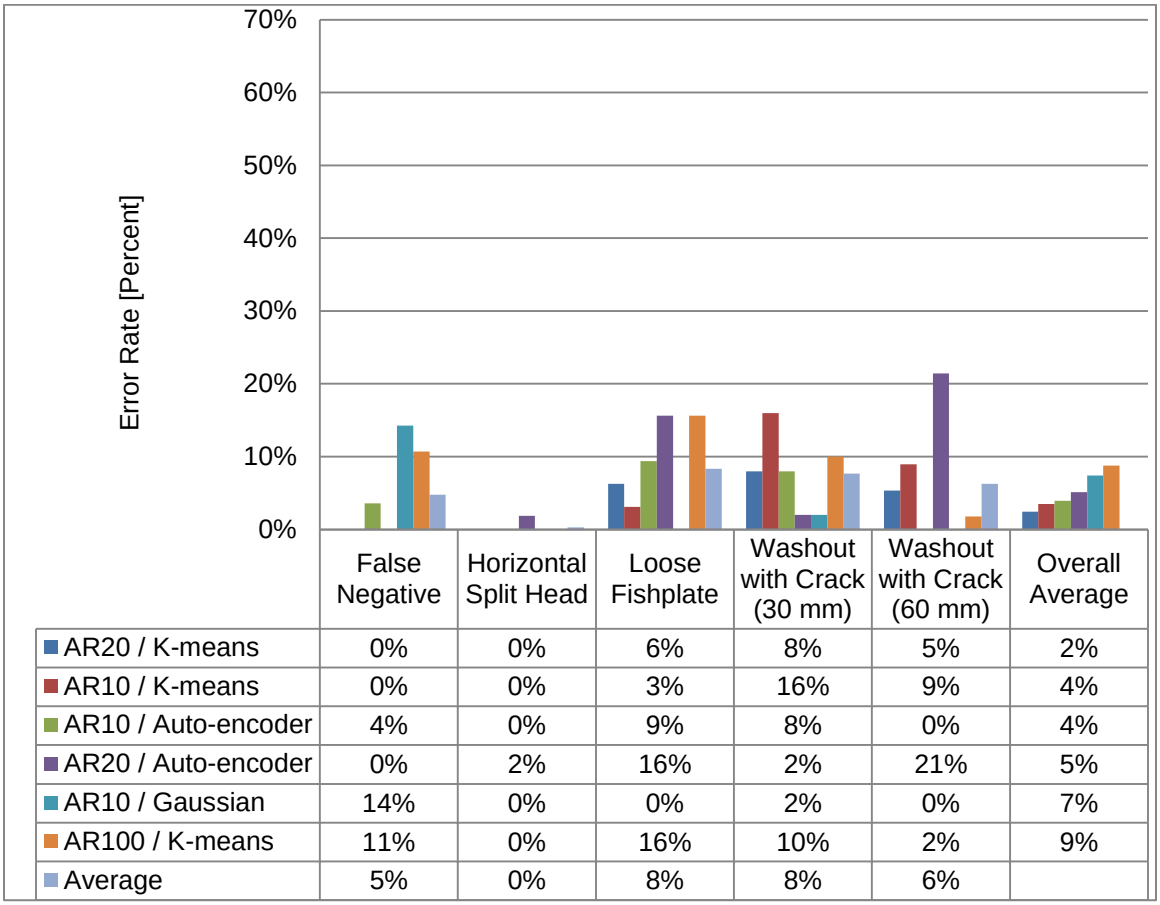
	False Negative	Horizontal Split Head	Loose Fishplate	Washout with Crack (30 mm)	Washout with Crack (60 mm)	Overall Average
Gaussian	39%	2%	18%	30%	20%	28%
PCA	25%	10%	26%	50%	35%	28%
K-means	41%	0%	23%	35%	25%	31%
Nearest Neighbour	22%	4%	33%	49%	34%	26%
Support Vector	31%	8%	32%	49%	34%	31%
Auto-encoder	25%	1%	38%	56%	38%	29%
Parzen	35%	2%	36%	45%	37%	32%
Average	31%	4%	30%	45%	32%	

Figure 5-2: Fault Detection Results – Comparing Classifiers

5.3 Best Overall Configuration of Feature Set and Classifier

Figure 5-3 presents the six best overall configurations of feature set and classifier. Each of the configurations yielded low overall error. The AR20 combined with the K-means classifier resulted in zero false alarms (false negatives) and always detected the horizontally split head faults. Roughly, one in twenty loose plates or washouts went undetected in this configuration. The AR10 feature set paired with the K-means or auto-encoder classifier also performed with excellent results; both exhibited only four percent error. In this case, the trade-off between false

negatives and false positives was demonstrated. The K-means was a better fit to the data, resulting in fewer false alarms; however, faults went undetected more often than the auto-encoder. The auto-encoder has a false alarm rate of four percent, versus zero with the K-means.



	False Negative	Horizontal Split Head	Loose Fishplate	Washout with Crack (30 mm)	Washout with Crack (60 mm)	Overall Average
AR20 / K-means	0%	0%	6%	8%	5%	2%
AR10 / K-means	0%	0%	3%	16%	9%	4%
AR10 / Auto-encoder	4%	0%	9%	8%	0%	4%
AR20 / Auto-encoder	0%	2%	16%	2%	21%	5%
AR10 / Gaussian	14%	0%	0%	2%	0%	7%
AR100 / K-means	11%	0%	16%	10%	2%	9%
Average	5%	0%	8%	8%	6%	

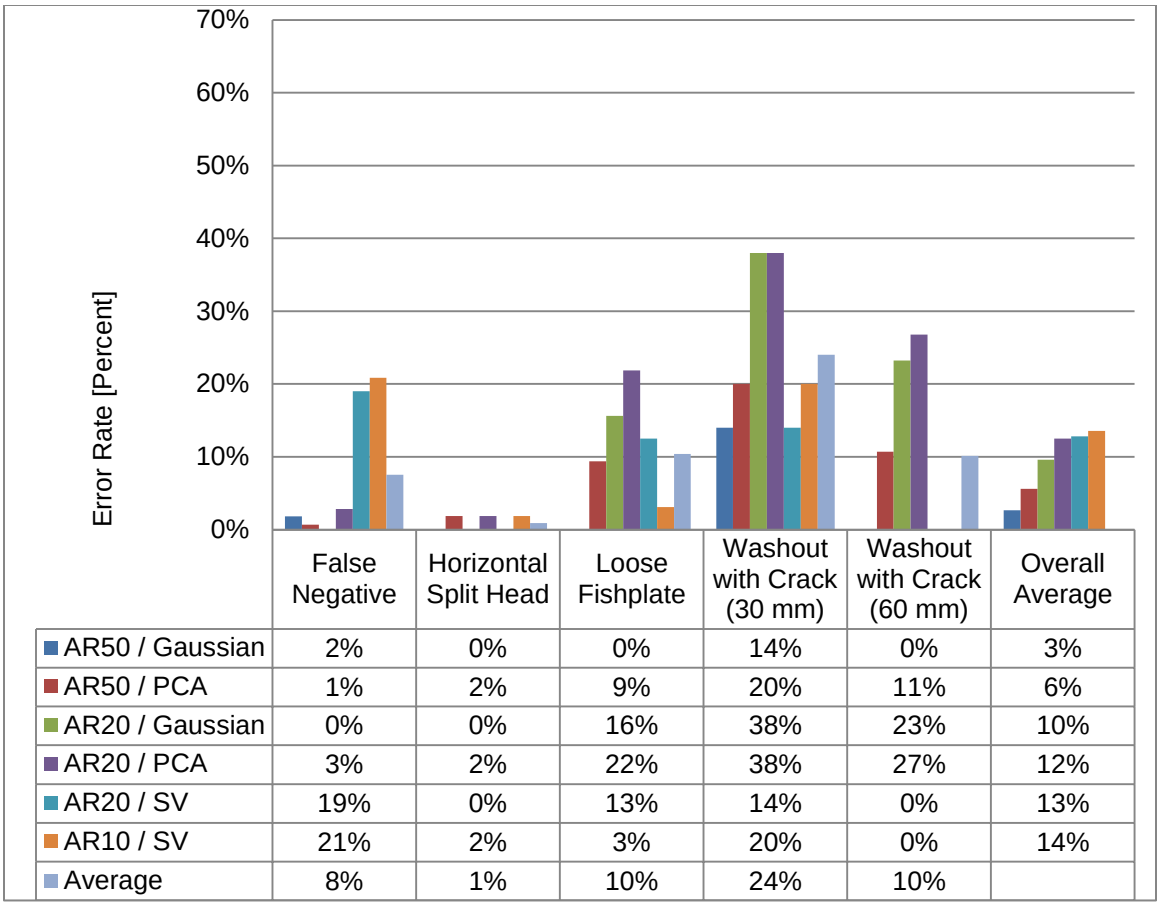
Figure 5-3: Fault Detection Results – Six Best Overall Configurations Using Specialized Segmentation

The overall trend of the results in Figure 5-3 showed the low-order AR feature set performing the best when presented individually, rather than averaged over classifier or feature set performance as a whole.

5.4 Performance Using a Quasi-Generalized Segmentation Approach

In this section, fault detection results were based on a more generalized segmentation approach using the straight section of track leading up to the fault location. This approach was investigated to determine the feasibility of this segmentation approach. The motivation for this study was the ability to acquire a large data set in much less time than the direct comparison (specialized) approach.

Figure 5-4 made it evident that the quasi-generalized segmentation approach could work well for fault detection in this application. The AR50 feature set with a Gaussian density-based classifier performed exceptionally, with an overall error rate of three percent. This configuration also showed very few false negatives, with a two percent error rate; the configuration resulted in fault detection of every fault except the washout with a small crack. Overall, this configuration was very sensitive to damage and resilient to operational variation.



	False Negative	Horizontal Split Head	Loose Fishplate	Washout with Crack (30 mm)	Washout with Crack (60 mm)	Overall Average
AR50 / Gaussian	2%	0%	0%	14%	0%	3%
AR50 / PCA	1%	2%	9%	20%	11%	6%
AR20 / Gaussian	0%	0%	16%	38%	23%	10%
AR20 / PCA	3%	2%	22%	38%	27%	12%
AR20 / SV	19%	0%	13%	14%	0%	13%
AR10 / SV	21%	2%	3%	20%	0%	14%
Average	8%	1%	10%	24%	10%	

Figure 5-4: Fault Detection Results – Six Best Overall Configurations Using Quasi-Generalized Segmentation

5.5 Mean Absolute Deviation of Classification Results

To determine if classification results were reproducible, the mean absolute deviation was taken across 15 training iterations. The mean absolute deviation represented the average deviation of the fault detection results if the classifiers were retrained on a new training set. Therefore, a low mean absolute deviation over 15 training iterations meant that the training stage was reproducible. The results of this investigation are presented in Table 5-1.

Overall, mean absolute deviations were nearly zero across all classifications. Low mean absolute deviations exhibited by classifiers across multiple training iterations were a promising indication that classification results were reproducible.

Table 5-1: Mean absolute deviation over 15 training iterations.

	Mean Absolute Deviation				
Classifier	Healthy	Horizontal Split Head	Loose Fishplate	Washout w/ Crack (30 mm)	Washout w/ Crack (60 mm)
Gaussian DD	0%	0%	0%	0%	0%
Auto-encoder DD	0%	0%	3%	0%	1%
Parzen DD	0%	0%	0%	0%	0%
K-means DD	0%	0%	1%	1%	1%
PCA DD	0%	0%	0%	0%	0%
Nearest Neighbour DD	0%	0%	0%	0%	0%
Support Vector DD	0%	0%	0%	0%	0%

Chapter 6

6 Conclusions and Future Work

6.1 The Problem Identification

The problem of damage detection in railway systems was investigated. It was determined that there was an industrial need in terms of modular function. Existing technologies use sophisticated techniques that require downtime to perform inspection. Furthermore, the need for a relatively low-cost industrial condition monitoring system that can be adapted to existing equipment was proposed.

6.2 Experimental Observations

During the data acquisition phase, transducer measurements were monitored and inspected. The adequate selection of instrumentation is imperative when developing a condition monitoring system. Without proper mode classification to establish healthy baseline data, the task of condition monitoring can become daunting.

6.3 Segmentation Approach Conclusions

The choice of segmentation approach proved to have a significant effect on the best overall configuration of feature set and classifier. Two segmentation approaches were investigated: a specialized approach that only compared exact locations along the track to each other, and, a more generalized approach that compared seemingly similar sections of track for fault detection. The more specialized approach exhibited better overall performance compared to the more

generalized approach; however, each approach had its advantages and disadvantages at a more practical level.

Implementation of a specialized segmentation approach would be more involved since precise localization is needed, and it would require a longer period of time in which training data is gathered; however, the high performance of fault detection associated with the findings of this study is the direct benefit to this approach. The more generalized segments did not perform as well as the specialized segments, but the implementation of this approach is more straightforward in terms of commissioning since accurate localization is not necessary.

From the results in this study, both the specialized segmentation approach and the quasi-generalized segmentation approach demonstrated excellent performance and seem to be feasible solutions for an industrial implementation of a vibration-based rail fault detection system.

6.4 Feature Selection Conclusions

Fault detection performance is highly linked to intelligent feature selection and optimization. Features that accurately characterize the data set are vital to successful fault detection.

When using the direct comparison segmentation approach, the transient and statistical features performed the best on average over all classifiers. These feature sets accommodated the small data set available when using the direct comparison segmentation approach. The transient nature of the fault signals was captured in these feature sets and not in the AR sets. The stochastic fluctuations compounded with a small data set caused the regressions to over fit the data set and generate a high rate of false negative errors (false alarms).

An important aspect of fault detector performance was resilience to false alarms. In practice, it would be essential for the fault detection system to perform well and not raise any false alarms. While the statistical and transient feature sets excelled in performance using the specialized segmentation approach, the AR feature sets performed better with the larger data set and signal variance of the quasi-generalized segmentation approach. Although the statistical and transient feature sets exhibited the lowest false negative error rates, they could not detect faults as efficiently as the AR sets with comparable false negative error rates.

Using a more generalized segmentation approach converged to an AR50 feature set as the best feature set on average over all of the classifiers. The results conveyed that the more generalized segmentation approach produced a larger data set with a higher variance than in the direct comparison approach to segmentation. The variance was handled well by the AR50, AR20, and AR100 feature sets.

The best feature set on average was not the best overall feature set; the overall best configurations used the AR20 and AR10 feature sets for specialized segmentation and the AR50 feature set for the quasi-generalized segmentation. Since the above conclusions present the best feature sets on average over all of the classifiers, they overlook cases where particular feature sets work better with specific classifiers. This was the case when the results were arranged by best performance before averaging classifier performance. These conclusions contradict the notion that the best feature set on average and the best classifier on average would produce the best configuration.

The performance trend as AR order increases had an interesting outcome: a balance between the order of the AR model and the benefit of additional features presents itself. That is, in some cases, a higher order AR model results in reduced performance since the additional features cause the classifier to over fit, while in other cases, the lower order AR models result in reduced performance since they do not adequately describe the dataset. Therefore, an AR model with a specific order may be the optimal choice; however, it was also shown that the classifier affects the performance results since the K-means favours the AR20 feature set, while the auto-encoder performs better with the AR10 set.

Feature vector length has an effect on overall classification performance; the curse of dimensionality is a prime example of why larger feature sets result more false alarms (false negatives) since the classifier model becomes over fit to the data set. An analysis of feature vector size comparison was outside the scope of this research, but is definitely a subject to be considered for future work.

6.5 Classifier Conclusions

A similar paradox was present in classifier performance, in which the best classifier on average over the feature sets was not the classifier used in the overall best configuration. In the specialized segmentation (direct comparison) approach, the nearest neighbour classifier had the lowest overall error, while the Parzen classifier produced the lowest error on average for the quasi-generalized segments. The K-means and Gaussian classifiers were used to generate the lowest overall error when combined with the AR feature sets in the specialized and quasi-generalized segments, respectively. The K-means classifier and AR20 feature set were the best

fault detectors for the specialized segment. The Gaussian classifier and AR50 feature set performed the best using the quasi-generalized segment.

6.6 Fault Detection Conclusions

The fault detection results show that specific feature set and classifier combinations could be used to identify faults using time series vibration signals. The results also indicate that the best overall configuration for fault detection was the AR20 feature set with a K-means classifier, when using the direct comparison segmentation approach. The best configuration for a more generalized segmentation approach used the AR50 feature set with a Gaussian classifier. These results form a paradox in that the best feature set on average, combined with the best classifier on average did not yield the best performance. Furthermore, they illustrate that the best feature set and classifier combination varies with the segmentation approach.

The fault detection results of this study revealed that successful implementation of a vibration-based fault detection system was feasible. The presence of a horizontally split head fault amongst the railway infrastructure was highly detectable in this analysis. Furthermore, loose fishplates and washouts proved to be faults in their early stages, and were not consistently detected; however, the ability to have a variable segment size may be a potential solution for more consistent detection of these faults. The results of this study are encouraging when considering the potential for its application in the railway industry.

6.7 Industrial Implementation

Industrial implementation of the fault detection system developed in this study would follow a number of commissioning phases. First, the train would be instrumented with vibration, speed, and indexing sensors as well as a data acquisition and processing system. The drive stations would require a trigger mechanism for the indexing sensors (i.e., a magnet located in a known location) so that the monitoring system has reference locations for dead reckoning. The system would then need to be put into training mode for a period while the system operates under normal conditions, gathering a feature vector data set. Once the monitoring system has a sufficient history of data, the classifier would be trained and begin to test new observations. If the fault detector begins to notify the operator of faults and no faults are present, the system should be retrained with the accumulated data. Therefore, as long as the train operates under no-fault conditions, the classifier can be retrained on the history of data, refining the classifier performance. If the system fails to detect faults, the training set would need to be rolled back to a period before the development of the fault or the data from that section of track could be omitted; each method would require additional metadata, namely a timestamp or a location reference, respectively. This method of implementation would also mitigate changes due to maintenance (track sections are replaced due to wear), seasonal changes (thermal expansion and contraction), and other gradual changes in system response since the operator can notify the monitoring system that everything has been operating properly, so it can retrain the classifier.

6.8 Future Work

The encouraging results from this research suggest that future work in this field of study be continued. The successful implementation of a rail monitoring system as described in this

research shows significant potential for economic and environmental savings, while the opportunity for safeguarding life remains invaluable.

6.8.1 Further Testing on Industrial Case Study

Data availability in condition monitoring applications is fundamental to a successful system. As such, under the conditions of this research, a database of historical fault data would have been extremely valuable. Therefore, data collection for condition monitoring candidates would be very useful in terms of research and development. A natural next step for this research is to develop a prototype system in order to field test the proposed general methodology at the test site. A possible intermediate step could involve the off-line testing of the system on previously collected fault data collected during actual operation. This type of testing would demonstrate the ability of the system to detect faults as well as fault progression that occurs during normal operation.

6.8.2 Localization of Cars on Track

Having an accurate, reliable, and easy to implement method of locating the train on the tracks would be of great value to a fault detection system. This data would be useful for training the fault detection system on specific track sections as well as locating specific faults. The task of localization for an industrial implementation remains quite difficult to implement. The problem would be further exacerbated in the underground mining case where wireless and GPS technologies are difficult to implement. One approach would be to integrate existing technologies such as GPS, inertial measurement unit (IMU), or dead reckoning, to accomplish the localization task.

6.8.3 Testing and Development of Variations of Proposed Methodology

This investigation was a proof of concept rather than an actual deployment of an optimized fault detection system. The conclusions of the investigation undoubtedly revealed the wide range of possible configurations and their associated performance. Moreover, the development of this fault detector by no means exhausted all potential methods for fault detection. Many other techniques such as sensor data fusion (the addition of other sensing techniques to the data set), the use of hybrid feature sets (combining different types of features into one feature set), the testing of other classifiers, and combinational classification could all enhance the fault detection performance. Future work would require investigation into these areas and optimization of the fault detector for full-scale deployment.

6.8.4 Novelty Detection

All of the proposed classification approaches discussed in this study fall under the category of novelty detection or outlier detection. The goal of novelty detection is to devise a means to detect new behaviour or novel events. Therefore, the function of novelty detection in relation to this work was solely based on comparing new elements to the healthy set. Future work might extend this idea of novelty detection to learn about the behaviour of different failure modes. Insight into probability density functions, boundary delimitations, or artificial neural network topology of specific failure modes could then be used to diagnose failures in addition to fault detection.

References

[1] M. Mortada. Applicability and interpretability of logical analysis of data in condition based maintenance. pp. 203. 2011.

[2] O. Krellis and T. Singleton. Mine maintenance-the cost of operation. 1998.

[3] Rail-Veyor® Technologies Global Inc. (2013, January 18). *Redefining Material Haulage* [Online]. Available: http://www.railveyor.com.

[4] S. W. Doebling, C. R. Farrar and M. B. Prime. A summary review of vibration-based damage identification methods. *Shock Vib Dig 30(2),* pp. 91-105. 1998.

[5] M. Mjit. Methodology for fault detection and diagnostics in an ocean turbine using vibration analysis and modeling. pp. 115. 2009.

[6] O. Salawu. Detection of structural damage through changes in frequency: A review. *Eng. Struct. 19(9),* pp. 718-723. 1997.

[7] K. R. Al-Balushi and B. Samanta. Gear fault diagnosis using energy-based features of acoustic emission signals. *Proceedings of the Institution of Mechanical Engineers.Part I: Journal of Systems and Control Engineering 216(3),* pp. 249. 2002.

[8] J. McBain. Condition monitoring of machinery subject to variable states: Monitoring of mobile underground mining equipment. (Doctoral dissertation). *ProQuest Dissertations and Theses* pp. 266. 2012.

[9] X. Li. A brief review: Acoustic emission method for tool wear monitoring during turning. *Int. J. Mach. Tools Manuf. 42(2),* pp. 157. 2002.

[10] Z. Sun and C. -. Chang. Vibration based structural health monitoring: Wavelet packet transform based solution. *Structure and Infrastructure Engineering 3(4),* pp. 313-323. 2007.

[11] Q. Ye, T. Wang and J. Ye. Vibration signal multi-wave-packets mode decomposition and application. *Dynamics of Continuous Discrete and Impulsive Systems-Series B-Applications & Algorithms 14*pp. 1374-1378. 2007.

[12] M. Antonopoulos-Domis and T. Tambouratzis. System identification during a transient via wavelet multiresolution analysis followed by spectral techniques. *Ann. Nucl. Energy 25(7),* pp. 465-480. 1998.

[13] S. Bhunia and K. Roy. A novel wavelet transform-based transient current analysis for fault detection and localization. *IEEE Transactions on very Large Scale Integration (VLSI) Systems 13(4),* pp. 503-7. 2005.

[14] A. Caprioli, A. Cigada and D. Raveglia. Rail inspection in track maintenance: A benchmark between the wavelet approach and the more conventional fourier analysis. *Mechanical Systems and Signal Processing 21(2),* pp. 631-652. 2007.

[15] B. Chen, X. Wang, S. Yang and C. McGreavy. Application of wavelets and neural networks to diagnostic system development, 1, feature extraction. *Comput. Chem. Eng. 23(7),* pp. 899-906. 1999.

[16] H. A. Toliyat, K. Abbaszadeh, M. M. Rahimian and L. E. Olson. Rail defect diagnosis using wavelet packet decomposition. *IEEE Trans. Ind. Appl. 39(5),* pp. 1454-1461. 2003.

[17] B. FjrIik, B. Czechyra and A. Chudzikiewicz. Condition monitoring system for light rail vehicle and track. *Key Eng Mat 518.* pp. 66. 2012.

[18] H. Tsunashima, T. Kojima, A. Matsummoto, and T. Mizuma. Condition monitoring of railway track using in-service vehicle. *Japanese Railway Engineering,* (161). pp. 333-356. 2008.

[19] M. Yang, K. A. Edge and D. N. Johnston. Condition monitoring and fault diagnosis for vane pumps using flow ripple measurement. Presented at Bath/ASME Symposium on Fluid Power and Motion Control 2008.

[20] M. A. Timusk. A unified method for anomaly detection in unsteady systems. (Doctoral dissertation). *ProQuest Dissertations and Theses* 2006.

[21] L. Chatfield. (2006, June 7). *Gone Fishing - A portrait of a fishplate on the Bluebell railway* [Online]. Available: http://www.flickr.com/photos/elsie/191124445/.

[22] Transportation Safety Board of Canada. (2013, February 1). *TSB Statistical Summary - Railway Statistics* [Online]. Available: http://www.tsb.gc.ca/eng/rail/index.asp.

[23] R. Clark. Rail flaw detection: Overview and needs for future developments. *NDT E Int. 37(2),* pp. 111-118. 2004.

[24] M. P. Papaelias, C. Roberts and C. Davis. A review on non-destructive evaluation of rails: State-of-the-art and future development. *Proc. Inst. Mech. Eng. Pt. F: J. Rail Rapid Transit 222(4),* pp. 367-384. 2008.

[25] D. Cannon, K. Edel, S. Grassie and K. Sawley. Rail defects: An overview. *Fatigue & Fracture of Engineering Materials & Structures 26(10),* pp. 865-886. 2003.

[26] U.S. Army Corps of Engineers. Unified Facilities Criteria (UFC) Railroad Track Maintenance and Safety Standards. *United States Department of Defence.* 2008.

[27] Campbell Scientific (2012). *Measurement and Control Products for Long-term Monitoring* [Online]. Available: http://www.campbellsci.ca/railway-monitoring.

[28] ESG Solutions. (2012). *Two Decades of Microseismic Monitoring* [Online]. Available: https://www.esgsolutions.com/english/view.asp?x=806.

[29] Innowattech. (2012). *Energy Harvesting Systems* [Online]. Available: http://www.innowattech.co.il/slnRailMonitoring.aspx.

[30] Struckton Rail. (2012). *A Higher Level of Reliability* [Online]. Available: http://www.struktonrail.com/systems/poss-online-monitoring/.

[31] R. Pohl, A. Erhard, H. Montag, H. Thomas and H. Wustenberg. NDT techniques for railroad wheel and gauge corner inspection. *NDT E Int. 37(2),* pp. 89-94. 2004.

[32] M. Timusk, M. Lipsett and C. K. Mechefske. Fault detection using transient machine signals. *Mechanical Systems and Signal Processing 22(7),* pp. 1724-1749. 2008.

[33] National Instruments. (2009, February 27). *Increase System Performance with New CompactRIO Offerings* [Online]. Available: http://www.ni.com/newsletter/50704/en/.

[34] NDT Resource Center (2012). *AE Signal Features* [Online]. Available: http://www.ndt-ed.org/EducationResources/CommunityCollege/Other%20Methods/AE/AE_Signal%20Features.htm.

[35] D. M. J. Tax. DDtools, the data description toolbox for MATLAB. 2012.

[36] University of Utah. (2007). *Parzen-Window Density Estimation*. [Online]. Available: https://www.cs.utah.edu/~suyash/Dissertation_html/node11.html.

[37] K. P. Bennett and C. Campbell. Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter 2(2),* pp. 1-13. 2000.

Appendix A – Fault Detection Results

Table 6-1: Fault Detection Results of a Specialized Segmentation Approach

Feature Set	Classifier	False Negative Error Rate	False Positive Error Rate				Overall Average Error Rate
			Horizontally Split Head	Loose Fishplate	Washout with Crack (30 mm)	Washout with Crack (60 mm)	
AR20	K-Means	0%	0%	6%	8%	5%	2%
AR10	K-Means	0%	0%	3%	16%	9%	4%
AR10	Auto-encoder	4%	0%	9%	8%	0%	4%
AR20	Auto-encoder	0%	2%	16%	2%	21%	5%
AR50	K-Means	5%	0%	0%	24%	4%	6%
AR10	Gaussian	14%	0%	0%	2%	0%	7%
AR100	K-Means	11%	0%	16%	10%	2%	9%
AR20	Support Vector	11%	2%	9%	20%	18%	11%
AR50	Support Vector	14%	0%	0%	36%	7%	13%
AR50	Auto-encoder	4%	4%	22%	38%	25%	13%
AR100	Auto-encoder	7%	13%	34%	12%	21%	14%
Statistical AR100	K-Means	0%	0%	50%	52%	14%	15%
AR10	PCA	0%	0%	3%	46%	75%	16%
AR200	Support Vector	29%	0%	6%	4%	4%	16%
AR200	K-Means	21%	0%	19%	20%	7%	16%
AR10	Support Vector	14%	6%	19%	32%	23%	17%
AR20	PCA	36%	0%	0%	2%	0%	18%
Statistical AR100	Auto-encoder	0%	6%	53%	78%	11%	18%
Statistical PCA	PCA	4%	0%	9%	84%	45%	19%
Statistical PCA	Auto-encoder	14%	0%	13%	56%	38%	20%
AR20	Nearest Neighbour	4%	22%	9%	48%	79%	22%
Statistical AR100	Gaussian	14%	0%	41%	66%	9%	22%
Statistical	PCA	18%	0%	3%	82%	21%	22%
AR50	Nearest Neighbour	0%	26%	31%	74%	50%	23%
AR200	Auto-encoder	4%	13%	59%	74%	21%	23%
Statistical PCA	K-Means	18%	0%	6%	74%	34%	23%
AR100	Support Vector	43%	0%	6%	14%	0%	24%
AR100	Nearest Neighbour	0%	20%	25%	84%	64%	24%
Statistical AR100	Support Vector	0%	7%	56%	80%	52%	24%
Statistical PCA	Gaussian	25%	0%	3%	62%	30%	24%
Statistical	K-Means	25%	0%	6%	64%	32%	25%
Statistical	Gaussian	32%	0%	0%	56%	29%	27%

Feature Set	Classifier	False Negative Error Rate	False Positive Error Rate				Overall Average Error Rate
			Horizontally Split Head	Loose Fishplate	Washout with Crack (30 mm)	Washout with Crack (60 mm)	
Statistical	Nearest Neighbour	0%	0%	75%	92%	55%	28%
Statistical	Support Vector	11%	0%	31%	90%	68%	29%
Statistical PCA	Support Vector	18%	0%	22%	88%	61%	30%
Statistical	Auto-encoder	25%	0%	9%	72%	66%	31%
Statistical PCA	Nearest Neighbour	4%	0%	72%	90%	75%	31%
AR10	Parzen	71%	0%	0%	0%	0%	36%
AR20	Gaussian	71%	0%	0%	0%	0%	36%
AR20	Parzen	71%	0%	0%	0%	0%	36%
AR50	Gaussian	71%	0%	0%	0%	0%	36%
AR50	Parzen	71%	0%	0%	0%	0%	36%
AR50	PCA	71%	0%	0%	0%	0%	36%
AR100	Gaussian	71%	0%	0%	0%	0%	36%
AR100	Parzen	71%	0%	0%	0%	0%	36%
AR100	PCA	71%	0%	0%	0%	0%	36%
AR200	Gaussian	71%	0%	0%	0%	0%	36%
AR200	Parzen	71%	0%	0%	0%	0%	36%
AR200	PCA	71%	0%	0%	0%	0%	36%
Statistical AR100	Parzen	71%	0%	0%	0%	0%	36%
Transient	Nearest Neighbour	0%	0%	100%	100%	89%	36%
Statistical	Parzen	75%	0%	0%	0%	0%	38%
Transient	Gaussian	4%	0%	94%	98%	96%	38%
Statistical PCA	Parzen	75%	0%	0%	10%	2%	39%
Transient	PCA	4%	2%	97%	100%	100%	39%
AR200	Nearest Neighbour	0%	63%	84%	90%	86%	40%
AR10	Nearest Neighbour	0%	96%	53%	94%	93%	42%
Transient	Support Vector	14%	0%	97%	96%	91%	43%
Transient	K-Means	25%	0%	69%	98%	84%	44%
Statistical AR100	Nearest Neighbour	0%	67%	94%	98%	96%	44%
Transient	Auto-encoder	25%	0%	91%	100%	95%	48%
Transient	Parzen	61%	0%	72%	72%	59%	56%

Table 6-2: Fault Detection Results of a Quasi-generalized Segment

Feature Set	Classifier	False Negative Error Rate	False Positive Error Rate				Overall Average
			Horizontally Split Head	Loose Fishplate	Washout with Crack (30 mm)	Washout with Crack (60 mm)	
AR50	Gaussian	2%	0%	0%	14%	0%	3%
AR50	PCA	1%	2%	9%	20%	11%	6%
AR20	Gaussian	0%	0%	16%	38%	23%	10%
AR20	PCA	3%	2%	22%	38%	27%	12%
AR20	Support Vector	19%	0%	13%	14%	0%	13%
AR10	Support Vector	21%	2%	3%	20%	0%	14%
AR100	PCA	23%	0%	13%	6%	0%	14%
AR100	Support Vector	27%	0%	9%	4%	0%	15%
AR50	Support Vector	30%	0%	0%	6%	0%	16%
AR100	Gaussian	30%	0%	9%	2%	0%	16%
Statistical AR100	Support Vector	19%	2%	44%	14%	5%	18%
AR20	Auto-encoder	1%	4%	31%	52%	55%	18%
AR50	Auto-encoder	2%	4%	16%	58%	70%	20%
AR10	Gaussian	4%	2%	19%	66%	66%	21%
AR200	Support Vector	35%	0%	22%	4%	0%	21%
AR50	K-Means	1%	7%	13%	74%	75%	22%
AR100	Auto-encoder	1%	17%	38%	70%	50%	22%
Statistical	Support Vector	30%	0%	19%	22%	23%	23%
Statistical PCA	Support Vector	30%	0%	13%	36%	20%	23%
AR10	Nearest Neighbour	23%	4%	25%	46%	32%	25%
AR20	K-Means	0%	6%	38%	80%	95%	27%
Statistical	Gaussian	0%	0%	38%	98%	86%	28%
Statistical PCA	Gaussian	0%	0%	38%	98%	88%	28%
Statistical PCA	K-Means	0%	0%	44%	94%	84%	28%
Statistical PCA	Nearest Neighbour	7%	0%	56%	80%	68%	29%
AR200	PCA	55%	0%	9%	0%	0%	29%
AR100	K-Means	0%	24%	53%	84%	73%	29%
Statistical	PCA	1%	0%	38%	100%	93%	29%
Statistical	K-Means	0%	0%	50%	98%	86%	29%
Statistical PCA	Auto-encoder	0%	0%	38%	100%	96%	29%
Statistical AR100	K-Means	0%	28%	72%	60%	79%	30%
Statistical PCA	PCA	1%	4%	31%	100%	100%	30%
Statistical	Auto-encoder	1%	0%	50%	98%	91%	30%
Statistical	Nearest Neighbour	7%	0%	56%	86%	71%	30%
AR10	PCA	2%	4%	41%	98%	95%	31%
AR10	K-Means	3%	7%	63%	80%	84%	31%

Feature Set	Classifier	False Negative Error Rate	False Positive Error Rate				Overall Average
			Horizontally Split Head	Loose Fishplate	Washout with Crack (30 mm)	Washout with Crack (60 mm)	
AR10	Auto-encoder	0%	33%	47%	90%	95%	33%
AR20	Nearest Neighbour	48%	4%	13%	38%	38%	35%
Statistical AR100	Nearest Neighbour	8%	28%	69%	82%	77%	36%
Transient	K-Means	1%	0%	97%	98%	100%	37%
Transient	Parzen	5%	0%	94%	94%	96%	38%
Transient	Auto-encoder	1%	0%	100%	100%	100%	38%
Transient	Gaussian	1%	0%	100%	100%	100%	38%
Transient	PCA	1%	2%	100%	100%	100%	38%
Transient	Support Vector	38%	0%	47%	60%	50%	39%
AR200	Gaussian	77%	0%	3%	0%	0%	39%
Statistical AR100	Auto-encoder	2%	67%	81%	72%	89%	40%
AR200	Auto-encoder	1%	67%	72%	90%	88%	40%
Transient	Nearest Neighbour	3%	31%	94%	98%	91%	41%
AR200	K-Means	0%	63%	84%	96%	88%	41%
Statistical PCA	Parzen	89%	0%	0%	0%	0%	45%
AR50	Nearest Neighbour	81%	6%	6%	22%	9%	46%
AR100	Nearest Neighbour	85%	2%	9%	10%	9%	46%
Statistical AR100	PCA	3%	81%	91%	100%	100%	48%
Statistical	Parzen	98%	0%	0%	0%	0%	49%
AR200	Nearest Neighbour	96%	0%	6%	0%	0%	49%
AR10	Parzen	100%	0%	0%	0%	0%	50%
AR20	Parzen	100%	0%	0%	0%	0%	50%
AR50	Parzen	100%	0%	0%	0%	0%	50%
AR100	Parzen	100%	0%	0%	0%	0%	50%
AR200	Parzen	100%	0%	0%	0%	0%	50%
Statistical AR100	Parzen	100%	0%	0%	0%	0%	50%

Appendix B – Additional Investigations

The following sections outline commonly used preprocessing techniques in the field of condition monitoring, relevant to the current research.

Fourier Analysis

Frequency domain analysis is a common technique used in condition monitoring to identify notable frequency components. One common method used to generate a frequency spectrum is the fast Fourier transform (FFT). The FFT is an algorithm used to produce the discrete Fourier transform in a computationally-efficient manner, which is defined by,

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k\frac{n}{N}}, \qquad k = 0, \dots, N-1, \qquad \text{Equation 6-1}$$

where X_k is the sequence of transformed N-periodic complex numbers, and x is a finite set of complex numbers. The FFT algorithm used in this section was the conventional Cooley-Tukey algorithm. The FFTs in this study were computed using MATLAB.

To analyze the time series data in the frequency domain, a suitable window size was first selected. Since the train speed was controlled at approximately 3 m/s, the fundamental wheel rotating frequency, F_w, should be observed at,

$$F_w = \frac{s}{C}$$

$$= \frac{3\ [m]}{[s]} \cdot \frac{1}{\pi \cdot d\ [m]} \qquad \text{Equation 6-2}$$

$$= 2.35\ Hz,$$

where the s is the train speed, and C is the train wheel circumference. Next, the sampling rate and signal length are entered as inputs for the FFT algorithm.

Figure 6-1 shows time series vibration data and its corresponding frequency spectrum under healthy operating conditions.

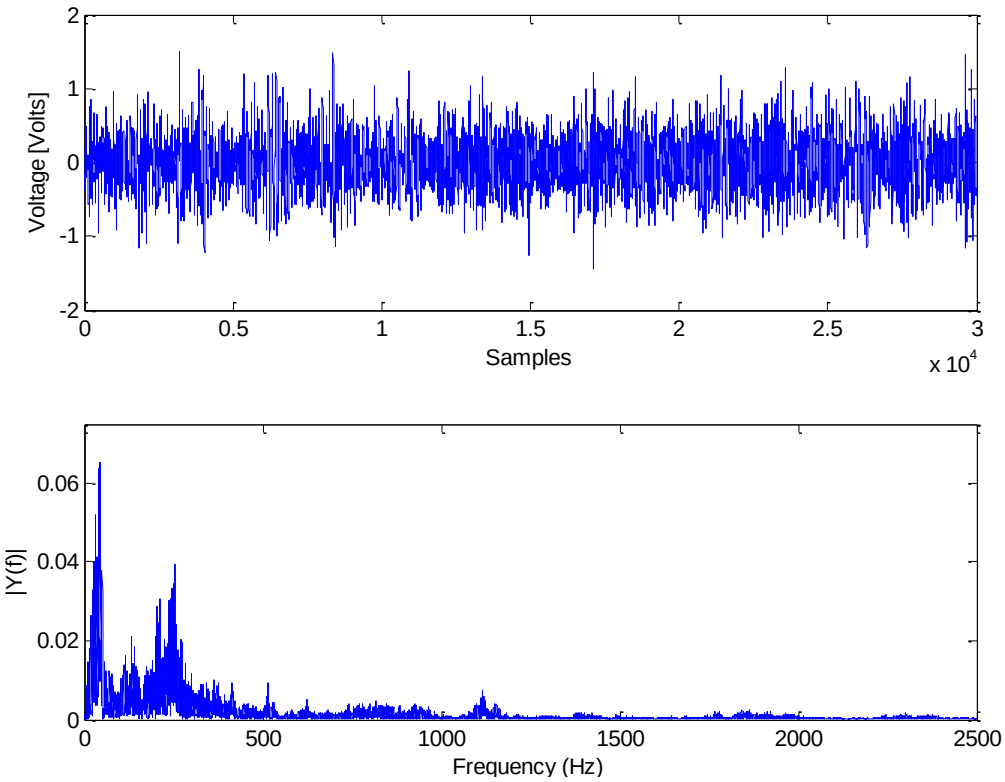


Figure 6-1: Top: Raw vertical accelerometer data in healthy conditions; Bottom: The FFT of the healthy accelerometer signal above.

Comparing these plots to faulted transducer data and frequency spectra, it became evident that little periodicity existed in these signals, and peaks in the frequency spectrum were stochastic and chaotic in nature.

Furthermore, when comparing the frequency spectra of healthy and unhealthy conditions in Figure 6-2, no inconsistencies were observable and both sets of spectra had a similar shape. This reinforced the notion that strong frequency content was absent and time domain analysis remained a more suitable candidate for data separation, as depicted in section 0.

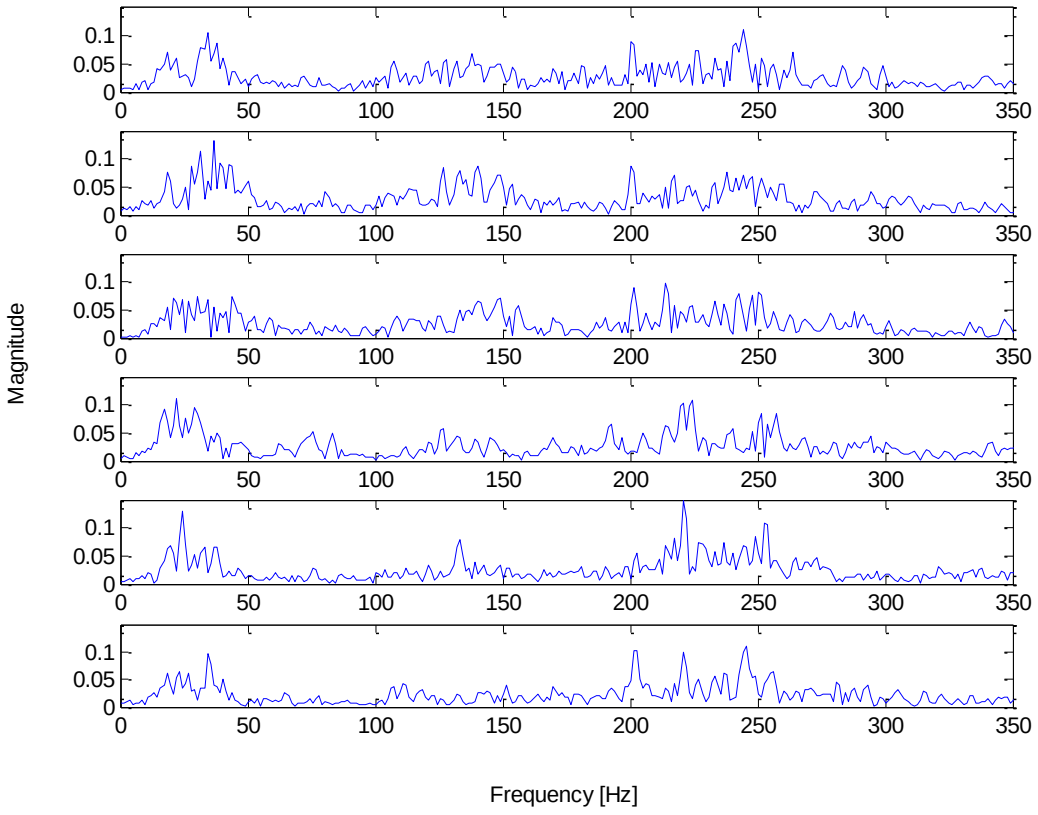


Figure 6-2: FFTs of three healthy signals (top three), and three loose fishplate signals (bottom three).

Based on the seemingly random behaviour of the transducer signals paired with the impulsive nature of the selected faults, envelope analysis presents itself as a suitable candidate for signal enhancement. The following section discusses the procedure and presents some results from an envelope analysis conducted in this study.

Envelop Analysis

In the case where the development of a fault creates an impulsive shock, such as a crack, the vibration modulation can be isolated using envelope analysis. The modulating signal is the result of the damage and can be analyzed in an envelope spectrum. Conventionally, envelope analysis is performed for fault detection in rotating machinery since the stationary nature of the signal amplifies dominant frequency components [37, 38]; however, the impulsive events that are the focus of fault detection in unsteadily operating machinery demonstrate similar behaviour when compared to unfaulted signals.

The envelope is obtained by first rectifying the signal, then applying a low-pass filter to remove high-frequency content, leaving the modulation signal. Conventionally, the enveloping would occur across a signal with periodic impulsive events; however, in the case of a rail-based system, the track loop or period is very large in comparison to typical rotating machinery. Therefore, the periodic nature usually observed in envelop analysis is absent, while the presence of the impulsive event corresponding with the fault remains. Consider the raw transducer signal in Figure 6-3; to enhance the signal to noise ratio, an envelope analysis is performed.

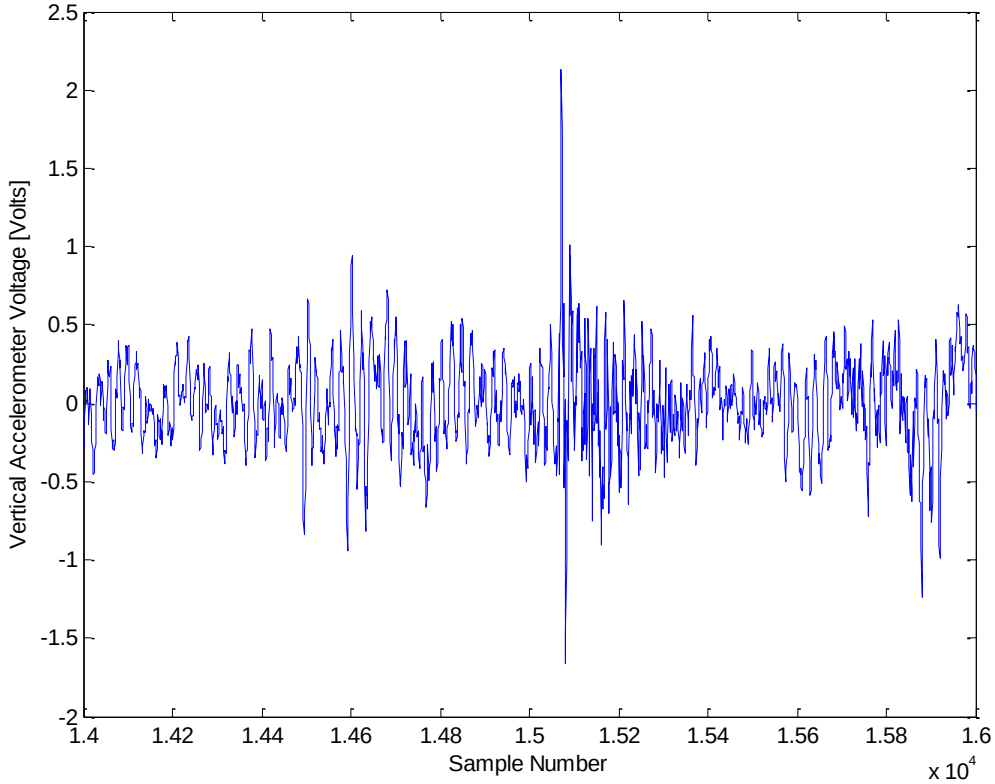


Figure 6-3: Vertical accelerometer response from loose fishplate fault. The fault impulse is located near sample 1.5×10^4.

Figure 6-4 shows the rectified version of Figure 6-3. The general shape of the upper envelope has remained the same.

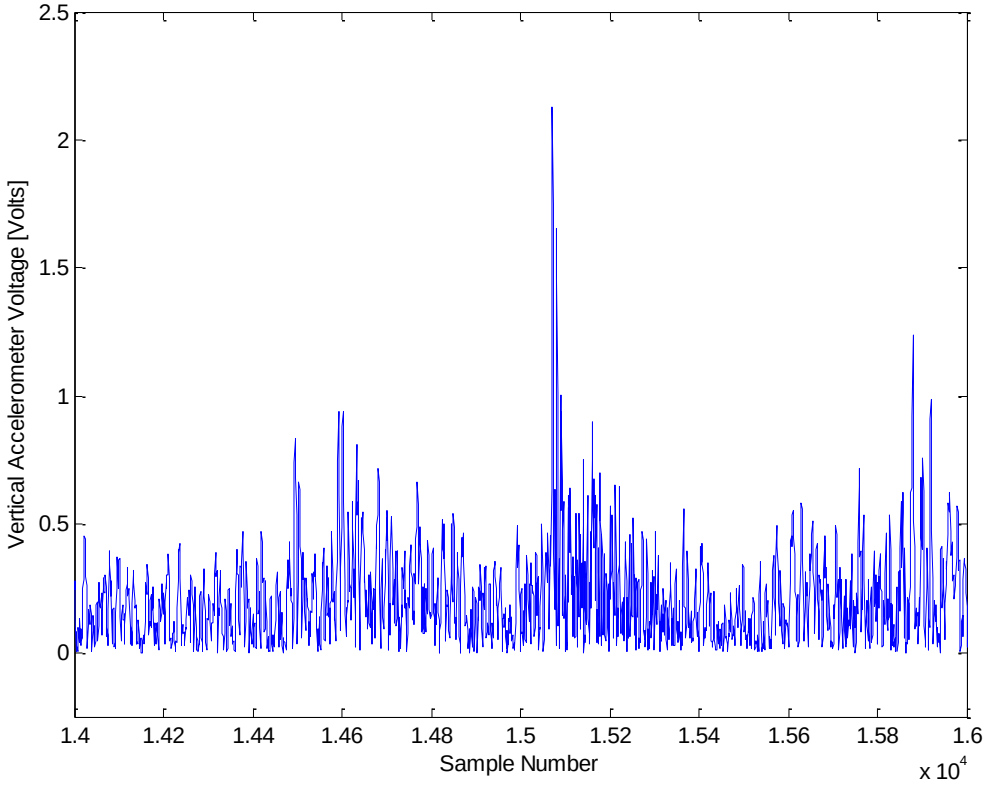


Figure 6-4: Rectified vertical accelerometer response from loose fishplate fault. The fault impulse is located near sample 1064.

Figure 6-5 shows the rectified signal after a simple moving average smoothing filter. Although the fault initiation at sample 1064 remains dominant in terms of amplitude, the signal to noise ratio has not been enhanced to the degree in which classification of the fault signal is trivial.

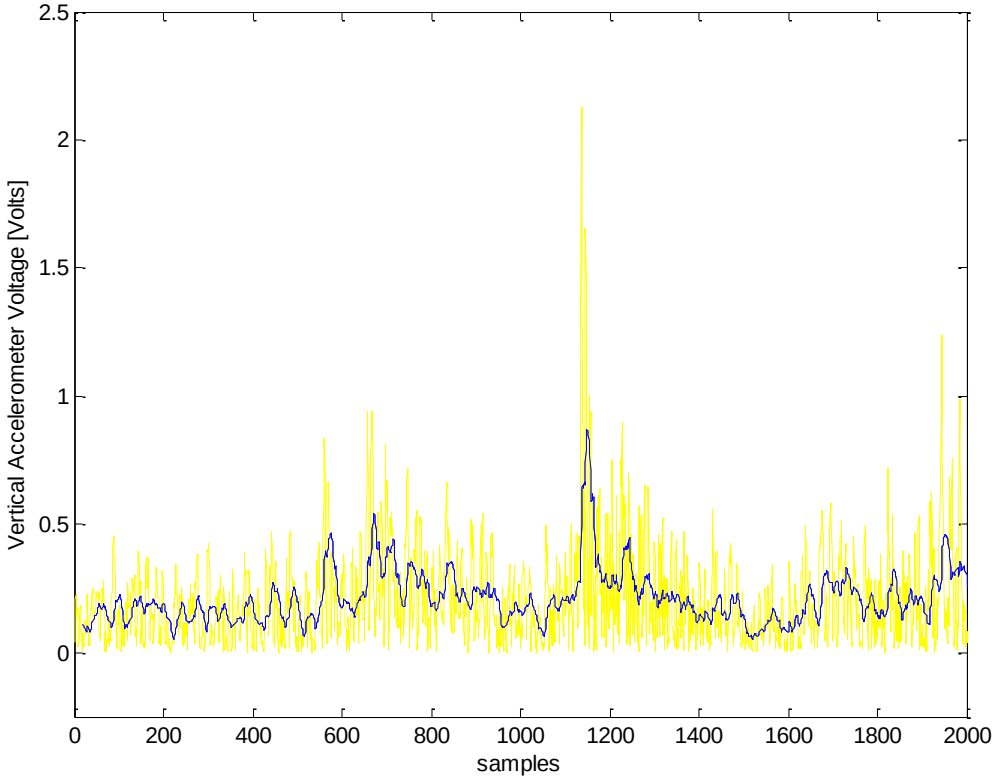


Figure 6-5: The rectified signal (loose fishplate condition) with a simple moving average applied for smoothing.

In the midst of preprocessing analysis, it became evident that signal enhancement would become a non-trivial task, and more attention was shifted to time domain feature selection. In addition, digital filtering techniques were also investigated; however, the methods were also cumbersome and did not lead to significant signal enhancement.