THREE ESSAYS ON RAILROAD SAFETY ANALYSIS USING NON-PARAMETRIC STATISTICAL METHODS

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Neeraj Dhingra

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department
Transportation, Logistics, and Finance

May 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

THREE ESSAYS ON RAILROAD SAFETY ANALYSIS USING NON-PARAMETRIC STATISTICAL METHODS

By

Neeraj Dhingra

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Raj Bridgelall
Chair

Dr. Pan Lu

Dr. Joseph Szmerekovsky

Dr. David Roberts

Approved:

06/21/2022
Date

Dr. Tim Peterson
Department Chair

ABSTRACT

The FRA mandated railroad companies to install a new monitoring system known as Positive Train Control (PTC). This system overlays sensors, signals, and transponders over existing track and other wayside infrastructure. Technologists designed the system to prevent accidents mainly caused by human negligence and communications. However, PTC will not address track-related defects, which is the second dominant cause of accidents.

A new track monitoring system called Railway Autonomous Inspection Localization System (RAILS) was proposed to address track-related accidents. RAILS is based on low-cost sensor technology that identifies defect symptoms, ranks their severity, classifies defect types, and localizes their positions. So, RAILS technology can augment the PTC by identifying track-related issues.

The main objectives of this dissertation are: (1) To compare the potential performance of RAILS with traditional inspection methods based on its fundamental theory of operation; (2) To identify factors contributing to railroad accidents; and (3) To determine and rank factors responsible for severe financial damages caused by railroad accidents.

The first two objectives will help compare the proposed technology and identify the major factors responsible for causing train accidents. The final objective will help to categorize accidents based on the potential financial damage severity. Categorizing such incidents would help to create a database that prioritizes issues and suggest possible countermeasure based on the problems.

The study's key findings are as follows: (1) RAILS is more efficient in conducting continuous inspection and identifying potential defects than traditional systems by 33%, with only two trains per day and a 50% first-pass detection probability; (2) Nonparametric methods

provide implicit information about rail accidents and function better than parametric methods by highlighting factors that are responsible for causing accidents rather than identifying the cause-and-effect relationship; (3) The most significant reasons for causing the financial damages are the number of derailed freight cars and the absence of territory signalization; and (4) Nonparametric methods automatically categorize rail accidents and, using text narratives, highlight causative factors responsible for a train derailment.

ACKNOWLEDGMENTS

Numerous people have contributed and helped me complete my study, and I would like to thank them by writing a few words.

First of all, I am deeply grateful to my advisor, Dr. Raj Bridgelall, for the guidance, support, encouragement, and advice throughout my graduate studies. And I am thankful for his flexibility and understanding at critical moments of my life. It would be hard to overstate how much I have learned and benefited from his support, knowledge, experience, and expertise.

I would like to thank Dr. Pan Lu for her continuous support and valuable assistance throughout this journey. She inspired me always and helped me work on different projects at the Upper Great Plains Transportation Institute (UGPTI). I would also like to express my appreciation to Dr. Joseph Szmerekovsky for his personal and professional guidance and Dr. David Roberts for his valuable support and being a member of my committee.

My gratitude is extended to Dr. Denver Tolliver for supporting my graduate studies at NDSU and Dr. Tim Peterson for being a true mentor. Dr. Peterson's support helped me excel in my career.

I am also grateful to Jody Bohn Baldock, Starla Morkassel, Megan Kortie, and all other staff members for their assistance and support at the UGPTI.

My friends and colleagues at the Upper Great Plains Transportation Institute and outside are acknowledged for their friendship and support.

My heartfelt gratitude goes out to my family for their encouragement and the many sacrifices they have made over the years. Finally, I am thankful, especially to my wife, Bhavana, for her patience and never giving up on me and our daughter, Eshana Dhingra.

DEDICATION

To my parents

Mr. M.M. Dhingra, and Mrs. Niru Dhingra

To my brothers and their families

Gaurav Dhingra, Ruchi Dhingra, and Anay Dhingra

Sourabh Dhingra, Aarti Malhotra, Zoe Dhingra, and Kai Dhingra

To my wife

Bhavana Dhingra

To our daughter

Eshana Dhingra

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION AND OBJECTIVES

1.1. Background and Motivation

The U.S. railway network spans seven Class I railroads (railroads with operating revenues of $433.2 million or more), 21 regional railroads, and 547 local railroads that operate 140,000 miles of tracks [1]. This vast connected rail network is a vital infrastructure and the most prominent mode of long-haul transportation because of its capacity, reliability, and efficiency. Today, more people and goods are moving through railroads than in the past. Consequently, the traffic increase exacerbates track geometry irregularities, which increases the risk of incidents and accidents over time. Hence, continuous track maintenance is necessary but increases maintenance costs.

FRA has distinctive practices, procedures, and guidelines for track inspections to achieve the timely detection and rectification of defects and maintain a safe rail network. However, current maintenance procedures include visual inspections and several automated techniques that require vast resources and increase the duration of traffic interruptions. It also limits the ability of FRA resources to enforce safety compliances with federal laws by monitoring the railway track geometry more frequently and across the entire network. Hence, more advanced and cost-effective inspection technologies are necessary for achieving the FRA's zero accident vision.

Railroad companies are currently working to implement a new mandated inspection system based on the intelligent condition monitoring known as Positive Train Control (PTC). PTC aims to reduce train accidents due to human error by improving communications between the system and train engineers. Hence, these technologies do not necessarily address accidents that occur from track related issues. In the last decade, derailments alone accounted for nearly 60% of the accidents, of which 42 % occurred because of track related issues [2].

There is a new advanced monitoring system proposed called Railway Autonomous Inspection Localization System (RAILS) to address accidents from track related issues [3]. The proposed RAILS system avoids expensive bogie sensor installation and adaptive sensor configurations. The sensors are similar to those available in smartphones, hence attachment direct to the train cars will enable low-cost and robust data collection. The sensors will compress and upload their geo-tagged inertial data periodically to a cloud-based system. Remote algorithms will combine and process the data from multiple train traversals to extract features. A statistical model built from the extracted features will estimate track geometry measurements such as profile, alignment, and warp. A central database will store and maintain all such information to help compare the current and the past track geometry conditions. The RAILS system will identify defect symptoms, rank their severity, classify defect types, and localize their positions in case of any discrepancy.

Some of the potential benefits of the RAILS system are defect classification, efficient screening, resource optimization, maintenance cost reduction, and enhanced railway safety. Other anticipated additional benefits are time savings, risk reduction, and a safer work environment for inspection personnel [4].

The proposed RAILS system aims to increase the defect identification rate and help to divert existing resources towards remediation. Subsequently, the railroads will be in a position to scale the RAILS system deployment over time to close the gap between defect formation and remediation rates to realize benefits in terms of accident risk reduction.

The main aim of this dissertation is to compare the potential performance of RAILS based on its fundamental theory of operation compared with traditional inspection methods and to evaluate how the system could use different methods to achieve its objective.

1.1.1. Overview

In this dissertation, I attempt to address the following three questions:

(1) How is the RAILS system more effective than traditional railway inspection methods based on their fundamental theory of operations?

(2) How to identify factors contributing to the significant railroad accidents?

(3) How to determine and prioritize factors responsible for severe financial damages caused by railroad accidents?

The introductory chapter provides an overview of the three essays

1.2. Research Objectives

1.2.1. Essay 1: Autonomous Railroad Track Monitoring System: An Onboard Instrumentation Technique

The USA's 140,000 miles rail network requires efficient track maintenance to minimize the risk of accidents. Federal Railroad Association (FRA) has distinctive practices, procedures, and guidelines for track inspections to timely detection and rectification of defects and to maintain a safe rail network. The significant challenges involved in following maintenance guidelines are limited finances, workforce & equipment, and a vast network to cover regularly.

Rail companies use nondestructive evaluation (NDE) technologies (including electromagnetic, acoustic, optical, and inertial sensing) to conduct track inspections for locating different abnormalities. The NDE methods are expensive, labor-intensive, slow, complicated, require vast resources, and increase the duration of traffic interruptions. Also, they have significant shortcomings in accuracy, precision, size, and costs limiting their deployment to specially constructed automated inspection vehicles that locate internal rail flaws and irregular

track geometry. Additionally, the amount of data these systems generate increases computational complexity and requires significant energy.

Moreover, FRA mandates railroads companies to install another technology known as Positive Train Control (PTC). Peters [5] explained early PTC designs intended to prevent train-to-train collisions, over-speed derailments, limit work zone accidents, and train movement through a switch left in the wrong position. Later, FRA requires companies to include Global Positioning System (GPS) under PTC, which helps distribute the train whereabouts to the entire network using digital signals. These instruments help locate trains with more precisions and improve coordination and communication.

Even with an improvised version of the technology, PTC is limited to improving communication between the system and training engineers to prevent accidents and minimize their severity caused by human-related factors or signal and communication problems. But these technologies will not address the defects related to the Track, Roadbed, and Structure (TRS), which are the second dominant cause of accidents. These growing numbers of TRS accidents necessitate more effective inspection and maintenance in less time by optimizing and automating these activities where possible [6].

A group of authors proposes a new sensor-based autonomous track geometry monitoring system called Railway Autonomous Inspection Localization System (RAILS). RAILS system uses inertial sensors on rolling stock to detect track irregularities automatically and continuously and characterize potential defects by analyzing the inertial dynamics of rolling stock. The sensors periodically compress and upload their geo-tagged data to a cloud-based system using the installed PTC communications network. Remote algorithms combine and process the data from multiple train traversals to extract features that identify defect symptoms, rank their severity,

classify defect types, and localize their positions. Also, RAILS proposes a methodology that can handle and analyze a massive amount of data with low computational complexity and high stability for practical use.

Consequently, this study aims to examine the accidents that occurred due to TRS and to extrapolate the potential of the proposed Railway Autonomous Inspection Localization System (RAILS) to reduce financial losses. The main contribution of this paper is a probabilistic method of comparing the performance of RAILS and its advantages over current NDE methods. RAILS is based on multiple passes or scans of railroad track segments per week, unlike traditional methods. The former uses specially equipped inspection cars with data post-processing and human evaluations, whereas the latter uses onboard sensors with real-time signal processing.

To achieve the objective, this study uses data from multiple scans of a track segment for the analysis and applied probabilistic models based on the theory of operations. Findings suggest that RAILS have the potential to outperform traditional NDE conditional monitoring systems by 33%, with only two trains per day and a 50% first-pass detection probability. The probability of detection advantage increases to 165% with 14 trains per week when the first-pass probability of detection drops to only 20%. A scenario analysis based on the proportion of track and roadbed problems that can generate a detectable inertial event suggests that the RAILS approach would have saved the industry $259 million in accident prevention over ten years.

1.2.2. Essay 2: Using Non-Parametric Methods to Identify Factors Contributing to Rail Derailment

The FRA requires companies involved in train accidents exceeding damages above the $10,700 threshold (2019, inflation-adjusted) to complete and submit detailed reports, which include numerous standardized fields that describe the accident conditions (temperature, time &

date, etc.), location (state, county, etc.), operations (speed, number of cars, etc.), and probable cause. Moreover, the reports also include a comprehensive text narrative for each accident, providing additional information about the accident's actual cause, responsible factors, and circumstances.

Previous research has statistically analyzed qualitative and quantitative data to highlight reasons for accidents, including derailments. These studies were limited by (i) data availability, (ii) details captured by the data, (iii) how the data are coded, and (iv) the assumptions of the statistical models. However, they overlooked detailed text narratives to determine causal factors. Analyzing these narratives using nonparametric machine learning models provides explicit information about these accidents as they are free from statistical assumptions, thus preventing limited and biased results [7] [8]. This study explicitly analyzed the text narratives using natural language processing techniques and machine learning models to determine train derailment causes based on each type of five accident categories.

My results indicate that the supervised algorithm and the random forest model performed best for conducting text-based classification and are used to identify local explanations highlighting factors associated with train derailments. Text explaining the local explanations does not necessarily have the same impact on the global corpora. Thus, I propose a new approach called GLIME, which converts the local explanations into the global explanations by aggregating the individual LIME explanation multiplied by the global TF-IDF values. I used these global explanations to conduct the association analysis on the top global words and the text narratives to find out the most frequent word pair used in a text narrative describing the accidents. My results indicate a strong association among a set of terms for each sub-category of the accidents. For instance, the correlation between the words 'journal', 'burn (0.46),' and overheat (0.27)' represents

that while defining the accidents related to axel and journal bearing, journal and bearing are generally used 46% of the time in the narration together and 27% times of with the word overheat. In addition, this methodology also provides solutions to counter errors, including typographical errors and data inconsistencies. Another significant contribution of this research is that it will help researchers and other personnel use this methodology to determine the causes of rail accidents and develop countermeasures at the micro and macro levels.

1.2.3. Essay 3: Ranking Risk Factors in Financial Losses from Railroad Incidents: A Machine Learning Approach

For a decade prior to 2019, nearly 25,000 accidents caused 446 deaths, 5,137 injuries, and more than $4.11 billion in financial loss seasonally adjusted to 2018 dollars [2]. Class I railroads accounted for 78% of those accidents, more than 72% of the resulting injuries and fatalities, and 81% of the total financial loss.

The consistently large number of accidents and the injuries and fatalities they cause place a substantial social and economic burden on the industry, environment, and society. Hence, it is vital to understand the dominant accident causes to guide strategies and policies that could minimize financial losses from accidents. Subsequently, this paper aims to apply data mining and machine learning techniques to 15 years of railroad accident data from 2004-2018 to reveal insights into the major contributing factors to financial losses from class I freight train accidents. The finding of this study provides the basis for developing effective maintenance strategies and efficient budget allocation. This research extends previous work in railroad safety in the following three ways:

- Considers 15 years of accident data reports, which is longer than the periods that other studies covered.

- Isolates factors that lead to financial losses from railroad accidents.
- Ranks the importance of the major factors in financial losses from railroad accidents.

Data between 2004 and 2018 from the railroad equipment accident (REA) database provided inputs to achieve the objective. The analysis compared results of six machine learning models, including Random Forest (RF), Gradient Boosting Model (GBM), eXtreme Gradient Boosting Model (XGBM), clustering with k-nearest neighbors (KNN), Support Vector Machine (SVM), Simple Linear Regression (SLR) using K-cross validation (K=10) with 3 repeats to improve the generalization. The final evaluation metric is a root-mean-squared error (RMSE) and means absolute error (MAE), which are the standard criterion for model selection.

Results showed that tree-based tree-based ensemble models performed best. Particularly, XGBM proved to be the best model for analyzing railroad accident data that is highly imbalanced. The XGBM model identified the significant contributors to railroad accidents. The results indicate that LOADF2 (number of derailed loaded freight cars), SIGNAL (Type of territory – signalization), and EMPTYF2 (number of derailed empty freight cars) are the top three significant factors that account for financial loss severity with the gains of 57.46%, 20.22%, and 10.12% respectively.

1.3. References

[1] ASCE, "2017 infrastructure report card," American Society of Civil Engineers, 2017.

[2] FRA, "Accident Data as reported by Railroads," Federal Railroad Administration, [Online]. Available: https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/on_the_fly_download.aspx. [Accessed 17 March 2018].

[3] L. Chia, B. Bhardwaj, P. Lu and R. Bridgelall, "Railroad track condition monitoring using inertial sensors and digital signal processing: A review," *IEEE Sensors Journal*, vol. 19, no. 1, pp. 25-33, 2018.

[4] R. Bridgelall, P. Lu, D. Tolliver, N. Dhingra and B. Bhardwaj, "Benefit Cost Analysis of Railroad Track Monitoring Using Sensors Onboard Revenue Service Trains, MPC-21-446," North Dakota State University - Upper Great Plains Transportation Institute, Fargo: Mountain-Plains Consortium, 2021.

[5] J. C. Peters and J. Frittelli, "Positive Train Control (PTC): overview and policy issues," Congressional Research Service Report, 2018.

[6] R. N. Ngigi, C. Pislaru, A. Ball and F. Gu, "Modern techniques for condition monitoring of railway," *Journal of physics: conference series.*, vol. 364, no. 1. IOP Publishing, p. 012016, 2012.

[7] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations," *Journal of Modern Transportation*, vol. 24, no. 1, pp. 62-72, 2016.

[8] C. Arteaga, A. Paz and J. Park, "Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach.," *Safety Science*, vol. 132, p. 104988, 2020.

2. AUTONOMOUS RAILROAD TRACK MONITORING SYSTEM: AN ONBOARD INSTRUMENTATION TECHNIQUE

2.1. Abstract

Human-Related Factors (HF) and Track, Roadbed, and Structure (TRS) related issues are the two primary causes of railroad accidents. Large railroad companies are installing a new wireless sensor-based system called Positive Train Control (PTC) to reduce accidents. However, PTC is limited to preventing accidents related to human factors or signal and communication. This paper reports on a new low-cost onboard system to identify consistent inertial events from locations that pose potential risks of TRS related accidents. The Rail Autonomous Inspection Localization System (RAILS) uses inertial sensors on rolling stock to continuously assess track irregularities and to characterize and categorize potential defects by analyzing the inertial dynamics of rolling stock. The system can communicate sensory data using the installed PTC communications infrastructure. Railroads are required to inspect all tracks in operation as often as twice per week; whereas the RAILS accomplish several scans based on the number of trains passing over a track segment of track each day. The primary contribution of this study is a probabilistic analysis that quantifies the effectiveness of RAILS over the current railway inspection methods. A scenario analysis found that when the first-pass fault detection probability of the sensor is 50%, the RAILS achieve 33% better chance of detecting a fault with only two train traversals per day. With a much lower first-pass fault detection probability of 20%, the RAILS advantage was 165% at 14 scans per week.

2.2. Introduction

The U.S. rail network is the most prominent mode of long-haul transportation, which is vital for the U.S. transportation system and economy. Currently, U.S. railroads deliver five

million tons of freight and carries approximately 85,000 passengers each day, which is expected to increase by 40% by 2040 [1]. Consequently, the traffic increase exacerbates track geometry irregularities, increasing the risk of incidents and accidents over time. Timely detection and rectifications of track irregularities and track defects are vital for maintaining a safe rail network. To comply with safety standards, railways spend close to 40% of their total revenue on inspections, capital expenditure, maintenance, and condition monitoring each year [2]. Despite those huge investments, there were still nearly 19,000 accidents that cost railroads $3.10 billion over the last decade. Human-Related Factors (HRF), and Track, Roadbed, And Structure (TRS) related issues are the two primary causes of accidents. HRF and TRS accidents accounted for more than 37% and 25% of the total number of accidents from 2009 to 2018, respectively. HRF and TRS accidents were responsible for approximately 33% and 30% of the total financial losses, respectively [3]. Figure 2.1 summarizes total number of pre-pandemic accidents from 2009 to 2018, due to different causes.

In recent years, many authors raised concerns about such a large number of accidents in spite of considerable investments in railway infrastructure [4], [5], [6]. Those authors highlighted that the traditional and existing Nondestructive Evaluation (NDE) methods (including electromagnetic, acoustic, optical, and inertial sensing) are expensive, labor-intensive, slow, complicated, requires vast resources, and increase the duration of traffic interruptions. The NDE methods also limit the ability of FRA personnel (and other resources) to enforce safety compliances with federal laws by monitoring the railway track geometry more frequently and across the entire network. Hence, more advanced and cost-effective inspection technologies are necessary for achieving the FRA's zero accident vision. These new technologies should comply

with the industry's current practices, facilitate inspections, help to better allocate budgets, provide efficient use of resources, and allow for uninterrupted railway services.

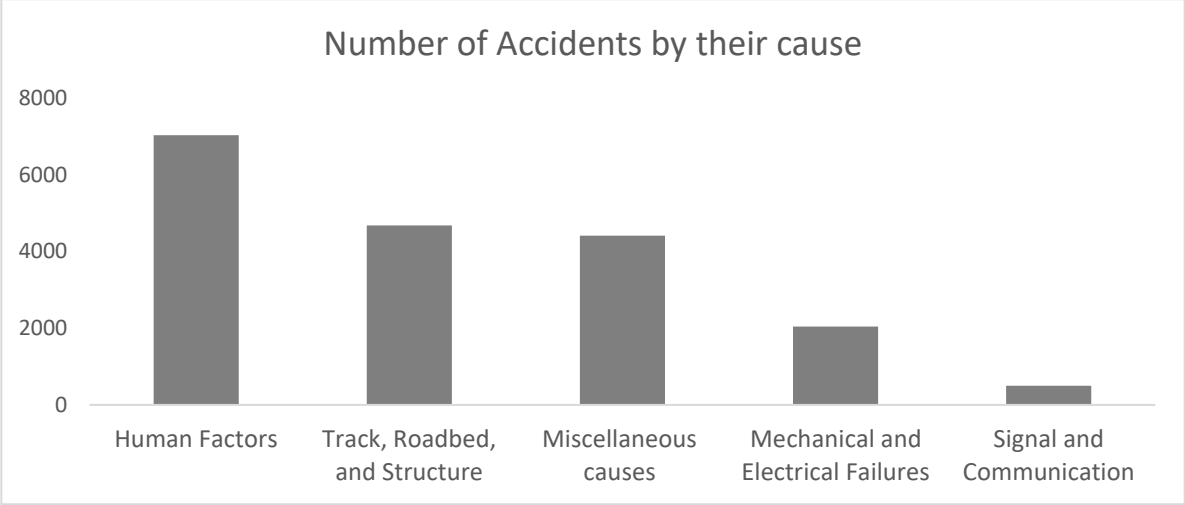

Figure 2.1. Factors responsible for causing rail accidents from 2009 to 2018

One technology that many of the large railroad companies implemented is Positive Train Control (PTC). As explained by Peters [7], early PTC designs intended to prevent train-to-train collisions, over-speed derailments, limit work zone accidents and the movement of a train through a switch left in the wrong position. Conversely, the system had some functional boundaries which limited the benefits. Such limitations include 1) the operating system does not track the real-time train location, 2) can only communicate when a train passes the wayside infrastructure, 3) limited data rates because the use of Wi-Fi is not practical in the PTC. Consequently, a more sophisticated and expensive variant of the early PTC system, Communications-Based Train Control (CBTC), emerged. CBTC is a more advanced system in which train information is sent to a central location which then distributes this information to the entire network. In this architecture, the installation of a Global Positioning System (GPS) helps to track the train location and speed, along with the other instrumentation. These instruments help to locate trains even in GPS denied environments. Similar to cell phone technology, the

advanced infrastructure communicates continuously using the digital signals which provide greater precision. The FRA later mandated railroads to install a CBTC system and placed all such safety technologies under the PTC umbrella.

Even with an improvised version of the technology, PTC is limited to improving communication between the system and train engineers to prevent accidents and minimize their severity when caused by HRF or signal and communication problems. But these technologies will not address the TRS related issues, which are the second dominant cause of accidents. Figure 2.2 summarizes the annual railroad accidents due to TRS and the resulting financial losses for the decade prior to 2019. Consequently, the **goal** of this study is to examine the accidents that occurred due to TRS, and to extrapolate the potential of the proposed Railway Autonomous Inspection Localization System (RAILS) to reduce financial losses due to those accidents.

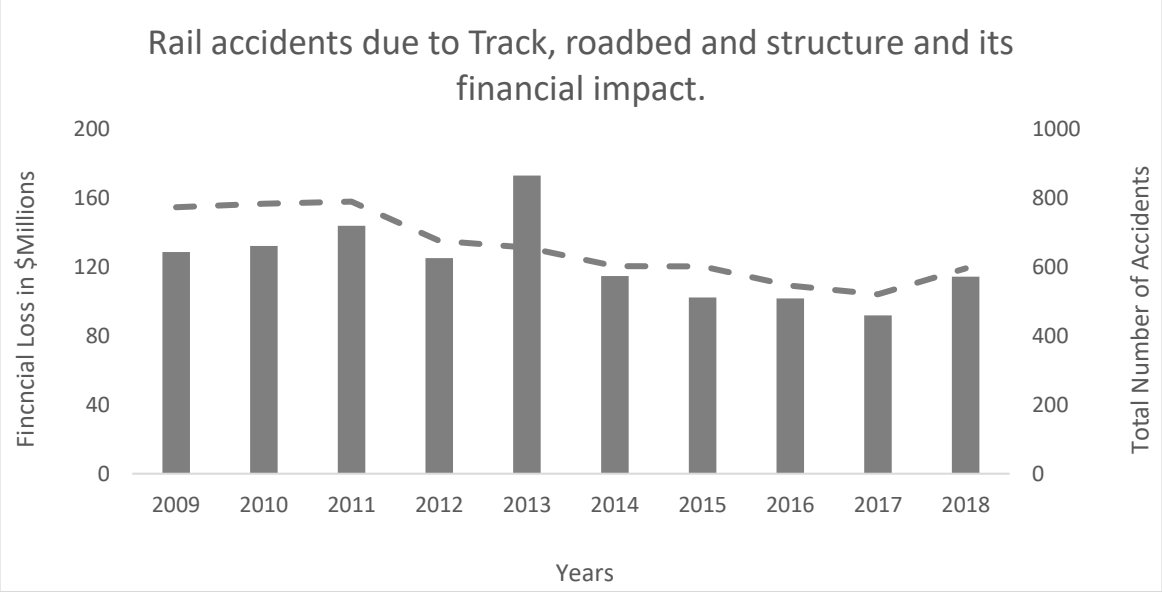

Figure 2.2. Railroad incidents due to TRS from 2009 to 2018 and the reported financial loss

A large number of TRS related issues necessitates more effective inspection and maintenance in less time by optimizing and automating these activities where possible [8]. In the recent past, many authors have proposed intelligent condition monitoring systems (ICMS). Such

methods include ultrasonic signaling [9] [10], vision-based inspections [11] [12], Ground-Penetrating Radar (GPR) [5], and wireless sensor-based systems that include fuzzy logic and surveys [4], [13], [14]. But these techniques require huge installation capital and maintenance costs. Also, most of these methods cannot classify and prioritize defects based on need.

The RAILS proposed in this research can utilize the installed PTC communications networks [15], [16]. RAILS will use inertial sensors on rolling stock to assess track irregularity automatically and continuously. Further, the system will characterize and categorize potential defects by analyzing the inertial dynamics of rolling stock. Defect classification will enable asset managers to allocate the appropriate specialists to scrutinize the location of the defect. This would help railroads focus inspections on high-risk areas without closing lines to search for developing issues. Efficient screening, optimized usage of resources, reduced maintenance costs, and improved railway safety are some of the key potential benefits of the RAILS [16].

The core focus of RAILS is to improve fault screening by identifying defects at the primitive stage and increasing the fault identification accuracy over time. Subsequently, the main **contribution** of this paper is a probabilistic method of comparing the potential performance of RAILS based on its fundamental theory of operation as compared with traditional NDE inspection methods. RAILS is based on multiple passes or scans of railroad track segments per week whereas traditional methods of inspections are based on substantially fewer scans per week. The former uses on-board sensors with real-time signal processing whereas the latter uses specially equipped inspection cars with data post processing and human evaluations. Other works by the authors describe in greater detail how the sensors combined with the algorithms detect and report on specific track defects [17], [18], [19], [20]. Rather, the focus of this work is to characterize the probabilistic performance difference between RAILS and current NDE

methods of track inspection. The remainder of the paper is structured as follows: the literature review discusses some of the previous related work. The methodology section presents the architecture and operational characteristics of RAILS. The experiment section quantifies the benefits of using RAILS over existing NDE methods. The conclusion section presents the final remarks and describes future work.

2.3. Literature Review

A condition monitoring tracks divergence from normal operating conditions to predict failures [21]. Over the years, condition monitoring techniques have evolved from measurement-oriented to computer-based strategies. Researchers have started placing more emphasis on an early warning system based on Wireless Sensor Networks (WSNs) for condition monitoring. WSN is a cooperatively wireless network of spatially distributed and autonomous devices that monitor infrastructure, structures, and machinery. WSNs facilitate low-cost monitoring of extensive infrastructure at a faster pace, require less infrastructure and maintenance, provides autonomous and near real-time data acquisition, and improve data management and accessibility [8]. Consequently, practitioners have begun to consider WSN as potential substitutes for the traditional tethered railway track monitoring system [22]. Important considerations for WSN installations are (i) sensor location and (ii) data analysis methods. The following subsections describe literature that considered those two decisions.

2.3.1. Sensor Types, Quantity, and Location

Many previous studies have conducted their research by using different sensors ranging from uniaxial accelerometer/gyroscopes to Inertial Measurement Units (IMUs) with 6 degrees of freedom, also by applying different numbers of sensors, including 3 [23], [24], 4 [25], 8 [26], 9 [27], and 15 [28] either on the track or on the vehicles.

Track-based monitoring systems have sensors installed at fixed locations such as bridges, tunnels, rail tracks, rail track beds, and other infrastructure locations. Some authors highlighted the importance of using fixed sensors to identify at specific areas of a rail infrastructure. For instance, Jang et al. [29], Moreu et al. [14], and Kołakowski et al. [30] used sensors to detect cracks and other internal structural damage to bridges. Zan et al. [31]; and Jenkins et al. [32] used WSNs to examine the health of tunnels. Wei et al. [33], Kouroussis et al. [34], and Filograno et al. [35] examined wheel condition with sensors embedded in rail tracks. Hodge et al. [4] covered numerous other related studies. However, there are a few functional limitations of track-based sensor systems. One is that the system functions only when a vehicle passes over the embedded sensors, which limits the detections to fixed locations. Hence sensors mounted on in-service vehicles are a solution to this functional problem.

Modern electronics, along with the development of robust sensors, facilitate constant condition monitoring through compact WSNs that can be installed on in-service vehicles [36]. Vehicle-based sensors collect data continuously. Sensors, typically include accelerometers, gyros, noise sensors (e.g., microphones) and GPS. Such sensors can identify track irregularities, dynamic vehicle behavior, vehicle location, and speed [37]. WSNs facilitate timely maintenance intervention and allows early detection of faults, which helps in effective planning for future track maintenance [26]. Many others authors have recommended different locations such as the train shell, wagon, bogies, axles, wheels, brakes, and pantographs. Installations included sensors for measuring track irregularities [36], [26], temperature sensors for evaluating anomalies [38], [39], and fiber Bragg grating sensors for analyzing bogie vibration [34], [40]. Some authors have applied sensors to monitor the stress on axels [41], [42] and on wheels [43], [44]. Ngigi *et al.* [8], Hodge *et al.* [4], and Chia *et al.* [45] highlighted many other studies that used WSNs.

However, none of the reviewed articles highlighted variations or impacts in their findings based on the type of sensors and the number of sensors used. Interestingly, some studies have found that the following factors are essential when selecting the sensor's location.

The linkage between sensor measurement and the train speed. Because the characteristics and the angular velocity would always be different from location to location, even traveling at the same speed. So in such cases, the sensor location is essential. The wheel's vibration may add more noise to the sensor data, resulting in false-positive or missed detection [45]. So the sensor's location could help avoid unnecessary corruption of the data.

Most previous studies highlighted that using the appropriate signal processing technique could help deal with the above two factors. As suitable signal filtering algorithms help reduce or eliminate unwanted signals features in some frequency ranges, that could help in further analysis. RAILS systems developed, applied, verified, and validated all algorithms required to process the sensor data and provide suitable results [17], [46], [47].

2.3.2. Analytical Methods

A reliable condition monitoring system requires maintaining proper coordination between data generation, data processing, and producing meaningful output. The sensors generate enormous amounts of data which often contains noise and other unwanted signal elements. Hence, it is crucial to select an appropriate technique for the successful implementation of a safe and functional condition monitoring system. In the previous related studies, authors have also provided some insight into using different techniques for various aspects of railway inspection and partitioned them into (i) model-based, and (ii) signal-based techniques.

Model-based techniques rely on mathematical models that characterizes a relationship between the input signals and signals from the vehicle response. In this method, sensor data are

used as an input to predict vehicle system dynamic behavior and then compared with the real-time measurements. The difference between predicted and measured data is used to find faults and anomalies. The accuracy of these models depends upon "*the selection of the initial values during partial linearization (or other approximation methods); unknown noises; model uncertainty caused by the nonlinear suspension parameters and structure flexibility*" [48]. Commonly used methods include Kalman filters [44], [49], [50], extended Kalman filter [51], [52], inverse modelling [53], [54], unscented Kalman filter [55], [56], Rao–Blackwellised particle filter [57], [58], and sequential Monte Carlo method [59], [8].

Signal-based methodologies apply where only output signals from vehicle response are available for analysis in response to some disturbance. The functioning of the system depends on the pre-built fault database through which signals compare the fault features and classify most similar fault conditions for identifying the fault type and level. Li *et al.* [48] suggested that many authors analyze the signal-based techniques in many ways including time-domain [26], [60], frequency- domain [60], [61], time-frequency approach [62], [63], correlation analysis [64], and simple peak magnitudes thresholding.

2.4. Proposed Methodology

In many cases, it would be impractical to install sensors on each vehicle. For example, the cost of monitoring the condition of a bogie could be more than the expense of repairing a fault [8], [65]. The proposed RAILS methodology does not rely on expensive bogie sensor installations and adaptive sensor configurations. The sensors, which are similar to those available in smartphones, may be installed directly in the train cars for low-cost and robust data collection. Also, the technology will use the installed PTC communications infrastructure to transmit the inertial and geospatial position data to a cloud-based system for subsequent analysis. Hence,

RAILS investments will develop PTC compatible equipment for installation, data processing, data handling, and maintenance [66], [67].

Major track surface abnormalities produce accelerated car movements in all directions [68]. Inertial sensors that monitor Vehicle-Track Interactions (VTI) can detect such responses from a railcar to help identify potential defects at a primitive stage. VTI sensors are widely used due to their small size, low cost, low power consumption, and robustness but they still have limitations [48], [15]. VTI sensors do not support two-way communication, which is crucial for classifying faults and identifying their location. In existing systems, inertial sensors detect symptoms of possible track and equipment defects through the inertial response of a railcar. These inertial sensors will only raise an alarm when acceleration magnitudes exceed a fixed threshold. However, there is no standard procedure to pre-configure the threshold values, and the system requires a technician to define such values based on their experience or intuition. Threshold adaptation is a complex process because the inertial responses of vehicles will vary with train speed, gross weight, suspension system design, and weather conditions. Hence, fixing thresholds must account for these exogenous circumstances to improve accuracy and reduce false positives and false negatives.

2.4.1. System Architecture

RAILS sensors on train cars or hi-rail vehicles compress the inertial data, geo-tag their inertial samples, and upload the data every second to a centralized processor. Figure 2.3 represents the operation of the RAILS currently under test at a local railroad. It shows steps to detect and classify possible track defects. Railway engineers, practitioners, or railroad companies can use the data to understand an enhanced situation to optimize inspection and maintenance

practice. Therefore, it will minimize the cost and safety risks while maintaining reliable track and equipment condition information.

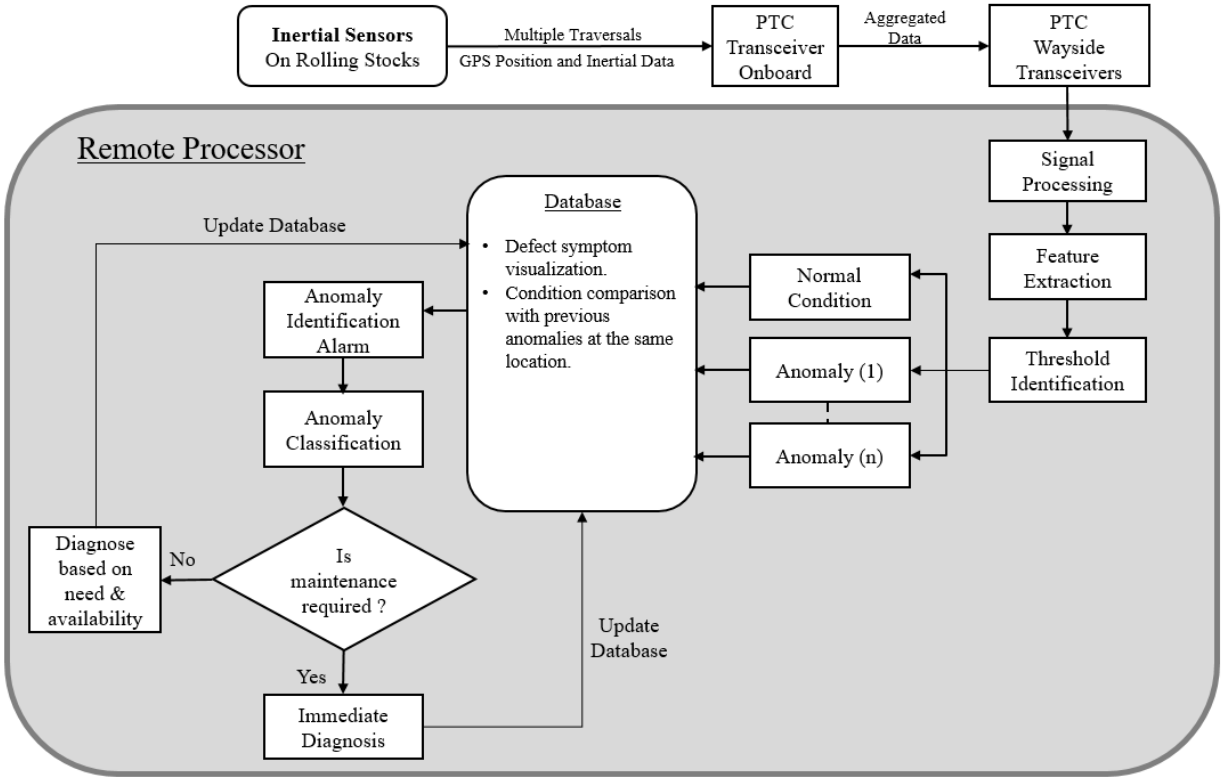

Figure 2.3. RAILS operating system architecture

Remote algorithms combine and compress three-dimensional linear acceleration, angular acceleration, and geospatial data from multiple train traversals to extract features that identify defect symptoms, rank their severity, classify defect types, and localize their position. This extracted feature helps to identify the fixed threshold values, which can be visualized based on a pre-defined color-coding scheme. Subsequently, a wide range of data mining techniques, such as machine learning, genetic algorithms, feature correlation, Bayesian analysis, and maximum likelihood, will perform the signal classification to determine the probable cause of the fault. The authors will extend the proposed algorithms to detect and classify a wide variety of possible track defects, such as broken rail, irregular geometry, fastening system defects (e.g. missing

spikes, displaced anchors), support structure defects (deteriorated ballast, mud spots, weak sub-grade), and vehicle defects (e.g. sticking brakes, wheel wear, suspension system misalignment) [15].

2.4.2. Theory of Operations

RAILS sensors onboard in-service vehicles will detect a variety of kinetic energy responses, speed, and position coordinates from a GPS receiver. Combining data across multiple train traversals will significantly enhance the Signal-To-Noise Ratio (SNR) by ensemble averaging. A higher SNR ratio will reduce false positive and false negative detections. The underlying theory is as follows. Let P_s be the probability of detecting a fault during the first pass scan. Therefore, the probability of not detecting the fault on the same attempt would be $P_{nd} = 1 - P_s$. Given that each future scan does not depend on the results of previous scans, the results are independent. Therefore, the probability of not detecting a fault after N scans would be the product of the individual probabilities such that

$$P_{ndn} = (1 - P_s)^N \tag{1}$$

and the probability of detecting the same fault after N attempts must be

$$P_D = 1 - (1 - P_s)^N \tag{2}$$

The equation evaluates to $P_D = 1$ as a function of N. Note that the extreme case of $P_S = 1$ represents the presence of a fault that produces a sufficiently high SNR for detection in a single attempt. Conversely, when a fault is not present ($P_S = 0$), the equation evaluates to zero as expected.

Federal track safety regulations require railroads to inspect all tracks in operation as often as twice weekly; whereas the number of scans with RAILS will be based on the number of trains passing over a segment of track each day. For example, if two trains per day traverse a segment of the track that contains a fault, then the system would have conducted 14 scans of the segment

per week without requiring track closure. Given the same probability of detection with a single attempt for both methods, RAILS will produce a higher probability of detection each week because of the larger number of scans conducted.

Figure 2.4 shows the impact of inspections with different probabilities of P_S ranging from 0 to 1. The probability of detecting faults with several values of *N* (trains per day) is compared with the results of two inspections per week using NDE methods (NDE2). RAILS14 represents two trains per day traversing a segment, thus resulting in 14 scans weekly. Similarly, RAILS21, RAILS28, RAILS35, represent scenarios of 21, 28, and 35 weekly scans, respectively. The results demonstrate that both methods provide identical results when the value of Ps is close to one. However, RAILS produce significantly better results when the chance of fault detection with a single scan is below 80%. For example, for a scenario where the success of fault detection is random (50%), RAILS produces a 33% better chance of detecting the fault with traversals of two trains per day across the segment. If the first-pass success of fault detection further reduced to 20%, RAILS has a 165% better chance of detecting a fault with 14 scans per week.

The percentage of the TRS related accidents that RAILS can address is

$$A_{RAIL} = A_{TRS} \times \eta_{TRS} \times P_D \tag{3}$$

where A_{RAIL} is the proportion of TRS related accidents that produce inertial events and η_{TR} is the proportion of those inertial events that RAILS can detect. In the decade between 2009-2018, TRS-related issues caused more than 6,500 accidents. Appendix B contains the list of factors responsible for causing these accidents. All these issues have the potential to generate inertial signals in the early stages, which could peak when the fault gets worst. However, with the usage of RAILS, combining multiple inertial events from several traversals helps detect the faults before they potentially cause an accident.

Figure 2.5 illustrates the three types of movements that mostly generates high enough PIEs that the RAILS system can detect to predict the type and severity of track and roadbed problems. Profile irregularities are vertical deviations from a flat surface. Alignment irregularities are lateral deviations from a straight line. Warp irregularities are uneven vertical displacements between the two rails that can cause rocking motion and lead to derailments. Table 2.1 contains the list of five possible issues that would generate at least one of those movements that RAILS could detect during the first scan. The statistics from Table 2.1 and Table B1 for the study period of a decade suggest that irregularities that generate PIEs causes almost 62% (4099/6555) TRS related accidents. For a scenario, η_{TR} = 60%, this yields P_D = 0.84 at two trains per day. Hence, A_{RAIL} = 0.62 x 0.60 * 0.84 = 0.3125. This result indicates that RAILS could have detected more than 31% of the TRS issues that produced detectable inertial events. Figure 2.6 plots A_{RAIL} as a function of the full range of η_{TR} scenarios and with N = 3, 4, and 5 scans.

Table 2.1. TRS potential issues generating PIEs

S.No	Factor Description	Count
1	Broken Rail (because of any reason)	1,794
2	Wide gage (because of any reason)	1,699
3	Defective or missing crossties	106
4	Cross level of track irregular	279
5	Track alignment irregular (buckled/sunkink)	221
	Total	4,099

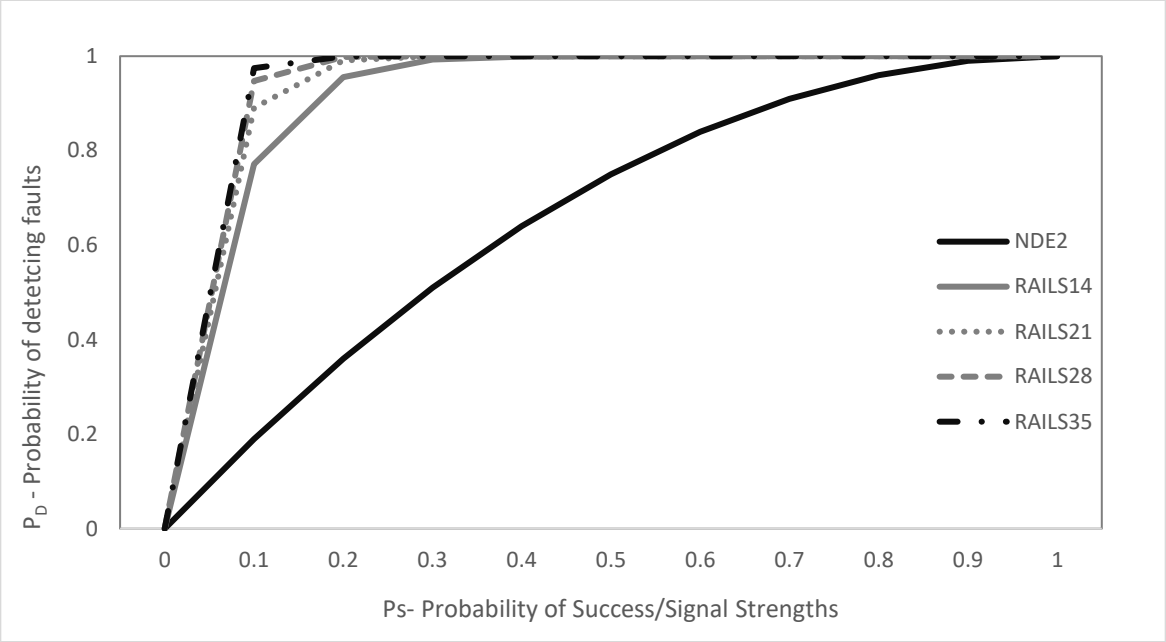

Figure 2.4. Fault detection rates based on different frequency of weekly inspection

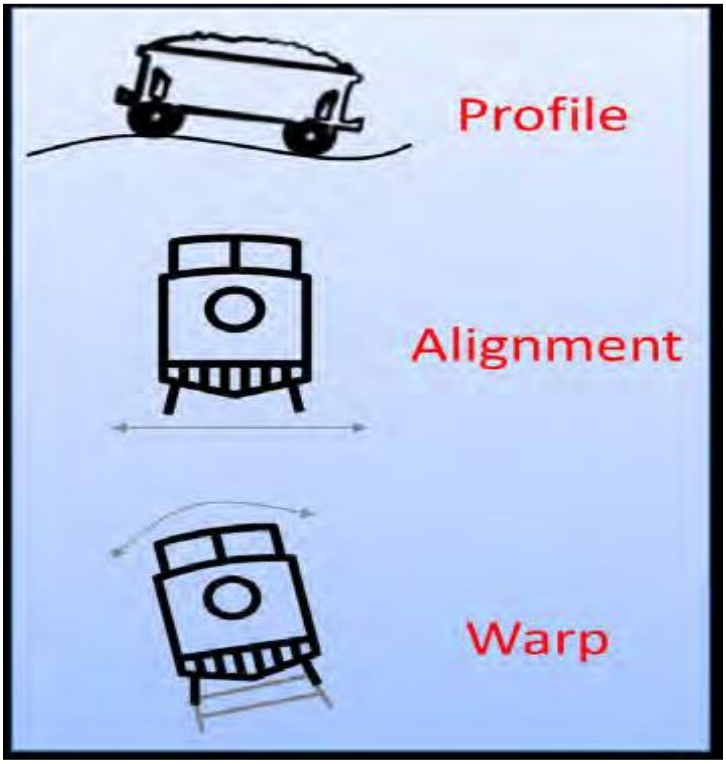

Figure 2.5. Movement detections due to track and roadbed problems

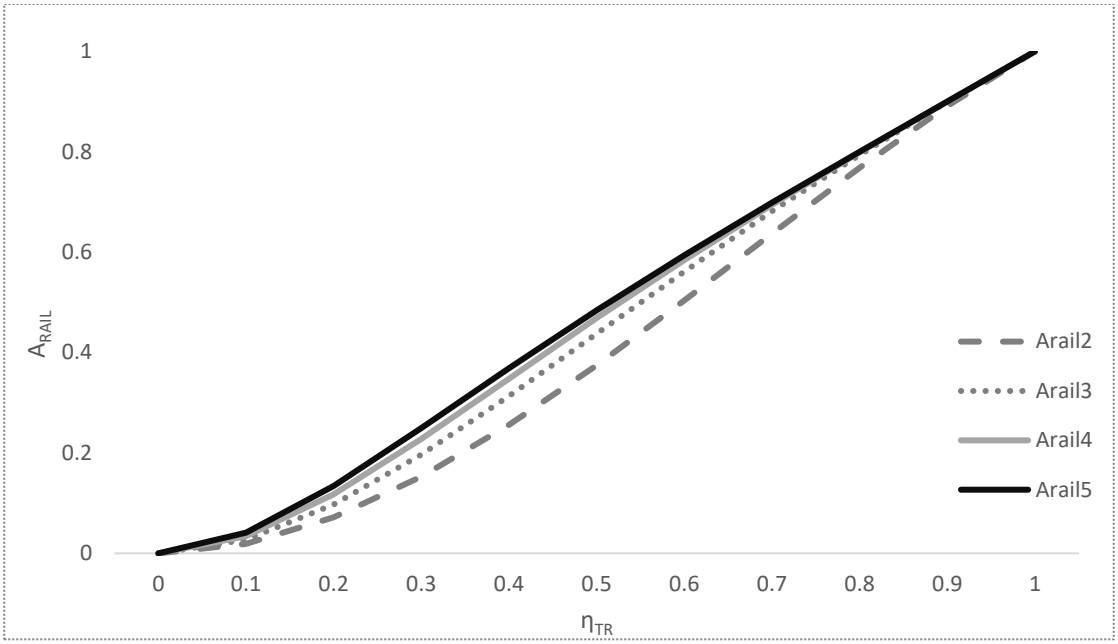

Figure 2.6. Accidents addressed by RAILS for a range of N values and η_{TR}

In the last decade, the railroad industry has lost more than $830 million due to TRS related accidents. Based on the above scenario of η_{TR} = 60%, RAILS would have resulted in a financial savings of $830 million × 0.312 = $259 million.

2.5. Conclusion

The railway industry would like to do track maintenance more efficiently by maximizing the use of limited resources to minimize the risk of accidents. The significant challenges involved in efficient maintenance are limited finances, workforce & equipment, and a vast network to cover at a regular interval of time. Along with these difficulties, the industry also requires continuous maintenance of rail geometry as tracks deteriorate due to weather, traffic density, and heavy load movements. Different track types and classes have distinct deterioration rates, which makes their maintenance more challenging. Unattended defects increase the risk of an accident over time. To minimize accident risk, the rate of detection and remedial measures

should be at least equal to the rate of defect formation. The rate of detection can improve with a quality inspection system that provides low false positive and low false negative detection rates.

This study describes a proposed RAILS system that use inertial sensors on rolling stock to detect track irregularities automatically and continuously, and to characterize potential defects by analyzing the inertial dynamics of rolling stock. The sensors used are similar to those available in smartphones, thus enabling low-cost and robust data collection. The sensors compress and upload their geo-tagged data periodically to a cloud-based system by using the installed PTC communications network. Remote algorithms combine and process the data from multiple train traversals to extract features that identify defect symptoms, rank their severity, classify defect types, and localize their positions. Defect classification will enable asset managers to allocate the appropriate specialists to scrutinize the high-risk locations flagged. Therefore, RAILS will transform existing condition monitoring strategies from "find and fix" to "predict and prevent" by focused follow-up inspections to fewer locations. Hence, the anticipated additional benefits of using RAILS are time savings, risk reduction, and a safer work environment for inspection personnel.

A probabilistic analysis based on the theory of operations from multiple scans of a track segment suggests that RAILS have the potential to outperform traditional NDE conditional monitoring systems by 33% with only two trains per day and a 50% first-pass detection probability. The probability of detection advantage increases to 165% with 14 trains per week when the first-pass probability of detection drops to only 20%. A scenario analysis based on the proportion of track and roadbed problems that can generate a detectable inertial event suggest that the RAILS approach would have saved the industry $259 million in accident prevention over ten years.

Future work will explore and evaluate an aggregated benefit-cost model for the deployment of RAILS on rolling stock. The results would provide some ground for assessing the financial viability of RAILS.

2.6. Reference

[1] ASCE, "2017 infrastructure report card," American Society of Civil Engineers, 2017.

[2] AAR, "Railroad 101- Freight Railroads Fact Sheet," 2020. [Online]. Available: https://www.aar.org/wp-content/uploads/2020/08/AAR-Railroad-101-Freight-Railroads-Fact-Sheet.pdf. [Accessed 23 May 2021].

[3] FRA, "Accident Data as reported by Railroads," Federal Railroad Administration, [Online]. Available: https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/on_the_fly_download.aspx. [Accessed 17 March 2018].

[4] V. J. Hodge, S. O'Keefe, M. Weeks and A. Moulds, "Wireless Sensor Networks for Condition Monitoring in the Railway Industry: A Survey," *IEEE Transactions on intelligent transportation systems,* vol. 16, no. 3, pp. 1088-1106, 2014.

[5] S. Fontul, E. Fortunato, F. De Chiara, R. Burrinha and M. Baldeiras, "Railways Track Characterization Using Ground Penetrating Radar," *Procedia engineering,* vol. 143, pp. 1193-1200, 2016.

[6] P. Lu and D. Tolliver, "Accident prediction model for public highway-rail grade crossings," *Accident Analysis & Prevention,* vol. 90, pp. 73-81, 2016.

[7] J. C. Peters and J. Frittelli, "Positive Train Control (PTC): overview and policy issues," Congressional Research Service Report, 2018.

[8] R. N. Ngigi, C. Pislaru, A. Ball and F. Gu, "Modern techniques for condition monitoring of railway," *Journal of physics: conference series.,* vol. 364, no. 1. IOP Publishing, p. 012016, 2012.

[9] M. Carboni and S. Cantini, "A model assisted probability of detection approach for ultrasonic inspection of railway axles," in *18th world conference on nondestructive testing*, 2012.

[10] C. Campos-Castellanos, Y. Gharaibeh, P. Mudge and V. Kappatos, "The application of long range ultrasonic testing (LRUT) for examination of hard to access areas on railway tracks," in *5th IET Conference on Railway Condition Monitoring and Non-Destructive Testing (RCM 2011),* 2011.

[11] H. Feng, Z. Jiang, F. Xie, P. Yang, J. Shi and L. Chen, "Automatic Fastener Classification and Defect Detection in Vision-Based Railway Inspection Systems," *IEEE transactions on instrumentation and measurement*, vol. 63, no. 4, pp. 877-888, 2013.

[12] E. Resendiz, J. M. Hart and N. Ahuja, "Automated Visual Inspection of Railroad Tracks," *IEEE transactions on intelligent transportation systems,* vol. 14, no. 2, pp. 751-760, 2013.

[13] F. Flammini, A. Gaglione, F. Ottello, A. Pappalardo, C. Pragliola and A. Tedesco, "Towards Wireless Sensor Networks for railway infrastructure monitoring," *Electrical Systems for Aircraft, Railway and Ship Propulsion*, pp. 1-6, 2010.

[14] F. Moreu, R. E. Kim and B. F. Spencer Jr., "Railroad bridge monitoring using wireless smart sensors," *Railroad bridge monitoring using wireless smart sensors.,* vol. 24, no. 2, p. e1863, 2017.

[15] P. Lu, R. Bridgelall, D. Tolliver, C. Leonard and B. Bhardwaj, "Intelligent Transportation Systems Approach to Railroad Infrastructure Performance Evaluation: Track Surface Abnormality Identification with Smartphone-Based App, MPC-19-384," North Dakota State University - Upper Great Plains Transportation Institute, Fargo: Mountain-Plains Consortium, 2019.

[16] R. Bridgelall, P. Lu, D. Tolliver, N. Dhingra and B. Bhardwaj, "Benefit Cost Analysis of Railroad Track Monitoring Using Sensors Onboard Revenue Service Trains, MPC-21-446," North Dakota State University - Upper Great Plains Transportation Institute, Fargo: Mountain-Plains Consortium, 2021.

[17] B. Bhardwaj, R. Bridgelall, L. Chia, P. Lu and N. Dhingra, "Signal Filter Cut-Off Frequency Determination to Enhance the Accuracy of Rail Track Irregularity Detection and Localization," *IEEE Sensors Journal,* vol. 20, no. 3, pp. 1393-1399, 2019.

[18] R. Bridgelall and D. D. Tolliver, "Railroad accident analysis using extreme gradient boosting," *Accident Analysis & Prevention*, vol. 156, p. 106126, 2021.

[19] B. Bhardwaj, R. Bridgelall, P. Lu, K. E. Nygard and N. Dhingra, "Architecture for an Intelligent Low-Cost Rail Track Condition Evaluation System," in *International Conference on Transportation and Development 2020*, Reston, VA: American Society of Civil Engineers, 2020.

[20] B. Bhardwaj, R. Bridgelall, P. Lu and N. Dhingra, "Signal Feature Extraction and Combination to Enhance the Detection and Localization of Railroad Track Irregularities," *Signal Feature Extraction and Combination to Enhance the Detection and Localization of Railroad Track Irregularities*, vol. 21, no. 5, pp. 6555-6563, 2020.

[21] A. Alemi, F. Corman and G. Lodewijks, "Condition monitoring approaches for the detection of railway wheel defects.," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit,* vol. 231, no. 8, pp. 961-981, 2017.

[22] F. Flammini, A. Gaglione, F. Ottello, A. Pappalardo, C. Pragliola and A. Tedesco, "Towards Wireless Sensor Networks for Railway Infrastructure Monitoring," *Electrical Systems for Aircraft, Railway and Ship Propulsion.,* pp. 1-6, 2010.

[23] H. Tanaka, A. Shimizu and K. Sano, "Development and verification of monitoring tools for realizing effective maintenance of rail corrugation," in *6th IET Conference on Railway Condition Monitoring (RCM 2014)*, IET., 2014.

[24] T. Uhl, K. Mendrok and A. Chudzikiewicz, "Rail track and rail vehicle intelligent monitoring system.," *Archives of Transport,* vol. 22, pp. 495-510, 2010.

[25] T. Real, J. Montrós, L. Montalbán, C. Zamorano and J. I. Real, "Design and validation of a railway inspection system to detect lateral track geometry defects based on axle-box accelerations registered from in-service trains.," *Journal of Vibroengineering,* vol. 16, no. 1, pp. 234-248, 2014.

[26] X. Wei, F. Liu and L. Jia, "Urban rail track condition monitoring based on in-service vehicle acceleration measurements," *Measurement,* vol. 80, pp. 217-228, 2016.

[27] M. Bocciolone, A. Caprioli, A. Cigada and A. Collina, "A measurement system for quick rail inspection and effective track maintenance strategy.," *Mechanical Systems and Signal Processing,* vol. 21, no. 3, pp. 1242-1254, 2007.

[28] P. F. Weston, P. Li, C. S. Ling, C. J. Goodman, R. M. Goodall and C. Roberts, "Track and Vehicle Condition Monitoring during Normal Operation Using Reduced Sensor Sets," *HKIE Transactions* , vol. 13, no. 1, pp. 47-54, 2006.

[29] Y. Jang, J. Kyu Han, S. Young Cho, G.-W. Moon, J.-M. Kim and H. Sohn, "Wireless Power and Data Transfer System for Smart Bridge Sensors," in *2016 IEEE Applied Power Electronics Conference and Exposition (APEC)*, IEEE, 2016.

[30] P. Kołakowski, J. Szelążek, K. Sekuła, A. Świercz, K. Mizerski and P. Gutkiewicz, "Structural health monitoring of a railway truss bridge using vibration-based and ultrasonic methods," *Smart Materials and Structures,* vol. 20, no. 3, p. 035016, 2011.

[31] Z. Yuewen, L. Zhilin, S. Guofeng and Z. Xiyuan, "An innovative vehicle-mounted GPR technique for fast and efficient monitoring of tunnel lining structural conditions.," *Case Studies in Nondestructive Testing and Evaluation,* vol. 6, pp. 63-69, 2016.

[32] M. D. Jenkins, T. Buggy and G. Morison, "An imaging system for visual inspection and structural condition monitoring of railway tunnels," in *2017 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS)*, IEEE, 2017.

[33] C. Wei, Q. Xin, W. H. Chung, S.-y. Liu, H.-y. Tam and S. L. Ho, "Real-time train wheel condition monitoring by fiber Bragg grating sensors," *International Journal of Distributed Sensor Networks,* vol. 8, no. 1, p. 409048, 2011.

[34] G. Kouroussis, D. Kinet, V. Moeyaert, J. Dupuy and C. Caucheteur, "Railway structure monitoring solutions using fibre Bragg grating sensors," *International journal of rail transportation,* vol. 4, no. 3, pp. 135-150, 2016.

[35] M. L. Filograno, P. C. Guillén, A. Rodríguez-Barrios, S. Martín-López, M. Rodríguez-Plaza, Á. Andrés-Alguacil and M. González-Herráez, "Real-time monitoring of railway traffic using fiber Bragg grating sensors," *IEEE Sensors Journal,* vol. 12, no. 1, pp. 85-92, 2011.

[36] P. Weston, C. Roberts, G. Yeo and E. Stewart, "Perspectives on railway track geometry condition monitoring from in-service railway vehicles," *Vehicle System Dynamics,* vol. 53, no. 7, pp. 1063-1091, 2015.

[37] L. Chunsheng, S. Luo, C. Cole and M. Spiryagin, "An overview: modern techniques for railway vehicle on-board health monitoring systems," *Vehicle system dynamics,* vol. 55, no. 7, pp. 1045-1070, 2017.

[38] J. Rabatel, S. Bringay and P. Poncelet, "Anomaly detection in monitoring sensor data for preventive maintenance," *Expert Systems with Applications,* vol. 38, no. 6, pp. 7003-7015, 2011.

[39] M. Grudén, A. Westman, J. Platbardis, P. Hallbjorner and A. Rydberg, "Reliability experiments for wireless sensor networks in train environment," in *2009 European Wireless Technology Conference*, IEEE, 2009.

[40] G. Qin, F. Gu, Y. Xu, F. Liu and A. Ball, "Bogie Speed Estimation and Signal Source Separation via Rail Vibration Analysis," in *30th International Congress & Exhibition on Condition Monitoring and Diagnostic Engineering Management*, 2017.

[41] P. Rolek, S. Bruni and M. Carboni, "Condition monitoring of railway axles based on low frequency vibrations," *International Journal of Fatigue,* vol. 86, pp. 88-97, 2016.

[42] R. Deuce and J. Rosinski, "Ultra miniature data loggers for testing and in-service monitoring of railway axles and other rail vehicle components.," in *IET International Conference on Railway Condition Monitoring*, 2006.

[43] W. Nan, M. Qingfeng, Z. Bin, L. Tong and M. Qinghai, "Research on linear wireless sensor networks used for online monitoring of rolling bearing in freight train," *Journal of Physics: Conference Series,* vol. 305, no. 1, p. 012024, 2011.

[44] P. Li, R. Goodall, P. Weston, C. S. Ling, C. Goodman and C. Roberts, "Estimation of Railway Vehicle Suspension Parameters for Condition Monitoring.," *Control engineering practice,* vol. 15, no. 1, pp. 43-55, 2007.

[45] L. Chia, B. Bhardwaj, P. Lu and R. Bridgelall, "Railroad track condition monitoring using inertial sensors and digital signal processing: A review," *IEEE Sensors Journal,* vol. 19, no. 1, pp. 25-33, 2018.

[46] R. Bridgelall and D. Tolliver, "Accuracy enhancement of anomaly localization with participatory sensing vehicles," *Sensors,* vol. 20, no. 2, p. 409, 2020.

[47] R. Bridgelall, C. A. Leonard, B. Bhardwaj, P. Lu, D. D. Tolliver and N. Dhingra, "Enhancement of signals from connected vehicles to detect roadway and railway anomalies," *Measurement Science and Technology,* vol. 31, no. 3, p. 035105, 2019.

[48] C. Li, S. Luo, C. Cole and S. Spiryagin, "An overview: modern techniques for railway vehicle on-board health monitoring systems," *Vehicle system dynamics,* vol. 55, no. 7, pp. 1045-1070, 55.

[49] S. Zoljic-Beglerovic,, G. Stettinger, B. Luber and M. Horn, "Railway suspension system fault diagnosis using cubature Kalman filter techniques," *IFAC-PapersOnLine,* vol. 51, no. 24, pp. 1330-1335, 2018.

[50] F. Fioretti, E. Ruffaldi and C. A. Avizzano, "A single camera inspection system to detect and localize obstacles on railways based on manifold Kalman filtering," in *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, IEEE, 2018.

[51] P. Pichlík and J. Zděnek, "Extended Kalman filter utilization for a railway traction vehicle slip control," in *2017 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM) & 2017 Intl Aegean Conference on Electrical Machines and Power Electronics (ACEMP)*, IEEE, 2017.

[52] P. Mercorelli, "A hysteresis hybrid extended Kalman filter as an observer for sensorless valve control in camless internal combustion engines," *IEEE Transactions on Industry Applications,* vol. 48, no. 6, pp. 1910-1949, 2012.

[53] H. A. M. Ruiz, P. J. Gräbe and J. W. Maina, "A mechanistic-empirical method for the characterisation of railway track formation," *Transportation Geotechnics,* vol. 18, pp. 10-24, 2019.

[54] Y. Q. Sun, C. Cole and M. Spiryagin, "Monitoring vertical wheel–rail contact forces based on freight wagon inverse modelling," *Advances in Mechanical Engineering,* vol. 7, no. 5, p. 1687814015585431, 2015.

[55] P. Pichlik and J. Zdenek, "Locomotive wheel slip control method based on an unscented Kalman filter," *IEEE Transactions on Vehicular Technology,* vol. 67, no. 7, pp. 5730-5739, 2018.

[56] J. S. Lee, I. Y. Choi, S. Kim and D. S. Moon, "Kinematic modeling of a track geometry using an unscented Kalman filter," *Measurement* , vol. 94, pp. 707-716, 2016.

[57] M. Roth, B. Baasch, P. Havrila and J. Groos, "Map-supported positioning enables in-service condition monitoring of railway tracks," in *2018 21st International Conference on Information Fusion (FUSION)*, IEEE, 2018.

[58] L. Fan, Z. Wang, B. Cail, C. Tao, Z. Zhang, Y. Wang, S. Li, F. Huang, S. Fu and F. Zhang, "A survey on multiple object tracking algorithm," in *2016 IEEE International Conference on Information and Automation (ICIA)*, IEEE, 2016.

[59] M. C. Jeong, S.-J. Lee, K. Cha, G. Zi and J. Sik Kong, "Probabilistic model forecasting for rail wear in seoul metro based on bayesian theory," *Engineering Failure Analysis*, vol. 96, pp. 202-210, 2019.

[60] A. N. Thite, S. Banvidi, T. Ibicek and L. Bennett, "Suspension parameter estimation in the frequency domain using a matrix inversion approach," *Vehicle system dynamics*, vol. 49, no. 12, pp. 1803-1822, 2011.

[61] X. Wei, L. Jia, K. Guo and S. Wu, "On fault isolation for rail vehicle suspension systems," *Vehicle System Dynamics*, vol. 52, no. 6, pp. 847-873, 2014.

[62] L. Bo, S. D. Iwnicki, Y. Zhao and D. Crosbee, "Railway wheel-flat and rail surface defect modelling and analysis by time–frequency techniques," *Vehicle System Dynamics*, vol. 51, no. 9, pp. 1403-1421, 2013.

[63] S. Azadi and A. Soltani, "Fault detection of vehicle suspension system using wavelet analysis," *Vehicle system dynamics*, vol. 47, no. 4, pp. 403-418, 2009.

[64] T. X. Mei and X. J. Ding, "A model-less technique for the fault detection of rail vehicle suspensions," *Vehicle System Dynamics*, vol. 46, no. S1, pp. 277-287, 2008.

[65] R. Lagnebäck, "valuation of wayside condition monitoring technologies for condition-based maintenance of railway vehicles," Doctoral dissertation, Luleå tekniska universitet, 2007.

[66] D. Barke and W. K. Chiu, "Structural health monitoring in the railway industry: a review," *Structural Health Monitoring*, vol. 4, no. 1, pp. 81-93, 2005.

[67] Y. Ouyang, X. Li, C. P. Barkan, A. Kawprasert and Y. Lai, "Optimal locations of railroad wayside defect detection installations," *Computer-Aided Civil and Infrastructure Engineering*, vol. 24, no. 5, pp. 309-319, 2009.

[68] FRA, "Track Inspector Rail Defect Reference Manual," Federal Railroad Administration, 29 July 2015. [Online]. Available: https://railroads.dot.gov/elibrary/track-inspector-rail-defect-reference-manual. [Accessed 25 May 2019].

[69] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations," *Journal of Modern Transportation*, vol. 24, no. 1, pp. 62-72, 2016.

[70] C. Arteaga, A. Paz and J. Park, "Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach.," *Safety Science*, vol. 132, p. 104988, 2020.

[71] W. Al-Nuaimy, A. Eriksen and J. Gasgoyne, "Train-mounted GPR for high-speed rail trackbed inspection," in *Proceedings of the Tenth International Conference on Grounds Penetrating Radar, 2004. GPR 2004.*, IEEE, 2004.

[72] U. Tadeusz, K. Mendrok and A. Chudzikiewicz, "Rail Track and Rail Vehicle Intelligent Monitoring System," *Archives of Transport,* vol. 22, pp. 495-510, 2010.

[73] Z. Xiaozhong, L. Jia, W. Xiukun and N. Ru, "Railway track condition monitoring based on acceleration measurements," in *The 27th Chinese Control and Decision Conference (2015 CCDC)*, IEEE, 2015.

3. USING NON-PARAMETRIC METHODS TO IDENTIFY FACTORS CONTRIBUTING TO RAIL DERAILMENTS

3.1. Abstract

Derailments are the most common rail accidents, accounting for more than 71% of freight train accidents in the United States between 2005 to 2019. A careful analysis of these accidents will provide essential information for developing safety countermeasures. However, most previous research has focused primarily on applying statistical analysis using quantitative data and often overlooked detailed text narratives. Analyzing such narratives can provide detailed information about the accidents and determine causal factors for these accidents. This study explicitly analyzed text narratives using natural language processing techniques and machine learning models to determine factors responsible for causing derailments based on each type of accident. Here, we compared supervised and semi-supervised machine learning algorithms with various models, including random forest, support vector machine, extreme gradient boosting, and K-nearest neighbor. Our results indicate that random forest was the best model in both algorithms but performed better with the supervised technique for text classification and predictions. We tested the validity of our model to analyze narratives' local interpretability by using the Local Interpretable Model-Agnostic Explanations (LIME). Furthermore, we propose a new method of converting and comparing the local results into global interpretability (GLIME) to identify possible causality factors responsible for a freight train derailment. We conducted the association analysis on the top global words and the text narratives to find out the most frequent pair of words used most often in a text narrative describing the accidents. Our results indicate a strong association among a set of terms for each sub-category of the accidents. For example, there are strong associations among journal, burn, and overheat; sill, break, and old; switch,

point, and gapped; between improperly and load; and many others. The results also suggest that the proposed methodology has great potential benefits for classifying rail accidents and developing countermeasures based on factors responsible for causing derailments.

3.2. Introduction

Freight train accidents are more frequent than those in passenger trains [3]. Derailments are the most common type of accident, accounting for more than 71% of U.S. freight train accidents from 2005 to 2019. Most of these accidents were not serious. However, some of these accidents caused more than $1M in losses and severe damages to railway infrastructure and rolling stock, resulting in long-time service disruptions [4]. There are many reasons for accidents, but they can be classified into five major causes. Table 3.1 includes the number of total accidents, freight train-related accidents, and freight train derailments from 2005 to 2019 by accident cause.

Table 3.1. Derailments by major cause from 2005 to 2019

Cause	Total Accidents	%	Freight Train Accidents	%	Freight Train Derailment	%
Mechanical and electrical failures	4,614	10.64	2,603	15.03	2,111	17.02
Miscellaneous	9,553	22.02	3,861	22.29	1,329	10.75
Track, roadbed, and structures	11,587	26.71	6,152	35.52	6,032	48.79
Signal and communication	1,052	2.43	100	0.58	77	0.62
Train operation - human factors	16,572	38.20	4,605	26.59	2,813	22.76
Total	43,378	100.00	17,321	100.00	12,362	100.00

The Federal Railroad Administration (FRA) requires involved companies to complete and submit a detailed report of all the rail accidents that exceed a specified monetary damage threshold (inflation-adjusted 2019 threshold - $10,700) [5] [6]. These accident reports include various standardized fields that describe the accident conditions (temperature, time & date, etc.), location (state, county, etc.), operational factors (speed, number of cars, etc.), and cause. The

reports also include a comprehensive narrative for each accident that provides more information about the actual cause, responsible factors, and circumstances under which the accident occurred. However, these narratives include railroad jargon that non-industry personnel will find difficult to understand [4].

Most previous studies have mainly focused on applying various statistical analyses using qualitative and quantitative data published in FRA accident reports to analyze and highlight reasons for derailments. Although such findings emphasize the causal effect based on explanatory variables, those studies were limited by (i) details captured by the data, (ii) data availability, (iii) how the data are coded, and (iv) the assumptions of the advanced statistical models. All these factors limit the use of quantitative data reports to highlight derailment causes. In contrast, research has overlooked detailed text narratives to determine causal factors not captured in the quantitative coded data. Analyzing these narratives using nonparametric machine learning models provides explicit information about these accidents. As the nonparametric machine learning algorithms are free from statistical intricacies or assumptions, thus preventing limited and biased results [1] [2]. By not making assumptions about data distribution, these algorithms are prepared to select any functional forms from the training data. Russell and Norvig [7] mentioned that non-parametric methods are useful for big data which do not have any prior knowledge about the nature of the data.

Another concern with previous research is that the work used comprehensive accident data as input to draw final conclusions addressing issues related to derailments. Such findings cannot be applied to all accident causes as each cause has a different frequency, severity, responsible factors, consequences, and distinctive level of risk. Individually evaluating and understanding each of the five classes of derailment causes would provide additional insight into

developing policies and safety countermeasures. Subsequently, this study aims to answer the following two research questions: (1) What factors play important roles in resulting derailment for each type of accident? (2) How can text narrative be used as data to reveal undiscovered issues, identify causal factors, and highlight problems related to data structuring?

The main challenge of dealing with accident narrative reports is their unstructured and qualitative nature which precludes statistical analysis and the time required to interpret the implicit meaning of the report. Text mining is at the interface of information systems and linguistics that converts qualitative data into a quantitative form using Natural Language Processing (NLP) techniques. This converted unstructured quantitative data is then used to drive machine learning algorithms for discovering and extracting implicit information to perform operations like retrieval, classification, and summarization [8]. Ge et al. [9] emphasized that, based on data type and research objectives, machine learning algorithms can be classified as unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning. The supervised learning technique trains on the labeled data and then applies the trained models to unseen data to predict labels. In unsupervised learning, the model trains on unlabeled data to classify observations. Semi-supervised models use both labeled and unlabeled data for training and prediction. Reinforcement Learning (RL) develops in an interactive trial and error environment and learns through rewards and penalties for correct and incorrect responses, respectively. RL is useful for gaming, navigation, and robotics, whereas the other three machine learning models are commonly used to facilitate text/data mining and analytics in various industries. In this study, unsupervised learning would not be useful because the accident data has suitable labels. Subsequently, this study utilizes supervised and semi-supervised methods which

will provide supportive evidence to answer the two research questions posed above by evaluating the following four objectives:

1. To extract patterns and implicit information from text narratives of freight train derailments from 2005 to 2019.
2. To compare, evaluate, and identify the best techniques between supervised and semi-supervised learning for text classification and predictions. Text classification will help automatically categorize all the future relevant text about train accidents, which would help determine the related root cause using the methodology described in the study.
3. To explore the factors causing train derailment using local level interpretability of different causes individually and comprehensively using LIME and comparing those results with global interpretability results.
4. To determine how different reporters, narrate similar accidents and highlight some possible limitations of describing and entering accident verbiages into the database.

The findings would help facilitate the processing of text descriptions in different ways to draw conclusions that would be impossible by merely looking at the accident fields. Also, identifying the inconsistencies in accident reporting would help to improve data recording procedures. The next section will discuss the studies that have used accident narratives to extract insights in various fields using supervised and semi-supervised techniques. Also, the next section will include some studies explaining the functionality of LIME and text narratives.

3.3. Literature Review

Data collection is exponentially increasing. Recent estimates suggest that up to 90% of the collected data is stored in semi-structured and unstructured forms such as text [9]. Thus,

novel approaches are needed to analyze and interpret ever-expanding data into meaningful information. Text mining is one such technique that has become prevalent, especially in industries where unstructured qualitative data could help provide implicit information that is useful for decision-making, such as in the fields of security [10, 11], risk management [12], and healthcare [13, 14]. On the other hand, only a handful of studies have applied text data mining techniques in the transportation sector, such as maritime [15, 16, 17, 18]; airline [19, 20]; road traffic, and road conditions [2, 21, 22]. However, the use of text mining in the railroad transportation industry, especially pertaining to railroad accidents, is limited. Thus, studying and evaluating text data provides a unique opportunity to gain insights into transportation operations, policy issues, and accidents.

3.3.1. Text Mining of Railroad Accidents

Soleimani et al. [23] analyzed railroad accident narratives related to Highway-Rail Grade Crossings (HRGCs) to classify the train-vehicle crashes into "train struck car" and "car struck train" using topic modeling and machine learning. Zhao et al. [24] used topic modeling to extract the fault feature from maintenance records with the arbitrary uncertainty and complexity of fault diagnosis using the Bayesian Network (BN). The study aimed to propose a fault diagnosis method for onboard vehicle equipment of a high-speed railway. Williams and Betak [25] used two distinct methods of topic modeling, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), to identify themes in railroad accidents. They concluded that identified switching accidents, hump yard accidents, and grade-crossing accidents were major accident-type topics. Brown [4] applied topic modeling to U.S. railway accident descriptions from 2001 to 2012. The aim was to demonstrate that the combination of text narrative with ensemble machine learning methods could improve the accuracy of accident severity prediction. The “number of

topics" selection, used as an input for running ensemble models, was arbitrary, but the final findings showed that the result was better than using a similar model on a single quantitative data set. In another similar study, Heidarysafa et al. [26] applied deep learning methods together with powerful word embedding on the U.S. railway accident text narrative that occurred between 2001 to 2016. The study's main aim was to extract simpler terms for describing the primary cause of accidents to improve the label accuracy. The results showed that the proposed supervised methodology could classify the cause of a reported accident with an overall 75% accuracy. However, the study did not highlight the root causes of the accidents.

3.3.2. Text Classification Using Supervised Learning

Text classification is one of the various ways of performing text mining. Text classification is a process of assigning text documents into pre-defined classes based on their content. In this process, textual data is initially labeled into pre-defined categories or classes; a learning algorithm is then used to learn a function from this labeled data to assign classes to unseen/unlabeled documents [27]. This approach of building a classifier using pre-labeled data to predict classes to future documents is called supervised learning [28]. Many authors have used supervised text classification to perform operations like topic modeling [29, 25], spam filtering-SMS [30] and email sorting [31], and sentiment analysis [32, 33]. Only a few authors have applied supervised text classification to classify accidents in various industries. For instance, text classification has been used for construction site accident classification [34, 35, 36], secondary crashes classification on roads [21], road accident severity [37, 38], and railway accident classification [4, 23, 26]. All these studies, including many others, have found that supervised machine learning models are efficient in performing text classification. Thus, given the availability of large text data and commonality of classification identifier (accident, cause, etc.),

we compared random forest, SVM, extreme gradient boosting, and KNN supervised machine learning algorithms to evaluate our model's performance.

3.3.3. Text Classification Using Semi-Supervised Learning

The two significant limitations of supervised learning are that it requires a large number of labeled training documents to build an accurate classifier [28], which leaves limited data to make predictions and validate findings, and it requires verification of pre-assigned labels for text classification. Consequently, the semi-supervised classification method in text classification becomes relevant.

Semi-supervised learning is a branch of machine learning which combines the functionality of unsupervised and supervised learning, where the classifier learns from both labeled and unlabeled data. The main objective of the semi-supervised method is to harness un-labeled data for the construction of better learning procedures. A variety of semi-supervised techniques have been applied for classification, clustering, and regression in the past. Also, some previous studies have mentioned that, under certain assumptions and with the limitations of supervised learning methods, semi-supervised learning performs better than supervised learning which is trained on the labeled data alone [39, 40]. Thus, studies have used semi-supervised text classification in various disciplines to provide solutions like automatic bug triage [41], automatic law text classification [42], cross-language text classification [43], identification of transportation mode [44], and anomaly detection in-flight data [45]. However, further research is needed to evaluate the effectiveness of supervised and semi-supervised learning because un-labeled data only performs better if it contains information that is not included in the labeled data or cannot be easily extracted from it [40]. Thus, this study will compare and evaluate the findings of semi-supervised and supervised learning algorithms to identify the best-performing model to

classify rail accidents. The authors are unaware of any research that previously compared the results of these two methods for classifying rail accidents.

3.3.4. Text Mining Using LIME

LIME, proposed by Ribeiro et al. (2016), is an algorithm that tries to explain the black box methods of machine learning at the local level [46]. One of the major challenges associated with machine learning is an incomplete understanding of the model's functionality, limiting confidence in the results. However, LIME uses a model-agnostic approach to perturb the data to generate an output and then, based on proximity, weighs the new data points to fit a local model to explain each local data point. Thus, the model becomes self-explanatory and does not rely on assumptions or the model choice.

LIME has already been applied by various researchers in different disciplines [47, 48, 49], but only a few have used LIME in machine learning-based text classification studies. Sari et al. (2018) applied text mining to reveal the authorship of the documents based on writing style. Along with the ML models, LIME was used to highlight factors associated with classifiers' predictions [50]. Arteaga et al. [2] proposed an analysis approach using similar techniques to identify factors associated with injury severity levels in traffic crashes. The study used yearly data from 2007 to 2017 of heavy vehicle crashes in Queensland, Australia. In their research, the data was initially transformed into vector format using text mining and then compared to the results of six machine learning algorithms. Finally, results showed Neural Networks outperformed the other algorithms, which were further used as a base model to apply global cross-validation-LIME (GCV-LIME). These results helped determine the factors associated with injury severity levels in traffic crashes at the local and global levels.

These previous studies, including others, have shown that LIME helps evaluate the problems at the local level. However, to the authors' best knowledge, no previous study has applied LIME to accident narratives to find the factors responsible for rail accidents. This study is intended to bridge this gap in the literature.

3.4. Methodology

This study consists of five different stages. The first one is cleaning and converting the text narrative into a structured qualitative format to be used as an input for performing machine learning models. The second stage is known as text weighting using The Term Frequency-Inverse Document Frequency (TF-IDF). The third stage consists of running, evaluating, and comparing different models using supervised and semi-supervised machine learning techniques. Subsequently, this part also includes a comparison of the overall performance between supervised and unsupervised results for the purpose of identifying the best classification techniques. In the fourth stage, the contribution of major variables (words, in this case) will be identified. In the final stage, these contributions will be compared with a local explanation using LIME which collectively coalesces to make the global variable important. An overview of the methodology used to perform each of the five stages is illustrated in Figure 3.1.

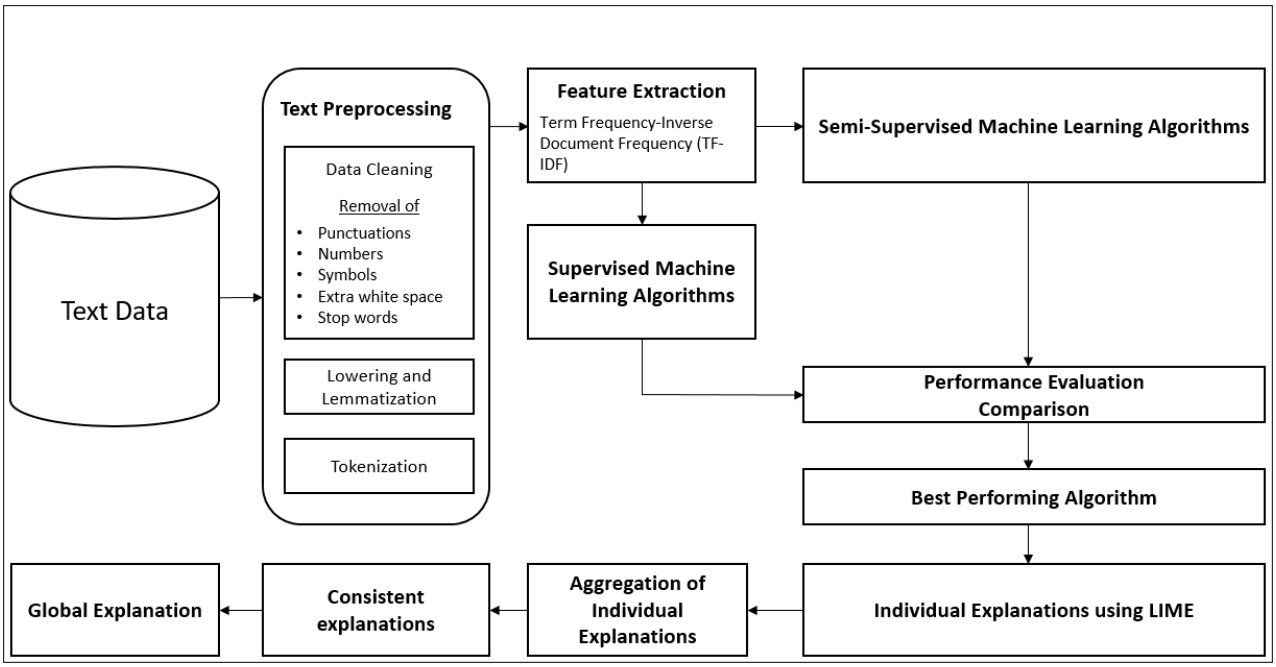

Figure 3.1. Study architecture

3.4.1. Data

The FRA maintains yearly accident data reports on all types of railroad accidents that exceed a specified monetary damage threshold. These reports are maintained in three significant databases. This study uses freight train derailment data from 2005 to 2019 extracted from the Rail Equipment Accidents Database. The extracted data is numeric, alphanumeric, binary, categorical, and free text form collectively recorded in 146 variables. Finally, the study uses text narratives and four-digit alphanumeric codes representing the cause behind these accidents. FRA publishes a separate file explaining the details of each code. The 389 unique codes have an individual description; these descriptions club into 28 different categories and further clustered into five accident cause categories. However, each accident cause does not occur that frequently, and some accidents tend to happen more than others. Subsequently, rare accident categories (frequency less than 100) were excluded from the analysis from all titles. Table 3.2 contains accident categories' names merged into each title and frequency of the freight train derailment in

the last 15 years. Category ID represents a code used in the analysis for the corresponding category under each title. Some of the accident categories were Not-Included (NI) in the analysis because of the rarity of the accidents. Henceforth, accidents related to title 4 (Signal and Communication) were also excluded from the analysis because of the low count. Finally, a complete dataset of 11,955 accidents was used for four titles of accidents.

3.4.2. Text Processing

Like any other text data, railway narratives are unstructured and cannot be used directly as an input for the machine learning models; hence it requires some pre-data processing and some post-structuring scientific adjustments to convert data into a qualitative format. The following are the steps required for the conversion.

3.4.2.1. Text Data Cleaning

Cleaning the raw data is the first step for any analysis as it strengthens data quality. Standard text cleaning begins by deleting unimportant typescripts (punctuations, numbers, and extra white spaces); and converting abbreviations, contractions, and symbols. Further, in the accident's narrative, some words add sense to the verbiage but do not contribute much to quantitative text analysis, commonly known as stop words. A list of pre-defined words from the English language (e.g., and, the, of, for) and the words that might have a distinct meaning for the railway expertise but do not have a general sense were combined as stop words and removed from the text. Excluding such words improves the quality and reduces the size of the data.

3.4.2.2. Normalization: Lowercasing and Lemmatization

Normalization for text processing refers to the transformation of words into a more uniform structure. This is essential in the text mining analysis as it requires algorithms to recognize when two words have a similar meaning, even if they are used differently or with a

different part of the speech [51]. Normalization begins with converting text into lowercase. It is essential; because without this conversion, algorithms would consider two identical words differently if one (capitalized) appears at the beginning of the sentence and another (lowercased) in the middle.

Table 3.2. Accident titles, categories, and frequency from 2005 to 2019

Title ID	Title Name	Categories	Counts	Category ID
1	Mechanical and Electrical Failures	Axles and Journal Bearings	461	10
		Body	157	11
		Brake	176	12
		Coupler and Draft System	207	13
		Doors	42	NI*
		General Mechanical Electrical Failures	7	NI*
		Locomotives	38	NI*
		Trailer Or Container On Flatcar	2	NI*
		Truck Components	462	18
		Wheels	559	19
	Total		2,111	
2	Miscellaneous Causes Not Otherwise Listed	Environment Conditions	381	20
		Loading Procedures	948	21
	Total		1,329	
3	Rack, Roadbed, and Structures	Frogs, Switches, and Track Appliances	918	30
		Other Way and Structure	85	NI*
		Rail, Joint Bar and Rail Anchoring	2414	32
		Roadbed	307	33
		Track Geometry	2308	34
	Total		6,032	
4	Signal and Communication	Signal and Communication	77	NI*
	Total		77	
5	Train operation - Human Factors	Brakes, Use of	107	50
		Employee Physical Condition	1	NI*
		Flagging, Fixed, Hand and Radio Signals	66	NI*
		General Switching Rules	604	53
		Main Track Authority	13	NI*
		Miscellaneous	76	NI*
		Speed	263	56
		Switches	829	57
		Train Handling / Train Make-Up	854	58
	Total		2,813	

*Not Included (counts <100)

The other normalization technique is to convert words into their base forms (stems), and the process of performing such conversion is known as stemming. Stemming is a rule-based technique that converts words into their base by removing the process of eliminating affixes, prefixes, suffixes, and prepositions. Another form of normalization is lemmatization, which works similar to stemming, but instead, it uses a dictionary to replace terms with their morphological root form [51]. In this study, the usage of stemming, lemmatization, and a combination of stemming and lemmatization were applied, but normalization with only lemmatization outperformed the other two possible methods.

3.4.2.3. Tokenization

Tokenization is a process of dividing the text into individual terms called tokens. Generally, tokens are a group of words represented by *n*-consecutive words or *n*-grams [52]. Creating a Document-Term Matrix (DTM) is the process of converting the text into a bag-of-words format. It is a matrix based on a one-row-per-document configuration, i.e., each row represents an individual document, each column with separate terms, and cells indicate frequency describing the number of times each word appears in each document. This procedure's main objective is to transform quantitative data into a categorical format that can then be analyzed using matrix algebra and advanced statistical techniques. However, converting text into a matrix has some challenges including the processing of low-frequency documents. Ignoring such words could result in the overfitting of models. Consequently, eliminating such words would improve the model performance, reduce the dimension of the DTM, and help generalize the results. In this study, the sparse argument is used to remove the lower frequency terms. Sparsity refers to the threshold above which the term will be deleted. The study uses

sparse=0.99, which would remove any tokens missing from 99% of the documents. In other words, each token must appear in at least 1% of the documents to retain.

3.4.3. Feature Extraction Using TF-IDF Term Weighting

DTM converts the data to binary term vectors; i.e., a vector element sets to "one" if the corresponding word appears in a document; zero if it does not. But, converting the words to boolean will make all the words equally important, and in the unstructured text format, all words/tokens are not equally informative for all documents, thus impacting the analysis and findings. Some tokens could also not appear across many documents, which creates a difference between corpora and increase model complexity. Subsequently, to improve the models' efficiency, it is vital to identify words based on their potential influence on the analysis and assign weights accordingly.

Weights can be assigned in two ways: the first is by the number of times a term appears in a document, known as Term Frequency (TF). In this method, each term is important proportional to the number of times it occurs in the text. But in such a weighting scheme, large weights will be assigned to words that appear more frequently but might contain lesser information than the rare ones. The second focuses on the inverse occurrence of the terms across a collection of texts known as Inverse Document Frequency (IDF). This method is based on the belief that terms that rarely occur over a collection of texts have higher importance. Subsequently, each word's importance is inversely proportional to the number of texts containing the term [53]. In other words, if a word frequently appears in one article and rarely in others, that word has good classification capabilities. Cai-zhi et al. [54] explain TF and IDF in their study.

Salton and Buckley [55] combined the concepts of TF and IDF to weight terms and concluded that TF-IDF performs better than using one of the two approaches. The main idea is that if a word or phrase frequently appears in an article and is rarely found in the other article within the same body of work, then the word has distinct classification capabilities.

$$TF - IDF = tf \times idf \quad (4)$$

This study also followed the TF-IDF mainly because of its characteristics of considering each word's frequency in every document, which reduces the effect of low-importance words for the classification task.

3.4.4. Supervised and Semi-Supervised Machine Learning

3.4.4.1. Supervised Learning

For supervised learning, data must be split into a training dataset and a test dataset. The training dataset contains pre-labeled data, which helps in building a classifier, and the test data act as new data, which is used to provide an unbiased evaluation of the classifier that is built using the training dataset. The cross-validation procedure is also used to avoid over-fitting and under-fitting test data. Arteaga et al. [6] explained that a larger number of cross-validation folds should not use in applying machine learning models because it would leave a very small number of records for the testing set. Subsequently, results were generated using cross-validation folds (k=5). This implies that 80% of the data is used for training at every fold, and 20% of the data is used for generating test data. Figure 3.2 displays the structure of the supervised machine learning followed in the analysis.

3.4.4.2. Semi- Supervised Learning

The basic concept behind semi-supervised learning is that a classifier learns from unlabeled data as much as from the labeled data to obtain more accurate models [56, 57]. There are many methods to perform semi-supervised learning for either classification or clustering. In

this paper, we use a wrapper method for the multi-classification task. Keyvanpour and Imani [58] explained that the wrapper method is based on the self-training technique in which the classifier first trained with the reduced set of labeled data to classify the unlabeled instances in the training sample. Pise and Kulkarni [27] further explained that these unlabeled points, together with their predicted labels, are added to the training set. The classifier is then re-trained to predict the labels on unseen or future test data. The first stage setting in the self-training technique is classed as transductive learning, and the second stage is inductive learning [39]. Figure 3.3 represents the functionality for performing the semi-supervised self-training method.

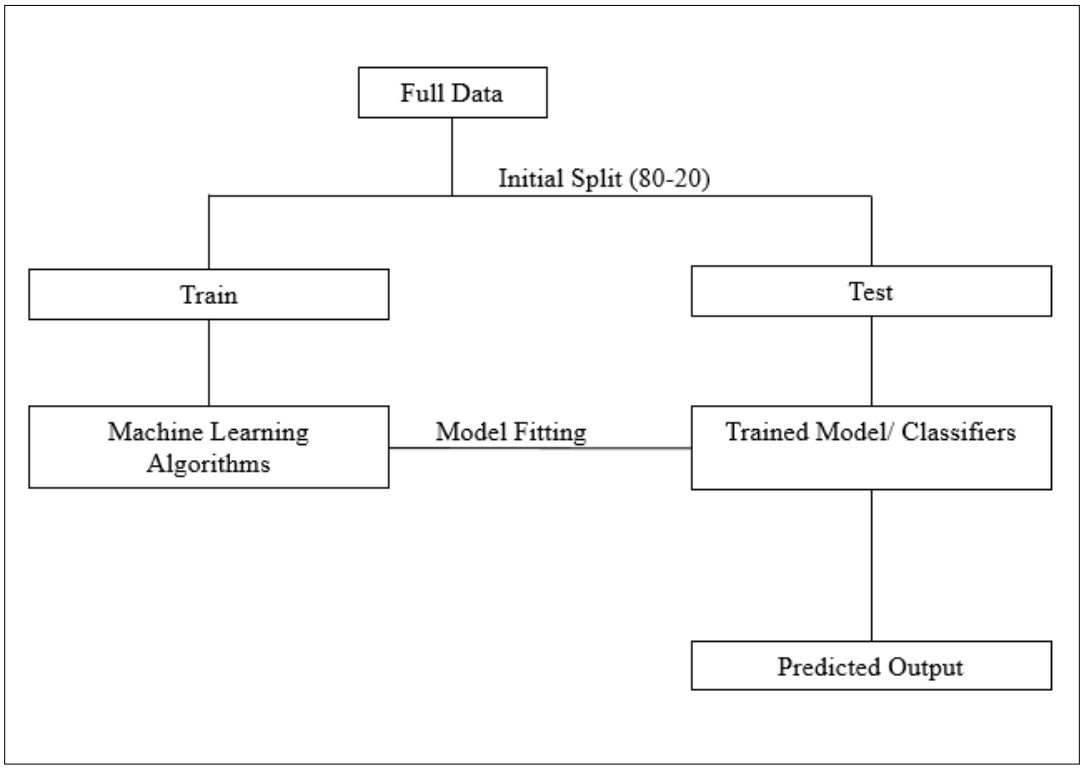

Figure 3.2. Supervised machine learning structure

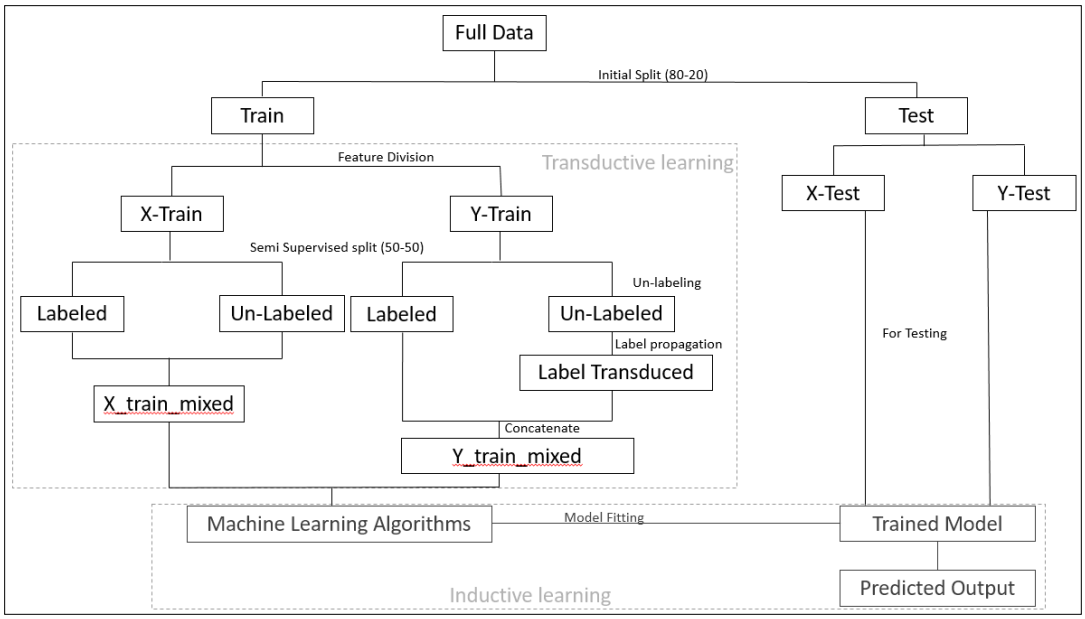

Figure 3.3. Semi-supervised machine learning structure

After the initial split of data into 80% training (Train) and 20% testing (Test), the classes/ labels were separated from the training and test sets and categorized as Y-Train and Y-Test, respectively. Next, both sets of the training dataset are further divided with an equal split of 50-50 again, such that 50% instances have labels and the other 50% becomes unlabeled. Using X-train and corresponding labeled Y-train, label propagation was conducted on the unlabeled data using scikit-learn from python (python-sklearn).

The newly labeled transductive predictions then combined with the initial labeled data to create a new class set called Y_train_mixed, and the corresponding data is then concatenated into X-train data to form X_train_mixed. The Y_train_mixed and X_train_mixed are two matrices that are finally used to perform inductive learning, i.e., testing the final prediction capabilities of the model. Similar to supervised learning, the k-cross-validation procedure is also used to avoid overfitting (k=5).

3.4.5. Machine Learning Algorithms

Machine learning techniques are based on the artificial intelligence system, which seeks to extract patterns from the historical data – known as learning, to make a prediction about the new/unseen data. In many previous research studies, machine learning models were shown to be efficient and effective for performing text classification. This study used several variant techniques for the analysis like random forest, support vector machine, extreme gradient boosting, and k-nearest neighbor.

3.4.5.1. Random Forest

Random forest is a supervised learning technique based on a bagging method in which decision trees are used as a base classifier. The basic concept behind RF is that the crowds' wisdom increases the overall results, i.e., aggregating the predictions of a large number of weak and uncorrelated decision trees will improve the prediction results.

In standard decision trees, each node is split by the best feature, among all other features, using splitting criteria. However, the RF algorithm first selects a random subset of k features for each node and then decides on the best split among these randomly chosen subsets of features. For unknown cases, the prediction is made by aggregating decision tree results using majority voting for classification and averaging for regression tasks [59]. Also, RF is strongly resistant to overfitting because of randomness applied in both sample and feature space [60].

3.4.5.2. Support Vector Machine (SVM)

As explained by Bridgelall [61], a SVM is a supervised non-probabilistic binary linear classifier. Unlike the probabilistic classifiers, where the model considers all data to provide a solution as the probability distribution for each class, SVM uses a small subset of the data (feature vectors) to separate data objects into only two classes - a positive class y_i= +1, and a

negative class y_i= -1. During training, the SVM classifier uses optimization techniques to find the decision boundary (a linear hyperplane) in the feature space to facilitate maximum separation between two classes.

3.4.5.3. Extreme Gradient Boosting (XGBoost)

Boosting algorithms were introduced by the machine learning community mainly to solve classification problems. The main characteristic of boosting is to combine several simple models (week learners) and convert them to a strong and robust classifier. Friedman [62] extended the boosting to the regression and developed a gradient boosting method (GBM). It is a sequential training method focusing on the gradient reduction of the loss function at each stage by adding a new decision tree. The algorithm continues until a given number of iterations is reached [63].

Chen and Guestrin [64] further extended the concept of GBM and introduced extreme gradient boosting (XGBoost). The core improvement is the normalization of the loss function to reduce the model variance [65]. Also, XGBoost has other advantages over the GBM. For instance, XGBoost is fast and more reliable than other decision tree-based machine learning techniques [65], less prone to overfitting [66]; and supports a linear classifier instead of a decision tree, which applies to both the classification and regression.

3.4.5.4. K-nearest neighbor (KNN)

KNN is a lazy learning technique that explores the group of K objects in the closest neighbors among the training data (similar) to objects in the new or test data [67]. Similarly, in the text classification, KKN classifiers discover closest neighbors within learning texts based on the classes defined and then compare them with the test text to give weighting to the texts for identifying their classes [68]. In general, K-NN uses Euclidean distance to categorize the texts into one or more pre-defined classes.

3.4.6. Model Selection

Although we have used a few different machine learning models for making predictions, each model's performance is unique and could be affected by factors like data structure, characteristics of the data, type of the analysis, data volume, and related complexities and many more. The selection of a particular model is by comparison of a selected performance metric required. Subsequently, we use the confusion matrix and a Receiver Operative Curve (ROC) for each set of classes and subclasses to evaluate model performance. The next subsections summarize the details of the confusion matrix and ROC.

3.4.6.1. Confusion Matrix

A confusion matrix is a standard machine learning phenomenon to measure any model's accuracy by comparing predictions versus actual values. In a binary classification problem (which has only two classes to classify), a confusion matrix consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative(FN), which helps in calculating the performance measures including precision, recall, accuracy, and F-1 Score.

- *Accuracy*: fraction of the total number of samples correctly classified by the model.
- *Precision*: fraction of positive predictions that are correct.
- *Recall*: fraction of all positive samples predicted as positive; also known as sensitivity or true positive rate (TPR)
- *F1 Score*: the harmonic mean of precision and recall. The value of the F-1 score lies between 0 and 1.

In multi-class classification, where the matrix has more than two classes to classify, there are no positive or negative classes. Instead, the above features are calculated based on the one-vs-rest transformation, which involves training a single classifier per class, with the samples of

that class as positive and the collective sample of all other classes as negative. The technique requires the base classifier to produce a real-values confidence score for its decision rather than just a class label [69] and can be presented as a micro F1 score.

3.4.6.2. Area Under the Curve (AUC) and Receiver Operative Characteristic (ROC)

In general, ROC is the probability curve, and AUC represents separability measures [69]. In classification problems, AUC- ROC is a performance indicator to visualize the quality of models graphically at various threshold settings. The graph is plotted to indicate the dichotomy of True Positive Rate (TPR) from the positive samples on the y-axis and False Positive Rate (FPR) from the negative samples on the x-axis [59]. Bridgelall [59] explained the concept of ROC in his study and highlighted that less than 0.5 AUC is worse than assigning the classes based on random guessing. Thus, the classifier's main objective is to maximize the AUC and move the operating point as close to the perfect classification as possible.

3.4.6.3. Matthews Correlation Coefficient (MCC)

Accuracy, recall, and F1-score computed on confusion matrices are appropriate methods to evaluate model performances, especially for a binary classification problem. However, such statistical measures make over-optimistic predictions/estimations on the multiclass dataset because of the imbalanced data set between classes (where the number of samples is not the same). An effective solution to overcoming the class imbalance problem is the Matthews Correlation Coefficient (MCC). The MCC calculates the Pearson product-moment correlation coefficient between actual and predicted values ranging from -1 and +1 [70]. A coefficient of +1 represents a perfect classification, -1 as perfect misclassification, and MCC = 0 is as good as random predictions and is calculated using the following equation.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

The MCC considers the balanced ratios of all categories of the confusion matrix (TP, FP, TN, FN) and generates a higher score only if the majority of the positive instances are predicted as positives and negatives instances as negative.

3.4.7. Local Interpretable Model-Agnostic Explanations (LIME)

LIME provides instance-based explanations to the prediction of a classifier f by fitting a simpler, interpretable explanation model g locally around the data point x on which classification is to be explained [71]. In other words, this algorithm provides an explanation of every record in the dataset. The output of each class is predicted using machine learning models, then a simpler model is fitted, and the attributes of this simple model are used as explanations. The simple learned model has a good local approximation of the machine learning model predictions but does not have a good global approximation [72].

For the text classification, LIME basically tests the changes to classification when introducing variations of the data into machine learning models. Variation of the data is generated by randomly removing words from the original text and tracking the performance. The model returns each word's predictive probability in each of the sentence variations, known as the LIME score, along with each word's frequency.

Local explanations are useful in many ways, but in the perspective of the study's objective, identifying keywords that could highlight the factors responsible for causing accidents, the global explanation is more important and adds more logic to the conclusion. LIME has an option of a sub-modular pick (SP-LIME) that offers a global explanation. SP-LIME is the process of running the explanation model on a diverse but representative set of instances that could be considered a global representation of the model. However, for the text analysis of the rail accidents, estimating the true global representation would be more useful than using the

representative sample. Arteaga et al. [2] proposed a new global cross-validation LIME (GCV-LIME) technique for estimating the global explanation. In their study, GCV-LIME is based on an intuitive aggregation strategy of the individual LIME scores, as mentioned in equation 6.

$$Importance\ (t) = \sum_{e \in E} \alpha(t, e) \tag{6}$$

The overall importance of word t in the set of LIME explanation E is calculated as the aggregation of the individual LIME score $\alpha(t, e)$ for word t in explanation $e \in E$. Arteaga et al. [2] concluded that this aggregation strategy is ultimately combining the word frequency and the average of the individual LIME score returned by the LIME explanations and can be represented as equation 7.

$$Importance\ (t) = \sum_{e \in E} I(\alpha(t, e) \neq 0) \times \frac{\sum_{e \in E} \alpha(t, e)}{\sum_{e \in E} (\alpha(t, e) \neq 0)} \tag{7}$$

Where $I(\alpha(t, e) \neq 0$ is the indicator function that returns one when the LIME score of t in e is non-zero. Thus, the first half of equation 7 aggregates the value representing the number of times a word was returned as an explanation. The second half of equation 7 calculates the average of the individual LIME scores.

Although equation 7 helps convert local explanations to the global scale, such results are inappropriate to be considered global influence scores to rank the terms as performed by Arteaga et al. [2]. Such a score represents each word's average impact in local explanations based on each local record. Thus, each word's predictive probability in each sentence variation does not necessarily have the same effect over the variation in the global dataset. Also, in a global dataset, each word's importance is inversely proportional to the number of texts containing the term. Hence, this study will expand the work of Arteaga et al. [2] by including the global word importance (TF-IDF) within the context body of words and called GLIME, which is represented as equation 8

$$Importance\ (t) = \sum_{e \in E} I(\alpha(t,e) \neq 0) \times \frac{\sum_{e \in E} \alpha(t,e)}{\sum_{e \in E}(\alpha(t,e) \neq 0)} \times TF - IDF_G \quad (8)$$

Where $TF - IDF_G$ is the global inverse weight of each term within the corpus identified in stage 3. Including global inverse weights would provide the two benefits. First, the output will represent the importance of each word to a document in a collection or corpus (globally). Second, the findings will be compared in the true sense with the results of the other machine learning models' variable importance at a global scale which otherwise is not possible.

3.5. Results

3.5.1. Influential Words Based on TF-IDF

The TF-IDF weights define the importance of the words in the corpus. However, initial results highlighted some obvious words, including but not limited to “the,” “found,” “see,” and “more” as the most important; however, they have no actual contribution to the analysis. Thus, removing such prominent discriminate words is recommended to evaluate the impact of unique words as factors associated with accidents [2]. Figures 3.4, 3.5, 3.6, 3.7, and 3.8 represent the top 5 tokens/words based on the TF-IDF weights for title 1, title 2, title 3, title 5, and collective data.

The results show that the TF-IDF weights are independent of the number of observations. Additionally, most words appropriately represent their sub-category of accident classes and can be used as tags to categorize the content. For instance, tornado, weather, and gust represent category-20 (environmental conditions); and the coupler, gear, and drawbar symbolize category-13 (coupler and draft systems). Moreover, a few words highlighted the plausible cause of the issue like tread, buildup for causing wheel related accidents. And other several words do not provide any insights or information which could be helpful in either categorizing the accidents or identifying the potential cause. For instance, words such as amt, downloaded, web, and brc.

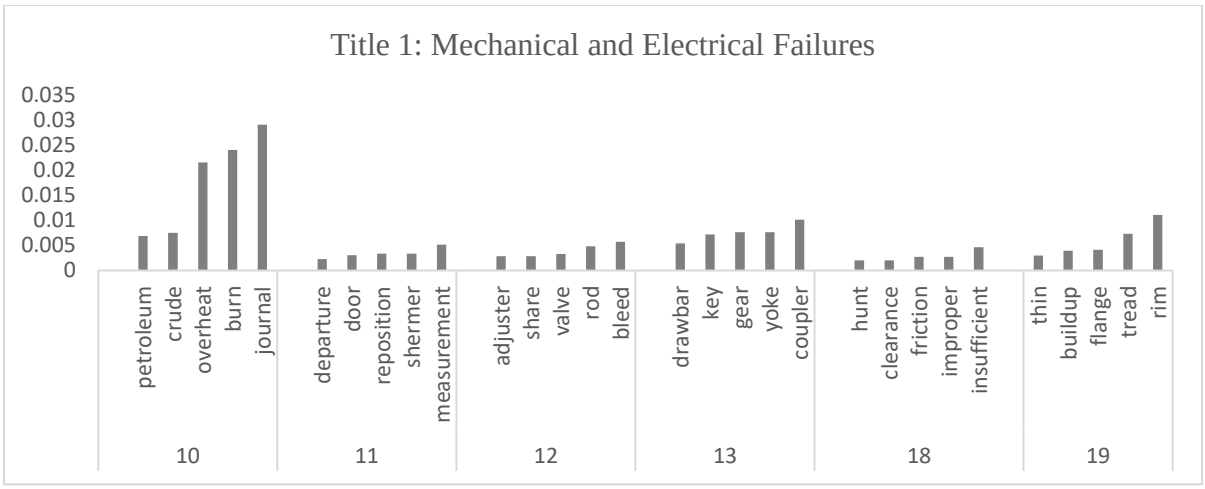

Figure 3.4. Title 1 most influential words based on tf-idf weights

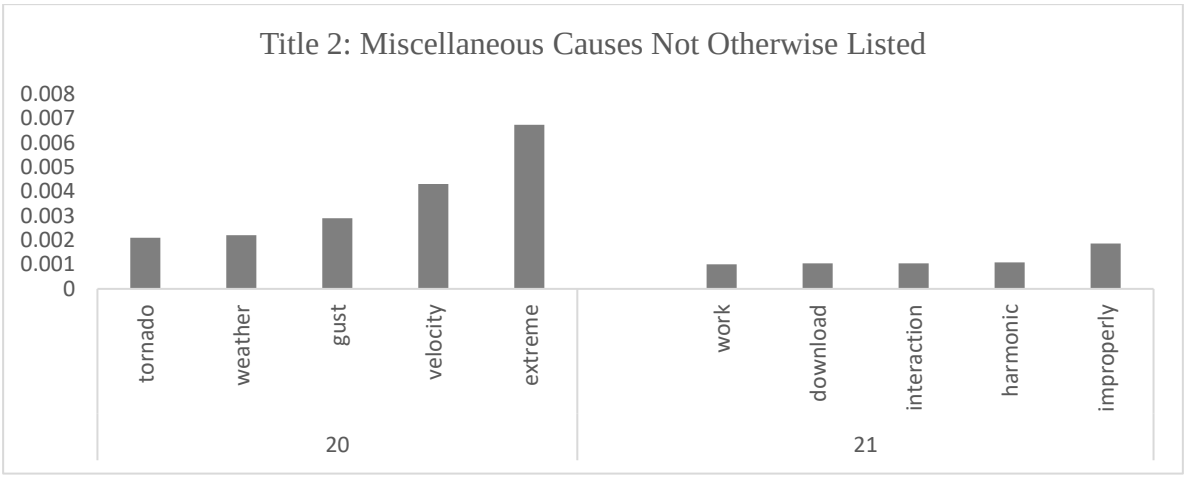

Figure 3.5. Title 2 most influential words based on tf-idf weights

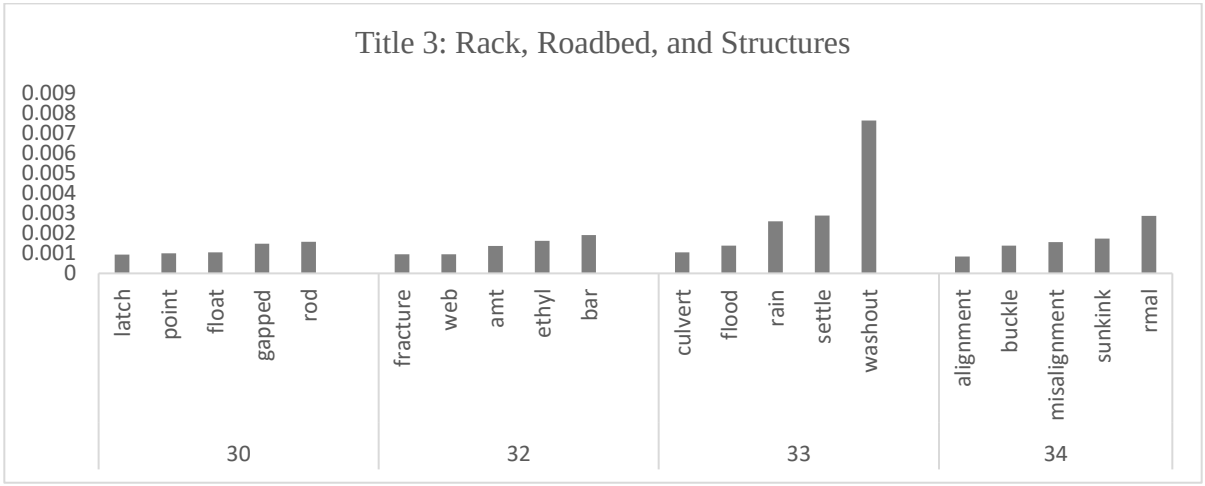

Figure 3.6. Title 3 most influential words based on tf-idf weights

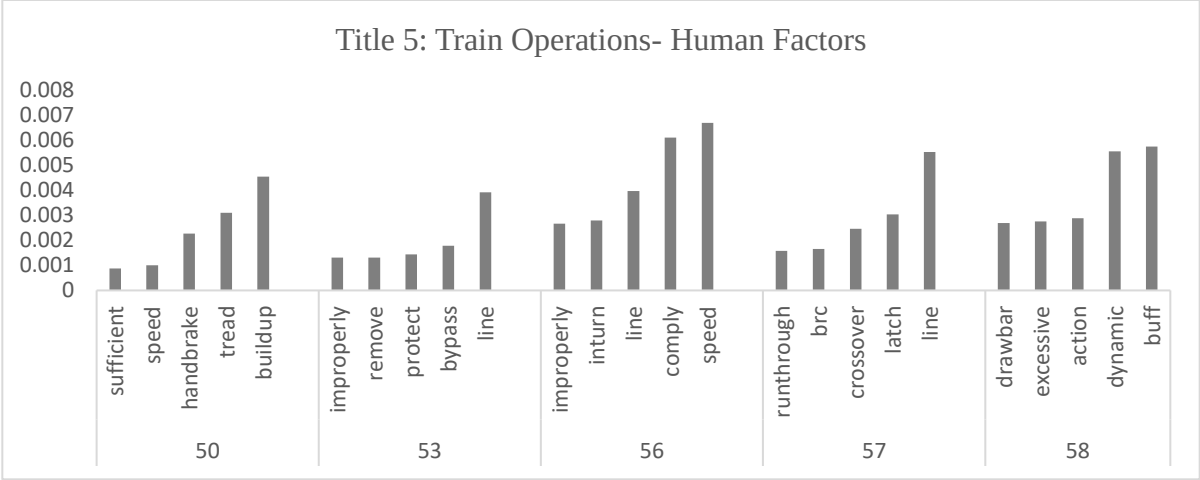

Figure 3.7. Title 5 most influential words based on tf-idf weights

3.5.2. Machine Learning Algorithm Selection

A python library, Scikit-Learn, was used to compare the effectiveness of RF, SVM, XGBoost, and KNN for semi-supervised and supervised techniques. Different hyper-parameters were tuned using a grid search process for the different algorithms. We used stratified splitting for each category to ensure the same proportion of samples were available for the training and testing data.

Table 3.3 compares the predictive performance of supervised and semi-supervised machine learning algorithms using Micro F1, averaged ROC, and MCC. Results are consistent for the different classifiers. There are no explicit explanations available for using the averaged ROC to evaluate algorithm performances for classifying multiclass imbalanced data, hence the final evaluation was deemed to be based on the Micro F1 and MCC.

As per the results in Table 3.3, supervised learning techniques performed better than semi-supervised learning, indicating that the inclusion of unlabeled data in the classifier degrades the algorithms' prediction performances. The prevalence of such performance degradation is likely under-reported due to publication bias [73, 40]. Additionally, RF models outclassed all other algorithms in every category; therefore, this technique was used to perform LIME analysis.

3.5.3. Lime

The local explanations were generated using RF and k= 5 as the cross-validation folds. The train and test dataset split ratio is consistently maintained as 80-20%. Figure 3.8 shows an example of the prediction of an individual explanation using LIME. The output in this figure includes the prediction probability of an observation belonging to a correct class (positive) or otherwise (negative). In addition, the explanation includes the words contributed in favor of and against the overall prediction probabilities and their weights. For instance, in Figure 3.8, we can see that the words “winds,” “severe” “weather,” and “articulated” all contributed to the final prediction score of 0.58.

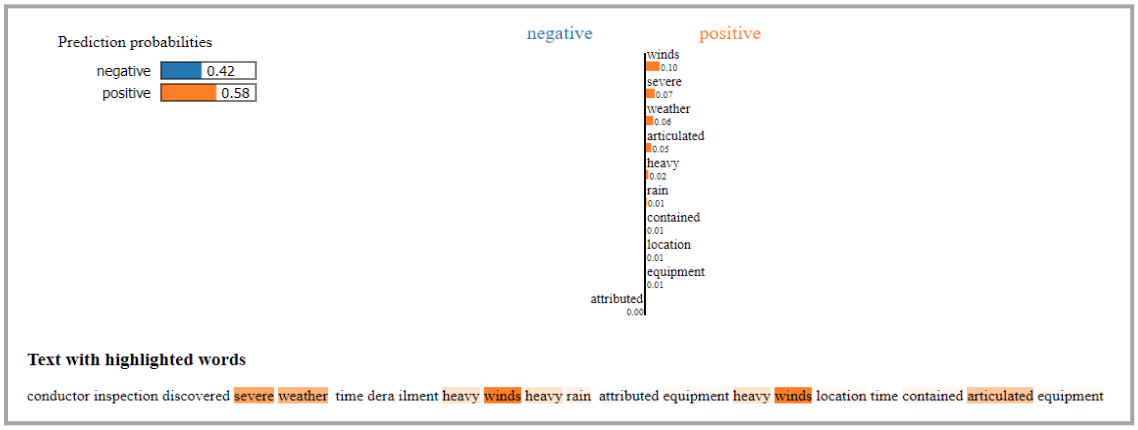

Figure 3.8. Prediction results of an individual explanation using LIME

Similarly, in figure 3.9, “lateral,” “forces,” “vertical,” and “pulling” contributed against the prediction probabilities for the positive classification, which means that this single variable (a text explanation in this case) should not be classified under the tested category.

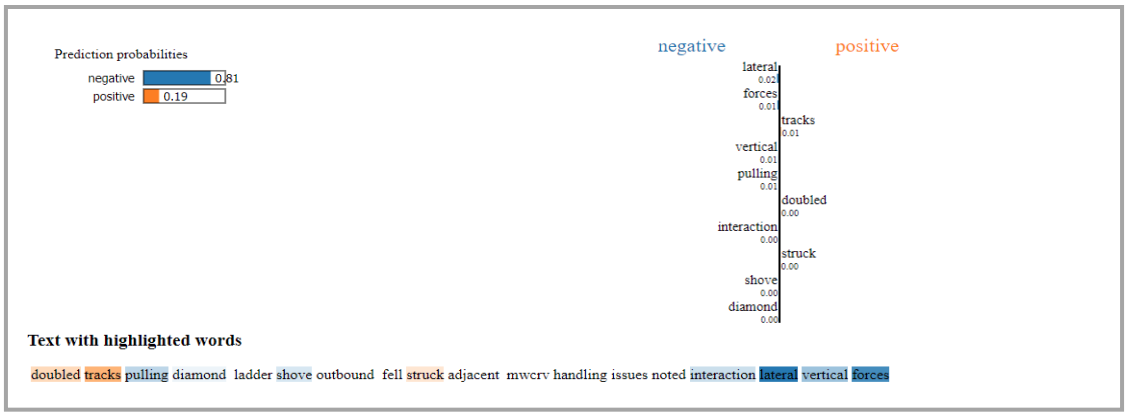

Figure 3.9. Prediction results of an individual explanation using LIME

Table 3.3. Machine learning algorithm results

	Supervised Learning			Semi-Supervised Learning		
Cause 1						
	Micro F1	MCC	Avg ROC	Micro F1	MCC	Avg ROC
RF	**0.71**	**0.66**	**0.86**	0.66	0.59	0.9
SVM	0.63	0.56	0.86	0.58	0.5	0.9
XGBoost	0.68	0.62	0.86	0.63	0.56	0.9
KNN	0.58	0.49	0.86	0.51	0.41	0.9
Cause 2						
	Micro F1	MCC	Avg ROC	Micro F1	MCC	Avg ROC
RF	**0.86**	**0.71**	**0.89**	0.78	0.56	0.78
SVM	0.83	0.65	0.88	0.77	0.54	0.79
XGBoost	0.86	0.7	0.89	0.8	0.59	0.79
KNN	0.74	0.47	0.89	0.76	0.53	0.78
Cause 3						
	Micro F1	MCC	Avg ROC	Micro F1	MCC	Avg ROC
RF	**0.71**	**0.61**	**0.86**	0.7	0.6	0.86
SVM	0.68	0.58	0.86	0.69	0.59	0.86
XGBoost	0.69	0.6	0.86	0.68	0.58	0.86
KNN	0.61	0.48	0.86	0.62	0.49	0.86
Cause 5						
	Micro F1	MCC	Avg ROC	Micro F1	MCC	Avg ROC
RF	**0.71**	**0.64**	**0.87**	0.62	0.53	0.85
SVM	0.65	0.57	0.88	0.62	0.52	0.87
XGBoost	0.69	0.61	0.88	0.64	0.54	0.87
KNN	0.61	0.51	0.88	0.53	0.42	0.87

Table 3.4 shows the sample of aggregation of LIME explanations for each type of accident. The actual results include words, the number of times each word is returned as an

explanation from all of the folds (count), the summation of the LIME score (sum), and an average of the LIME score. Note that due to space limitations, Table 3.4 shows only the top five words of each rail accident category. However, the complete list of words would be able to highlight more insights highlighting different causes of rail accidents.

The global scores for each subcategory were calculated using Equation 4 and are represented in Table 3.5. Several independent words are sufficient enough to highlight the real issues, as shown in Table 3.5. For instance, “extreme,” “gust,” “tornado,” and “weather” are clearly highlighting the accidents associated with the bad weather. Similarly, “settle,” “washout,” “flood,” and “soft” are highlighting the impact of floods on the track conditions. On the other hand, there are some ambiguous words that do not provide additional insights about the core issues. For example, “protect,” “stand,” and “pass.” One of the possible reasons is that the study is using word frequency to extract explanations which could be improved in the future by implying more semantic information, as highlighted by Arteaga et al. [2] and Cambria and White 2014 [74] .

To add significance to the analysis, this study also found the commonly associated terms frequently used with each other. Although various correlation coefficients were tested, 0.2 was used as the threshold. A lower threshold resulted in identifying relationships mainly with misspelled words (spelling mistakes, short forms, jargon) which were not useful for understanding the causal variables. Greater than 0.2, we risk losing the significant hidden relationships between words. Due to space limitations, Table 3.6 represents only a sample of the most significant words (from the LIME results) and their relationships with two variables (words). The full list of the top LIME words and associated variables are presented in appendix.

The results in Table 3.6 highlight some words, when used together or in a single sentence that are useful in identifying the cause of train accidents. For instance, the, the words "journal," "burn," and "overheat" represent the accidents that happened because of the problem of overheating, which resulted in burnt journal and caused the derailment. These narratives from the raw data could help us understand this relationship better.

- *Narrative 1 "DERAILMENT OF CSX DETOUR TRAIN DTR-29 AT MP 101.9 BECAUSE OF BURNT OFF JOURNAL (ROLLER BEARING-FAILURE FROM OVERHEATING)"*
- *Narrative 2 "MRVP0T-17 TRIPPED THE DRAGGING EQUIPMENT DETECTOR AT MILE POST 372.9. UPON INSPECTION, IT WAS DISCOVERED THAT CAR UTLX640581 HAD DERAILED AND HAD AN AXLE JOURNAL BURNED OFF."*

Similarly, "stiff," "bolster," and "improper" represents derailment due to stiff bolster and improper swiveling (refer to appendix). The narratives from the raw data could also help in relating the words with the accident narratives.

- *Narrative 3 "GSC5TU-25 HAD LEADING WHEELS ON CAR UP72635 DERAIL DUE TO A TRUCK BOLSTER STIFF, IMPROPER SWIVELING."*
- *Narrative 4 "CREW DERAILED ONE CAR WHILE DRAGGING OUT OF YARD BECAUSE DERAILED CAR HAD A STIFF BOLSTER AND WAS NOT SWIVELING PROPERLY. TRACK IS NOT CWR."*

The other group of words mentioned in table 3.6 also highlights valuable information about factors that are responsible for causing the derailment. These results could be useful in many different ways. The first is that results could be useful in categorizing the accidents based on each subcategory. Second, the results help in creating a database of the factors responsible for

causing derailments based on micro level. The database could be useful in developing some countermeasures to reduce the number of accidents and the severity of these accidents. The third methodology could also be helpful in identifying mistakes while labeling the accidents categories by the engineer. For instance, the text description in narrative 5 is labeled under C558 (Train handling/Train-Makeup), but the text highlights a different reason for the accident.

- *Narrative 5 "BURNT OFF JOURNAL ON 65TH CAR CAUSED DERAILMENT OF CARS."*

All the analyses done in the study helped illustrate the benefits of conducting text mining and helped in answering the first three objectives of the study. So, now we can proceed to the final question of the study to determine how different reporters highlight the same factors differently and how this study is useful in handling a different part of speech challenge. For instance,

- *Narrative 6 "GSBWMD-17 SPOTTING TRAIN TO FRONTIER COOP (GRAIN ELEVATOR) WHEN FCTX1028 CLIMBED THE SWITCH POINT AND DERAILED. 4 MORE CARS ALSO DERAILED. FRONTIER COOP MAINTANS TRACK. UP'S ML TRK ALSO DAMAGED BY JACKKNIFED CARS."*

Table 3.4. Aggregation of LIME explanations for each type of accident

LIME Results for C110		
Words	Count	Average
journal	46	0.1630
axle	38	0.1201
burn	25	0.1545
bear	27	0.0791
failure	24	0.0882
LIME Results for C111		
Words	Count	Average
sill	11	0.1668
ton	18	0.0319
plate	13	0.0315
body	3	0.0657
old	5	0.0282
LIME Results for C112		
Words	Count	Average
brake	31	0.0867
rig	7	0.1732
air	21	0.0483
hose	8	0.1244
valve	4	0.1020
LIME Results for C113		
Words	Count	Average
drawbar	19	0.1598
coupler	8	0.1461
pin	8	0.1093
fall	13	0.0482
miss	10	0.0533
LIME Results for C118		
Words	Count	Average
truck	61	0.0628
side	66	0.0317
climb	23	0.0851
pull	68	0.0274
car	285	0.0055
LIME Results for C119		
Words	Count	Average
wheel	106	0.0779
break	93	0.0512
rim	14	0.1598
tread	18	0.1238
flange	18	0.1216

LIME Results for C220		
Words	Count	Average
wind	20	0.1546
snow	9	0.1219
extreme	8	0.1319
release	30	0.0295
high	14	0.0527
LIME Results for C221		
Words	Count	Average
switch	54	0.0334
active	1	0.0215
force	10	0.0206
bottom	1	0.0195
yard	28	0.0185
LIME Results for C330		
Words	Count	Average
switch	185	0.0838
point	111	0.0788
yard	207	0.0139
climb	32	0.0585
shove	150	0.0124
LIME Results for C332		
Words	Count	Average
rail	408	0.0713
break	285	0.0934
emergency	176	0.0176
head	161	0.0169
car	653	0.0041
LIME Results for C333		
Words	Count	Average
roadbed	23	0.0842
soft	18	0.0900
settle	11	0.1320
rain	14	0.0569
heavy	11	0.0296
LIME Results for C334		
Words	Count	Average
wide	98	0.0891
gage	69	0.0771
gauge	53	0.0782
irregular	32	0.0984
level	28	0.0814

LIME Results for C550		
Words	Count	Average
roll	29	0.0794
hand	14	0.0640
secure	8	0.0895
handbrake	8	0.0805
sufficient	4	0.0882
LIME Results for C553		
Words	Count	Average
shove	201	0.0482
track	287	0.0117
fail	73	0.0313
control	25	0.0563
protect	18	0.0723
LIME Results for C556		
Words	Count	Average
speed	34	0.0985
restrict	14	0.0767
run	117	0.0083
mph	23	0.0309
failure	32	0.0208
LIME Results for C557		
Words	Count	Average
switch	239	0.0809
run	124	0.0523
pull	199	0.0237
line	142	0.0313
movement	70	0.0275
LIME Results for C558		
Words	Count	Average
brake	60	0.0757
excessive	39	0.0662
slack	27	0.0838
use	33	0.0677
curve	29	0.0741

Table 3.5. Results of word association based on the global LIME score

C110			**C220**			**C550**		
journal	burn (0.46)	overheat (0.27)	extreme	velocity (0.71)	environmental (0.65)	hand	brake (0.88)	apply (0.45)
burn	journal (0.46)	mile (0.26)	velocity	extreme (0.71)	environmental (0.46)	secure	properly (0.5)	sufficiently (0.44)
C111			**C221**			**C553**		
sill	break (0.48)	old (0.38)	improperly	load (0.34)		failure	control (0.72)	improperly (0.42)
plate	rigid (0.59)	center (0.42)	load	improperly (0.34)	empty (0.35)	Shove	track (0.33)	protect (0.32)
C112			**C330**			**C556**		
hose	air (0.78)	separation (0.56)	gapped	point (0.62)	switch (0.42)	speed	excessive (0.44)	rock (0.37)
valve	malfunction (0.81)	tread (0.37)	stock	rail (0.73)	point (0.56)	improperly	line (0.66)	switch (0.63)
C113			**C332**			**C557**		
gear	draft (0.85)	broken (0.8)	break	rail (0.62)	bar (0.41)	switch	line (0.48)	point (0.4)
yoke	broken (0.72)	drawbar (0.62)	bar	joint (0.57)	break (0.41)	through	run (0.74)	previously (0.31)
C118			**C333**			**C558**		
stiff	bolster (0.45)	improper (0.41)	settle	roadbed (0.49)	single (0.38)	buff	force (0.49)	excessive (0.34)
insufficient	clearance (0.57)	bear (0.51)	temporary	speed (0.45)	mph (0.33)	brake	use (0.53)	dynamic (0.48)
C119			**C334**					
buildup	slag (0.45)	tread (0.4)	thermal	misalignment (0.82)	traverse (0.32)			
tread	build (0.52)	buildup (0.4)	buckle	track (0.41)	sunkink (0.35)			

- *Narrative 7 "MAMSKCK1 08A CREW WAS IN THE PROCESS OF SETTING OUT 57 CARS FROM 8102 TRACK TO 8101 WHEN CEFX 350002, B-END WENT BEHIND THE SWITCH POINT OF THE 8106 TRACK SWITCH. IT ENCOUNTERED THE 8105 SWITCH POINT AND THE OUTSIDE WHEEL STRUCK THE SWITCH POINT CAUSING SWITCH TO BE LINED FROM 8105 TRACK. CEFX 35 0002 STRATTLED THE TRACKS FROM THE LEAD TO 8105 DERAILING THE WWLX 961142, NO HAZARDOUS MATERIALS LE AKING."*

Narrative 6, and Narrative 7describe the same subcategory of the accident but very differently. Without a careful reading of these narratives or without talking to the safety

engineer, it would be difficult to examine the reasons behind these accidents. Consequently, a further micro-level analysis, as presented in this study, is required. Also, the results herein help in identifying many typographical errors, incomplete descriptions, overuse of jargon, and even missing text information which requires specific attention to improve the identification of causal factors resulting in these accidents.

3.6. Summary and Conclusion

This study uses a text narratives-based analytical approach using text mining applications, machine learning models, and interpretable machine learning for each sub-category to identify factors responsible for rail accidents. Thus, we propose a new approach called GLIME, a technique used to convert the local into the global explanations by aggregating the individual LIME explanation multiplied by the global TF-IDF values.

Using different machine learning models, we used text narratives describing freight train derailments of Class I railroad accidents from 2005 to 2019 to compare supervised and unsupervised machine-learning algorithms. Our results indicate that the supervised algorithm and the random forest model performed best for conducting text-based classification. A plausible reason that Random Forest was the best model is that it uses both ensemble methods and feature randomness to create an uncorrelated forest of decision trees. Both procedures collectively add more randomness to the model and allow splitting the node based on the best feature among a random subset of features while growing the trees. This reduces the propensity to overfit to the training data, which yields a more generalized model for improved predictions on unseen data.

Then, we used the best-performing algorithm and model in running LIME to identify local explanations highlighting factors associated with train derailments. LIME results are helpful in explaining the words' weights and level of effects at the local level. Still, these words

do not necessarily have the same impact on the global corpora, and thus, we used GLIME to explain the importance of each word to a document in a collection or corpus (globally). In essence, we found a correlation between the top variable (Table 3.4) and global explanatory variables (Table 3.6). For instance, the correlation between the words “journal,” “burn” (0.46), and “overheat” (0.27) represents that while defining the accidents related to axle and journal bearing, journal and bearing are generally used 46% of the time in the narration together and 27% times of the time with the word overheat. It also helps understand that the occurrence of accidents because of overheating, which resulted in a burnt journal, is relatively high.

A possible limitation of the study is using TF-IDF weights for the analyses rather than schematics from the narrative, which could highlight some unimportant words to be significant in predicting classification tasks. In the future, we will counter this issue by transforming the text narrative into more structured data or by using further advances in natural language processing (NLP). Another challenge is the variation and uncertainty inherent with machine learning, text mining techniques, and LIME. To minimize the variation, we will use tabular data and text narratives to formally characterize the results.

Importantly, this study showed that as compared to previous methods, text narratives can more robustly identify accident contributors even at the local level, as there is a lower systemic error rate in the text narrative compared to tabular data. In addition, the methodology also provides solutions to counter errors, including typographical errors and data inconsistencies. Moreover, this methodology will help researchers, industry engineers, and other personnel identify the main factors responsible for causing rail accidents and develop better-targeted countermeasures to prevent these accidents to move toward achieving FRA's zero accident vision.

3.7. Reference

[1] E. Bernal, M. Spiryagin and C. Cole, "Onboard Condition Monitoring Sensors, Systems and Techniques for Freight Railway Vehicles: A Review," *IEEE Sensors Journal,* vol. 19, no. 1, pp. 4-24, 2018.

[2] D. E. Brown, "Text Mining the Contributors to Rail Accidents," *IEEE Transactions on Intelligent Transportation Systems,* vol. 17, no. 2, pp. 346-355, 2015.

[3] L. Xiang, C. P. Barkan and M. R. Saat, "Analysis of derailments by accident cause: evaluating railroad track upgrades to reduce transportation risk.," *Transportation research record* , vol. 2261, no. 1, pp. 178-185, 2011.

[4] FRA, "Monetary Threshold Notice," Federal Railroad Administration, [Online]. Available: https://railroads.dot.gov/forms-guides-publications/guides/monetary-threshold-notice. [Accessed 15th August 2020].

[5] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations.," *Journal of Modern Transportation* , vol. 24, no. 1, pp. 62-72, 2016.

[6] C. Arteaga, A. Paz and J. Park, "Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach.," *Safety Science,* vol. 132, p. 104988, 2020.

[7] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Pearson Education Limited, 2014, p. 757.

[8] A. Khan, B. Baharudin, L. H. Lee and K. khan, "A review of machine learning algorithms for text-documents classification.," *JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY,* vol. 1, no. 1, pp. 4-20, 2010.

[9] G. Zhiqiang, Z. Song, S. X. Ding and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *Ieee Access,* vol. 5, pp. 20590-20616, 2017.

[10] E. W. Ngai, Y. Hu, Y. H. Wong, Y. Chen and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature.," *Decision support systems,* vol. 50, no. 3, pp. 559-569, 2011.

[11] G. L. Gray and R. S. Debreceny, "A taxonomy to guide research on the application of data mining to fraud detection.," *International Journal of Accounting Information Systems,* vol. 15, no. 4, pp. 357-380, 2014.

[12] M. Jin, Y. Wang and Y. Zeng, "Application of Data Mining Technology in Financial Risk Analysis.," *Wireless Personal Communications,* vol. 102, no. 4, pp. 3699-3713, 2018.

[13] G. Mihaela and R. Petre, "Integrating data mining techniques into telemedicine systems.," *Informatica Economica,* vol. 18, no. 1, p. 120, 2014.

[14] K. Srinivas, B. K. Rani and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks.," *International Journal on Computer Science and Engineering (IJCSE),* vol. 2, no. 02, pp. 205-255, 2010.

[15] L. Wang, J. Wang, M. Shi, S. Fu and M. Zhu, "Critical risk factors in ship fire accidents.," *Maritime Policy & Management,* pp. 1-19, 2020.

[16] A. Galieriková, "The human factor and maritime safety.," *Transportation research procedia,* vol. 40, pp. 1319-1326, 2019.

[17] S. Xin, D. Xu, H. Zhuang and C. Liu, "How unsafe acts occur: an automatic text mining study.," *Maritime Policy & Management,* pp. 1-11, 2021.

[18] S. Tirunagari, "Data mining of causal relations from text: analysing maritime accident investigation reports.," *arXiv preprint arXiv:1507.02447.,* 2015.

[19] V. De Vries, "Classification of Aviation Safety Reports using Machine Learning.," in *2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT).,* 2020.

[20] O. Sjöblom, "Data Mining in Promoting Aviation Safety Management.," in *International Conference on Well-Being in the Information Society.,* 2014.

[21] X. Zhang, E. Green, M. Chen and R. R. Souleyrette, "Identifying secondary crashes using text mining techniques.," *Journal of Transportation Safety & Security,* pp. 1-21, 2019.

[22] R. Nayak, N. Piyatrapoomi and J. Weligamage, "Application of text mining in analysing road crashes for road asset management.," *In: Kiritsis D., Emmanouilidis C., Koronios A., Mathew J. (eds) Engineering Asset Lifecycle Management.,* pp. 49-58, 2010.

[23] S. Soleimani, A. Mohammadi, J. Chen and M. Leitner, "Mining the Highway-Rail Grade Crossing Crash Data: A Text Mining Approach," in *In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA),* 2019.

[24] Y. Zhao, T.-h. Xu and W. Hai-fend, "Text mining based fault diagnosis of vehicle on-board equipment for high speed railway," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC).,* 2014.

[25] T. Williams and J. Betak, "A Comparison of LSA and LDA for the Analysis of Railroad Accident Text.," *Procedia computer science,* vol. 130, pp. 98-102, 2018.

[26] M. Heidarysafa, K. Kowsari, L. E. Barnes and D. E. Brown, "Analysis of Railway Accidents' Narratives Using Deep Learning.," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA),* 2018.

[27] N. N. Pise and P. Kulkarni, "A Survey of Semi-Supervised Learning Methods," *2008 International conference on computational intelligence and security.* , vol. 2, pp. 30-34, 2008.

[28] B. Liu, W. S. Lee, P. S. YU and X. Li, "Partially supervised classification of text documents.," *ICML*, vol. 2, no. 485, pp. 387-394, 2002.

[29] Z. Weizhong, C. J. James , R. Perkins, Z. Liu, W. Ge, Y. Ding and W. Zou, "A heuristic approach to determine an appropriate number of topics in topic modeling," *BMC bioinformatics*, vol. 16, no. 13, pp. 1-10, 2015.

[30] M. T. Nuruzzaman, C. Lee and D. Choi, "Independent and personal SMS spam filtering.," in *2011 IEEE 11th International Conference on Computer and Information Technology. IEEE*, 2011.

[31] D. Gaurav, S. M. Tiwari, A. Goyal, N. Gandhi and A. Abraham, "Machine intelligence-based algorithms for spam filtering on document labeling.," *Soft Computing*, vol. 24, no. 13, pp. 9625-9638, 2020.

[32] S. Zeenia, S. Randhawa and S. Jain, "Sentiment analysis of customer product reviews using machine learning.," in *IEEE*, 2017.

[33] P. Baid, A. Gupta and N. Chaplot, "Sentiment analysis of movie reviews using machine learning techniques.," *Sentiment analysis of movie reviews using machine learning techniques.*, vol. 179, no. 7, pp. 45-49, 2017.

[34] M.-Y. Cheng, D. Kusoemo and R. A. Gosno, "Text mining-based construction site accident classification using hybrid supervised machine learning," *Automation in Construction*, vol. 118, p. 103265, 2020.

[35] F. Zhang, H. Fleyeh, X. Wang and M. Lu, "Construction site accident analysis using text mining and natural language processing techniques.," *Automation in Construction*, vol. 99, pp. 238-248, 2019.

[36] Y. M. Goh and C. U. Ubeynarayana, "Construction accident narrative classification: An evaluation of text mining techniques.," *Accident Analysis & Prevention*, Vols. 122-130, p. 108, 2017.

[37] J. Li, J. Guo and M. Qiu, "Injury Severity Analysis of Secondary Incidents. In 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)," in *IEEE*, 2020.

[38] C. Arteaga, A. Paz and J. Park, "Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach.," *Safety Science*, vol. 132, p. 104988, 2020.

[39] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning.," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1-130, 2009.

[40] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning.," *Machine Learning*, vol. 109, no. 2, pp. 373-440, 2020.

[41] J. Xuan, H. Jiang, Z. Ren, J. Yan and Z. Luo, "Automatic bug triage using semi-supervised text classification.," *arXiv preprint arXiv:1704.04769*, 2017.

[42] P. Li, F. Zhao, Y. Li and Z. Zhu, "Law text classification using semi-supervised convolutional neural networks.," in *IEEE*, 2018.

[43] L. Shi, R. Mihalcea and M. Tian, "Cross language text classification by model translation and semi-supervised learning.," in *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[44] S. Dabiri, C.-T. Lu, K. Heaslip and C. K. Reddy, "Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data.," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 5, pp. 1010-1023, 2019.

[45] M. Memarzadeh, B. Matthews and T. Templin, "Multi-Class Anomaly Detection in Flight Data Using Semi-Supervised Explainable Deep Learning Model.," *In AIAA Scitech 2021 Forum*, p. 0774, 2021.

[46] M. T. Ribeiro, S. Singh and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier.," *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144, 2016.

[47] M. R. Zafar and N. M. Khan, "DLIME: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems.," *arXiv preprint arXiv:1906.10263* , 2019.

[48] T. Peltola, "Local interpretable model-agnostic explanations of Bayesian predictive models via Kullback-Leibler projections.," *arXiv preprint arXiv:1810.02678*, 2018.

[49] C. Amrit, T. Paauw, R. Aly and M. Lavric, "Identifying child abuse through text mining and machine learning.," *Expert systems with applications*, vol. 88, pp. 402-418, 2017.

[50] Y. Sari, M. Stevenson and V. Andreas, "Topic or style? exploring the most useful features for authorship attribution.," in *In Proceedings of the 27th International Conference on Computational Linguistics*, 2018.

[51] K. Welbers, W. V. Atteveldt and K. Benoit, "Text analysis in R.," *Communication Methods and Measures*, vol. 11, no. 4, pp. 245-265, 2017.

[52] X. Zhang, E. Green, M. Chen and R. R. Souleyrette, "Identifying secondary crashes using text mining techniques," *Journal of Transportation Safety & Security*, vol. 12, no. 10, pp. 1338-1358, 2020.

[53] T. Takenobu and I. Makoto, "Text categorization based on weighted inverse document frequency.," Tokyo, 1994.

[54] C.-z. Liu, S. Yan-xiu, W. Zhi-qiang and Y. Yong-Quan, "Research of Text Classification Based on Improved TF-IDF Algorithm," *In 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE),* pp. 218-222, 2018.

[55] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management,* vol. 24, no. 5, pp. 513-523, 1988.

[56] M. González, O. Rosado, J. D. Rodríguez, C. Bergmeir, I. Triguero and J. M. Benítez, "ssc: An R Package for Semi-Supervised Classification.".

[57] O. Chapelle, B. Scholkopf and A. Zien, Semi-supervised learning, MIT press Cambridge, 2006.

[58] M. R. Keyvanpour and M. I. Bahojb, "Semi-supervised text categorization: Exploiting unlabeled data using ensemble learning algorithms.," *Intelligent Data Analysis,* vol. 17, no. 3, pp. 367-385, 2013.

[59] S. I. Dimitriadis, D. Liparas, M. N. Tsolaki and Alzheimer's Disease Neuroimaging Initiative., "Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: From the alzheimer's disease neuroimaging initiative (ADNI) data," *Journal of neuroscience methods,* vol. 302, pp. 14-23, 2018.

[60] Z. H. Kilimci and S. Akyokus, "Deep learning-and word embedding-based heterogeneous classifier ensembles for text classification.," *Complexity 2018,* 2018.

[61] R. Bridgelall, "Upper Great Plains Transportataion Institute SMARTSe Resources," [Online]. Available: https://www.ugpti.org/smartse/resources/downloads/support-vector-machines.pdf. [Accessed 02 March 2021].

[62] J. H. Friedman, "Greedy function approximation: a gradient boosting machine.," *Annals of statistics,* pp. 1189-1232, 2001.

[63] T. Samir, J. Granderson and S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings.," *Energy and Buildings,* vol. 158, pp. 1533-1543, 2018.

[64] T. Chen and G. Carlos, "Xgboost: A scalable tree boosting system.," in *n Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* 2016.

[65] C. Yung-Chia, K.-H. Chang and G.-J. Wu, "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions.," *Applied Soft Computing,* vol. 73, pp. 914-920, 2018.

[66] "Very high resolution object-based land use–land cover urban classification using extreme gradient boosting.," *IEEE geoscience and remote sensing letters,* vol. 15, no. 4, pp. 607-611, 2018.

[67] Okfalisa, Mustakim, I. Gazalba and G. I. N. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification.," in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE). IEEE,* 2017.

[68] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification.," *Artificial Intelligence Review,* vol. 52, no. 1, pp. 273-292, 2019.

[69] Wikipedia, "Multiclass classification," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Multiclass_classification. [Accessed 06 03 2021].

[70] C. Davide and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC genomics,* vol. 21, no. 1, pp. 1-13, 2020.

[71] T. Peltola, "Local Interpretable Model-agnostic Explanations of Bayesian Predictive Models," *arXiv preprint arXiv:1810.02678.,* 2018.

[72] C. Molnar, " Local Surrogate (LIME)," in *Interpretable Machine Learning*, Lulu.com, 2020.

[73] X. J. Zhu, "Semi-supervised learning literature survey.," University of Wisconsin-Madison., Madison, 2008.

[74] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural," *IEEE Computational intelligence magazine,* vol. 9, no. 2, pp. 48-57, 2014.

4. RANKING RISK FACTORS IN FINANCIAL LOSSES FROM RAILROAD INCIDENTS: A MACHINE LEARNING APPROACH

4.1. Abstract

The reported financial losses from railroad accidents since 2009 have been more than $4.11 billion dollars. This considerable loss is a major concern for the industry, society, and the government. Hence, identifying and ranking the factors that contribute to financial losses from rail accidents would inform strategies to minimize them. To achieve that goal, this paper evaluates and compares the results of applying different non-parametric statistical and regression methods. The models compared are random forest, K-nearest neighbors, support vector machines, stochastic gradient boosting, extreme gradient boosting, and stepwise linear regression. The results indicate that these methods are all suitable for analyzing non-linear and heterogeneous railroad incident data. However, the extreme gradient boosting method provided the best performance. Hence, the analysis used that model to identify and rank factors that contribute to financial losses, based on the gain percentage of the prediction accuracy. The number of derailed freight cars and the absence of territory signalization dominated as contributing factors in more than 57% and 20% of the accidents, respectively. Partial dependence plots further explore the complex non-linear dependencies of each factor to better visualize and interpret the results.

4.2. Introduction

Every year, railroads invest an average of 40% of their revenue on capital expenditures, maintenance, and condition monitoring [1]. In spite of these investments, the high number of accidents falls far short of the ultimate goal of the Federal Railroad Administration (FRA) to reduce rail-related accidents, injuries, and fatalities to zero [2]. For a decade prior to 2019, nearly

25,000 accidents caused 446 deaths, 5,137 injuries, and more than $4.11 billion in financial loss seasonally adjusted to 2018 dollars [3]. Class I railroads accounted for 78% of those accidents, more than 72% of the resulting injuries and fatalities, and 81% of the total financial loss. Figure 4.1 summarizes the annual class I railroad accidents and the financial losses for the decade prior to 2019.

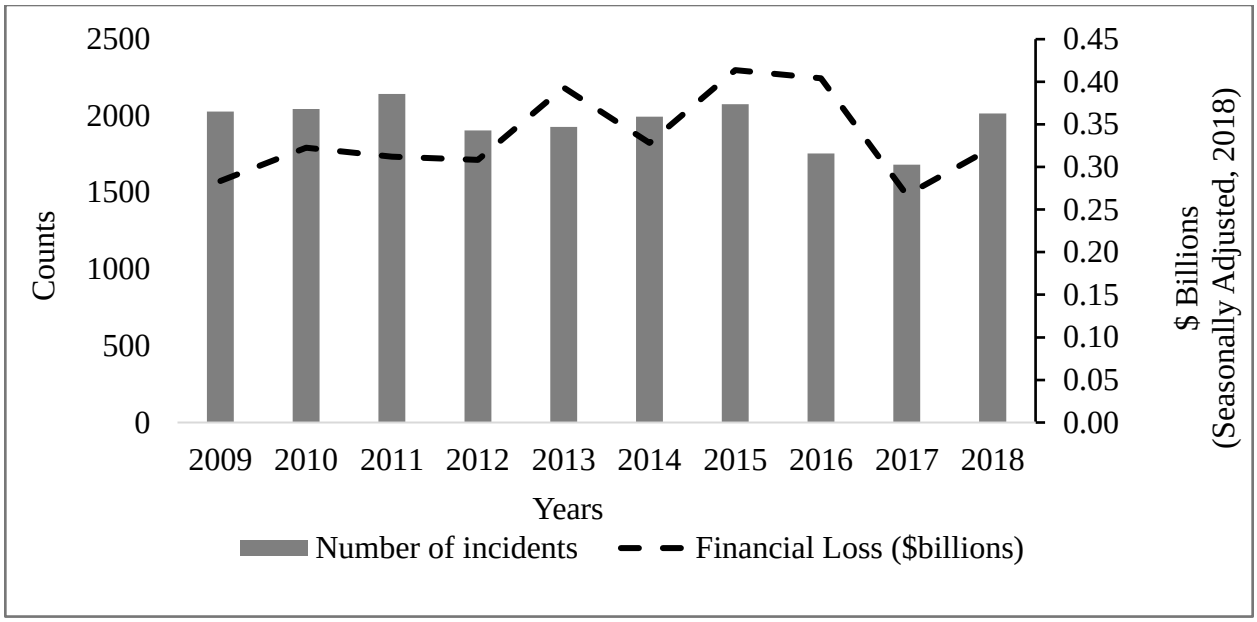

Figure 4.1. Class I railroad incidents from 2009 to 2018 and the reported financial loss

The consistently large number of accidents and the injuries and fatalities they cause place a huge social and economic burden on the industry, environment, and society. Hence, it is vital to understand the dominant accident causes to guide strategies and policies that could minimize financial losses from accidents. Subsequently, the goal of this paper is to apply data mining and machine learning techniques to 15 years of railroad accident data from 2004-2018 to reveal insights about the major contributing factors to financial losses from class I freight train accidents.

The FRA maintains historical data of railway accidents in three primary databases. These datasets are huge and have grown far beyond the ability of humans and commonly used software

tools to capture, manage, and process data within a "tolerable elapsed time" [4]. The available accident data are in non-uniform formats. The data includes heterogeneity, variety, unstructured features, missing values, incorrectly formatted values, and redundancy [5], [6]. Hence, it is not possible to apply standard statistical methods directly to the raw data. Therefore, advanced techniques such as data mining (DM) and machine learning (ML) are necessary to prepare the data for processing.

DM is helpful in analyzing vast amounts of data by using many different techniques to discover useful patterns and relationships among features [7], [8]. Kohavi [9] specified that insight and prediction are the two primary goals of DM. Insights identify patterns and trends that are useful, whereas prediction leads to the identification of a model that provides reliable forecasts based on new input data. Many researchers have applied different DM/ML methodologies to analyze factors that cause accidents on roadways [10], [11], [12], at highway rail-grade crossings (HRGC) [13], [14], and [15], and on railways [16], [17]. For instance, Sohn and Lee [12] compared the results of neural networks, Bayesian fusion, decision tree, bagging, and clustering models on Korean road accident data. Their results indicate that the clustering-based classification works better than the other methods. Depair et al. [11] also examined clustering techniques to identify homogenous accident types. They used vehicle types as the basis for segmentation and evaluated the relationship to injuries caused by different segments.

Some researchers used DM techniques to analyze road-related factors and linked them to accident severity. Beshah and Hill [18] compared different DM models to investigate the role of road-related factors in accident severity in Ethiopia and concluded K-nearest neighbors performed best. Mousa et al. [19] compared ability of tree-based ensemble methods to predict the onset of lane changing maneuvers using connected vehicle data and found that extreme gradient

boosting (XGBM) was the best models. The highest accuracy was 99.7%, and that was better than methods using decision trees, gradient boosting, and (random forest) RF ensemble methods.

Other related areas of research focused on HRGC accidents. Hu et al. [20] evaluated the relationship between crash frequency and the relevant attributes of highway and railroad systems. Ghomi et al. [13] used DM techniques to identify some of the main factors associated with injury severity of road users involved in HRGC accidents. Kang and Khattak [14] investigated the severity of HRGC accidents by clustering the data using a combination of DM and statistical methods. Researchers also use DM/ML techniques to examine railroad accident data. Brown [16] conducted text mining for identifying the contributors to rail accidents. Mirabadi and Sharifian [17] used association rule mining to reveal the relationships and patterns in Iranian Railway accident data. Many other researchers have conducted studies that use other analytical criteria to discover relationships between accident risk and contributing factors [8], [21], [22], and [23].

All research that analyzed rail or road accidents using DM techniques focused on identifying contributing factors that relate to attributes of the respective infrastructure. There is a gap in research to identify and rank risk factors in financial loss from railroad accidents. Subsequently, the main contribution of this research is a comparison of the ability of different non-parametric, tree-based DM methods, and a regression model to identify the risk factors for the financial loss by analyzing 15 years of railroad class I freight train accident data. The authors then use the best predictive model to rank the major factors based on their influence on financial loss. This research extends previous work in railroad safety in the following two ways:

1. It isolates factors that lead to financial losses.
2. It ranks the importance of the major contributors.

The remainder of the paper is structured as follows: Section 2 introduces the models used to identify the factors that influence financial loss. Section 3 describes the data structure, variables, cleaning, and handling. Section 4 compares the model outputs for selection, variable ranking, and the marginal effect of the variables. Section 5 presents the final remarks and describes future work.

4.3. Model Development

The ML methods use tree-based models (random forest, stochastic gradient boosting method, and extreme gradient boosting method), K-nearest neighbor method, and support vector machine to classify the data according to the selected features or factors. In addition, stepwise linear regression provides a baseline for comparison because of its proven effectiveness in previous research [24], [25]. This sections will provide basic descriptions of the six models, all of which are available from the caret package of the R Project for Statistical Computing.

4.3.1. Model Regularization

Model regularization involves trading off training data bias for a reduced variance on new data. This is achievable by partitioning the data appropriately into development and test sets. The former is used for cross-validation while tuning the model, and the latter is used to test the final regularized and tuned model [26]. Running the models with many different variations in partitioning revealed that a 70-30 split between development (training/validation) and testing datasets yielded the lowest variance.

4.3.2. Machine Learning Algorithms

4.3.2.1. K-Nearest Neighbor Method

K-nearest neighbor (KNN) is a supervised learning algorithm that uses non-parametric algorithm that does not require any assumptions on the underlying data distribution. This

algorithm predicts the class of an observation by searching through the entire dataset to identify K other observations that are most similar to it, and then takes the class associated with the majority. The measure of similarity is based on one of several available distance measures [27], [28]. This analysis selects the Euclidean distance because it is the most common.

4.3.2.2. Random Forest

Standard decision trees split the dataset by selecting an attribute and a threshold that maximizes the purity of subtrees. The purity of a node increases as the class imbalance of the dataset within that node increases. However, this tree-splitting strategy results in trees that tend to over-fit the data and subsequently fail to regularize by exhibiting a high variance on new data. Random forest (RF) addresses the regularization issue by introducing two levels of randomness—namely the random selection of learning data and the random selection of decision attributes for tree splitting. Such an adjustment results in better performance than many other classifiers models, and improved robustness against over-fitting [29], [30].

RF learns an ensemble of trees by bootstrapping the same dataset through random sampling with duplication, and then randomly selecting a predetermined number of attributes for subsequent tree splitting [30]. The selected class of observation is the majority vote from all trees created—also referred to as aggregation. Subsequently, the literature often refers to the combined methods of bootstrapping and aggregation as the bagging method. Bagging does not require pruning for regularization because averaging the results of all bootstrapped samples reduces the variance [31].

4.3.2.3. Stochastic Gradient Boosting Method

Stochastic gradient boosting (SGBM) is an extension of the Gradient Boosting (GB) technique. Gradient refers to model building optimization during the learning process. Boosting

refer to finding a more accurate hypothesis by combining the predictions of many weak hypotheses (learners), each of which is moderately accurate [32]. Most of the time, learners are nonlinear models (decision or regression trees), and for such cases, the literature refers to GB as "Gradient Tree Boosting" (GTB). The GTB algorithm builds an ensemble of weak prediction models by adding a sequence of trees, with successive trees grown on reweighted versions of the data. At each stage, GTB generates a new tree from the residual and adds to the existing group of trees. The algorithm builds the final ensemble with a weighted summation of the individual learners.

Motivated by Breiman's bagging phenomenon, Friedman [33], [34] augmented the gradient boosting procedure and incorporated randomness as part of the GB algorithm and called that phenomenon SGBM. Friedman recommended that instead of using the entire dataset to perform the boosting, it is more appropriate to select a random subsample from the training dataset at each step of the boosting process. The base learner then uses this randomly selected subsample.

4.3.2.4. Extreme Gradient Boosting Method

Extreme Gradient Boosting Method (XGBM) follows the gradient boosting method, but it is more efficient and accurate. Unlike the GB technique, the XGBM implements an additional regularization to avoid over-fitting by imposing additional control over model complexity [19]. The additional regularization term does not depend on the randomness. Instead, the focus of this additional term always remains on minimizing the model complexities based on some leaves and the sum-of-square scores of those leaves. For further reference, [35] presents a detailed study on XGBM.

4.3.2.5. Support Vector Machine

A Support Vector Machine (SVM) is a non-parametric statistical learning technique that requires no assumption on the underlying data distribution. The concept is to separate data across a decision boundary (planes) determined by a small subset of the data (feature vectors). The data subset that supports the decision boundary is called the support vector [36]. SVM assumes that the multi-feature data are linearly separable in the input space. However, in practice, data points of different planes overlap, which makes linear separability challenging [37]. A "kernel trick" overcomes the problem of the linearity restriction on the decision boundary. The kernel trick uses a transformation function to map the input vector into a higher dimension space by introducing new parameters [36]. The "trick" part is that the SVM operates only on the vectors in their ambient space, without actually transforming the vector into a higher dimension. This analysis uses the radial kernel. Various authors [36], [35], and [38] explain the use of the kernel trick in more detail.

4.3.2.6. Stepwise Linear Regression

Stepwise Linear Regression (SLR) is the process of building a model by successively removing or adding feature variables based on their relationship with the response variable. In other words, SLR is a method of regressing multiple variables in multiple stages. In each stage, the method removes or adds variables based on their correlation with the response variable.

4.3.3. Model Comparison

To minimize the potential for over-fitting or under-fitting, the machine learning procedure incorporates a K-fold cross-validation process with N repeats to identify the best model parameters. As explained by Jhangiri and Rakha [38], the K-fold algorithm segments the training data randomly into *K* parts or folds of approximately equal size. Subsequently, the

algorithm builds a model from the union of the remaining K-1 folds and evaluates the model performance on the validation fold. The algorithm repeats the cross validation K times so that each fold serves as the validation data exactly once. The algorithm repeats the K-fold process *N* times to introduce further randomization. The algorithm builds the final model by using those parameters that produce the best average performance across the *K* validations.

The K-fold cross-validation algorithm sets a uniform random seed before training each model to ensure consistency in the data partitions and repeats. Once trained, the process adds all the models to a list for re-sampling. This function verifies that the models are comparable and have used the same training scheme [39], [40]. Finally, the algorithm evaluates the performance of the models by comparing the Mean Absolute Error (MAE), the Root Means Squared Error (RMSE), and the R-squared metrics. The MAE is the unweighted average of the absolute differences between the predicted and actual observations. RMSE is the square root of the average of the squared differences between the predicted and actual observations. Hence, RMSE represents the average magnitude of the error. R-squared is a measure of the percentage of the variation in the response variable that the model explains.

4.4. Data

FRA requires that railroads maintain and submit a detailed report of all significant accidents or incidents associated with railroad operations. FRA compiles these reports into the Railway Equipment Accident (REA) database [94][94]. This study used 15 years of REA accident data from all railroads reporting all types of accidents between 2004 and 2018 [3]. This database records all accidents that exceed a specified financial cost (inflation-adjusted 2019 threshold - $10,700) from damages to on-track equipment, signals, track, track structures, and roadbed [41]. However, study uses class I freight train accident data for the analysis. The data

consists of more than 145 variables, such as the railroad identifier, accident location, speed, and other attributes that attempt to describe the nature of the event. A possible limitation of this database is that it may not have captured all the underlying factors that contributed to the level of financial loss. However, the models are based only on the available factors, and is likely to expose factors that are dominant in causing financial loss.

4.4.1. Cleaning and Structuring

The data cleaning followed a three-step process. The first step deleted variables that were not appropriate, such as text narratives, dummy variables, and duplicate variables. The second stage removed variables such as "number of engineers" and "location," that did not support the analysis objectives. The third stage modified some of the FRA-structured default variables as follows:

1. TIMEHR– changed the specific hours and minutes of the incident from the standard 12-hour, a.m.-p.m." format to a single variable in 24-hour military time format.
2. P_CARSDMG – a new variable that is equal to the percentage of cars carrying hazmat that were damaged or derailed.
3. TRKCLAS – changed the FRA track classes of A-E to a numeric categorical variable for compatibility across the data mining techniques used.
4. TRKDNSTY – imputed missing values and replaced zero values based on the maximum reported for that county.
5. Ospeed – restructured "train speed" as a categorical variable over speed where the value is '1' if the train was traveling faster than the track class limit, and '0' otherwise.

6. P_LocoDe – a new variable that contains the percentage of locomotives derailed is estimated using the same dataset.
7. Tloco – a new variable that contains the total number of locomotives is obtained using the same dataset.
8. Cause – changed the primary cause of an accident to a categorical variable with five classes based on their alphabetic order. '1' = 'Mechanical and Electrical Failures'; '2' = 'Miscellaneous Causes Not Otherwise Listed'; '3'= Rack, Roadbed and Structures; '4'=Signal and Communication; and '5'= Train operation - Human Factors.
9. EQATT – '1' if someone was attending the equipment and '0' otherwise.
10. R_ Amount – a modified dependent variable containing the total reported financial damage. The modifications are as follows.
    a. Time value normalization: adjusted the total reportable damage from the variable ACCDMG to the average consumer price index seasonally adjusted amount of 2018.
    b. The REA databases should include only those accidents that exceed financial losses of $5,000. Therefore, this adjustment deleted records with lower amounts because they would be outliers and not representative of the majority of accidents that occurred.
    c. Analysis of the distribution of the reported financial losses revealed that those beyond the 95-percentile are outliers and are eliminated.

4.4.2. Handling Correlation and Missing Values

Missing values do not cause a problem for Decision Tree (DT) models because the method imputes those values based on the values of other observations that are in similar classes.

However, models such as Linear Regression (LR) cannot use data that contain missing values, thereby making the size of the dataset inconsistent for uniform comparison of models [26]. Model comparison is most appropriate between models that are fitted using the same set of observations [26]. Hence, it is necessary to impute missing values before fitting models for comparison of performance. This analysis replaced missing values using an approach based on k-nearest neighbors, referred to as KnnImputation. For each missing value, the model identifies 'k' (k=10) closest observations based on the Euclidean distance and computes the weighted average as the missing value.

Highly correlated variables with the dependent variable are truly redundant and does not contribute additional information in the model [42]. Therefore, the procedure removed those variables that had a correlation coefficient above a commonly selected threshold of 0.75 [43].

4.4.2.1. Dataset for Model Comparison

The final dataset contained 23 variables (Table 4.1) and approximately 12,500 observations of the freight train accidents of the class I railroads.

4.5. Results and Discussion

4.5.1. Model Selection

Table 4.2 summarizes the evaluation metrics for the six machine learning models, using 10-Fold cross-validations with 3 repeats. In general, the ensemble tree-based models outperformed the other models. Among tree-based ensemble methods, XGBM provided the best predictive capability based on the lowest RMSE and MAE metrics, and the highest R-squared metric.

4.5.2. Variable Importance Using XGBM

After identifying XGBM as the best model for the data, the analysis focused on identifying the significant contributors to the accident. Table 4.3 summarizes the results.

Table 4.1. List of variables and their description

Variable	Description	Variable Type
R_Amount	Seasonally adjusted financial loss based on 2018 prices (dependent variable)	Continuous
MONTH	Month of incident	Categorical
DAY	Day of the incident	Categorical
TIME	Time of the accident (military standard time)	Continuous
TYPE	Type of accident: (1-13)	Categorical
P_CARSDMG	% of hazmat cars damaged or derailed	Continuous
TEMP	Temperature in degrees Fahrenheit	Continuous
VISIBLTY	Daylight period: (1-4)	Categorical
WEATHER	Weather conditions (1-6)	Categorical
Ospeed	Boolean of train traveling over the speed limit	Categorical
TONS	Gross tonnage, excluding power units	Continuous
EQATT	Boolean for equipment attended by a human	Categorical
TRKCLAS	FRA track class (0-9)	Categorical
TRKDNSTY	Annual track density - gross tonnage in millions	Continuous
POSITON1	Car position in train (first involved)	Categorical
POSITON2	Car position in train (causing)	Categorical
Tloco	Total number of locomotives	Categorical
P_LocoDe	Percent of locomotives derailed	Continuous
LOADF2	Number of derailed loaded freight cars	Categorical
EMPTYF2	Number of derailed empty freight cars	Categorical
CAUSE	Primary cause of incident	Categorical
TOTKLD	Total killed for the railroad as reported	Categorical
SIGNAL	Type of territory – signalization	Categorical

Table 4.2. Model comparison evaluation

Models	Label	RMSE	MAE	R^2
GBM	Gradient boosting model	87,131.87	139,295.8	0.4596048
KNN	K-nearest neighbors	122,391.3	189,069.6	0.0335639
SVM	Support vector machine	102,771.3	204,053.4	0.0476689
RF	Random forests	88,939.76	143,402.5	0.4534554
STEPWISE	Stepwise regression	95,392.14	149,052	0.3979852
XGBM	Extreme gradient boosting	85,989.11	137,646.9	0.4618731

The results indicate that the number of loaded freight cars derailed is the most crucial factor in financial damages from accidents by a proportional contribution of 57%. Territory signalization (SIGNAL) is the second-best predictor contributing over 20%. The number of empty freight cars derailed is next, which improves the predictability by more than 10%. Accidents on track class 4 are the next factor associated with more than 4% of the total damage. Table 4.3 summarizes the rank of the other variables.

Table 4.3. Results of variable importance

Feature	Description	Gain	Frequency	Cover
LOADF2	# of derailed loaded freight cars	0.57459	0.2900	0.51493
SIGNAL1	Type of territory – signalization (mandatory)	0.20220	0.1700	0.05337
EMPTYF2	# of derailed empty freight cars	0.10124	0.1366	0.26967
TRKCLAS4	FRA track class: 1-9	0.06536	0.1726	0.02682
TONS	gross tonnage, excluding power units	0.02092	0.0757	0.06610
TRKCLAS3	FRA track class: 1-9	0.01018	0.0460	0.00454
TRKCLAS2	FRA track class: 1-9	0.01008	0.0320	0.00698
TYPE3	type of accident: 03=Rear-end collision	0.00599	0.0197	0.02691
P_LocoDe	% of locomotive derailed	0.00370	0.0263	0.01545
CAUSE	Contributing cause of incident	0.00233	0.0091	0.00182
POSITON1	Car position in train (first involved)	0.00194	0.0089	0.01257
POSITON2	Car position in train (causing)	0.00069	0.0043	0.00037
TRKDNSTY	Annual track density-gross tonnage in millions	0.00062	0.0063	0.00033
MONTH12	month of incident	0.00014	0.0023	0.00012
Tloco	Total number of locomotive	0.00001	0.0003	0.00001

4.5.3. Marginal Effect of Predictor Variables

Advanced machine learning models can significantly improve predictions and classifications, but understanding the impact of one or more predictor variables on the response variable is not feasible even with these advanced models. Partial-dependence plots (PDPs) can show the marginal effect of a single contributor on the predicted outcome of a machine learning model [45]. The PDPs show the distinct impact of the most influential variables after controlling for the average effects of all other variables in the model [46]. Therefore, this study uses PDPs to explore and visualize the complex non-linear global relationship between each factor and the predicted outcome.

Figure 4.2 shows that, except for the effects of the binary signal variable, the PDPs from the XGBM model exhibits non-linear patterns. The yhat ($\hat{y}$) variable represents the predicted financial loss from a XGBM regression of each variable.

Per results, Financial damage generally increased with the number of derailed cars (LOADF2) and peaked at 40. Non-signaled territories (SIGNAL = 2) are associated with higher financial losses than with territories that are signaled. The partial dependency on EMPTYF2 suggests that financial damage could be most severe if 30 to 40 empty cars derail. Financial loss generally increases with track classification, and peaks for class 8 tracks. Trains that carry approximately 20,000 tons tend to have the most significant impact on the financial damage. Head-on collision (TYPE2) and Rear-end collision (TYPE3) are associated with higher financial losses than other accident types. Accident causes (CAUSE) category 5 (human factor related) are associated with the highest financial losses. P_LocoDe (percentage of locomotive derailed) exhibits a stepwise increasing trend in financial losses. POSITON1 (Car position in train (first

involved) and POSITON2 (causing car position in the train) from 125 to 135 are associated with the highest financial losses. These cars tend to be towards the rear of a typical Class I train [47].

By month (IMO), financial losses peak in the summer and subside in the winter. Grain harvesting and grain shipping generally peak in the summer when track conditions are favorable. Intuitively, peak demand leads to peak traffic with higher carloads and a resulting higher risk of accidents that contribute to losses. T_loco shows that accident damage increases with trains containing more than five locomotives. Weather shows that the financial impact of an accident increased by more than 0.03% while having snow compared to other weather conditions.

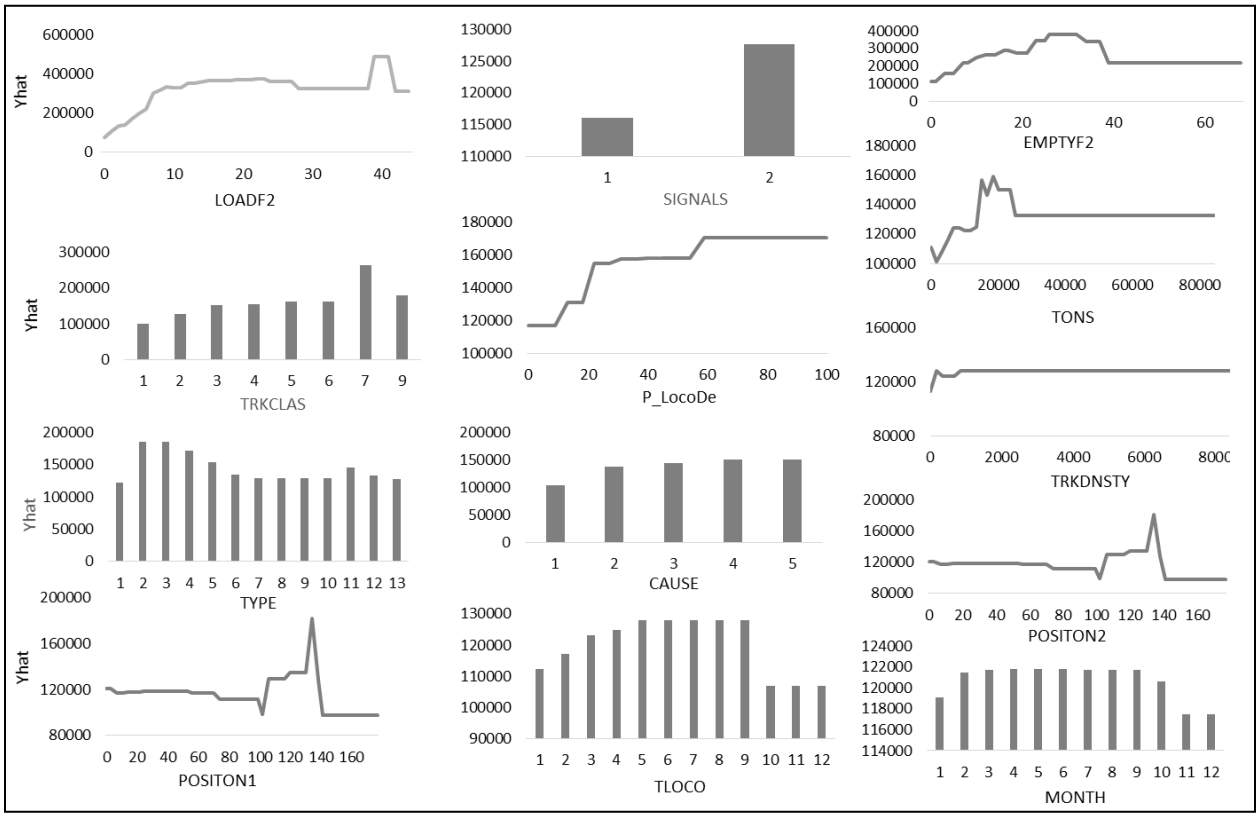

Figure 4.2. Partial dependence plots of the predictor variables in the model

4.6. Conclusion

The primary objective of this study was to determine the significant contributing factors to Class I railroad financial losses from railway accidents and to rank those factors by using DM

and ML techniques. Data between 2004 and 2018 from the railroad equipment accident (REA) database provided inputs for the analysis. To achieve the primary objective of the study, a comparative analysis of six machine-learning algorithms determined the best model for the dataset. Finally, the best performing model showed that tree-based ensemble models performed best. Particularly, XGBM proved to be the best model for analyzing railroad accident data that is highly imbalanced. The XGBM model identified the significant contributors to railroad accidents. The results indicate that LOADF2 (number of derailed loaded freight cars), SIGNAL (Type of territory – signalization), and EMPTYF2 (number of derailed empty freight cars) are the top three significant factors that account for financial loss severity with the gains of 57.46%, 20.22%, and 10.12% respectively. These results demonstrate the effectiveness of applying DM and ML techniques to high volume and non-uniform data formats. The results suggest that railroads should prioritize safety investments for situations where a greater number of trains carry freight and for infrastructure that implements signals.

Future work will explore and evaluate additional exogenous contributors to rail accidents using a similar approach. The results of these two studies will provide an opportunity to conduct a comprehensive assessment of rail accident contributors.

4.7. Reference

[1] AAR, "Railroad 101- Freight Railroads Fact Sheet," 2020. [Online]. Available: https://www.aar.org/wp-content/uploads/2020/08/AAR-Railroad-101-Freight-Railroads-Fact-Sheet.pdf. [Accessed 23 May 2021].

[2] FRA, "Monetary Threshold Notice," Federal Railroad Administration, [Online]. Available: https://railroads.dot.gov/forms-guides-publications/guides/monetary-threshold-notice. [Accessed 15th August 2020].

[3] FRA, "Accident Data as reported by Railroads," Federal Railroad Administration, [Online]. Available: https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/on_the_fly_download.aspx. [Accessed 17 March 2018].

[4] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering,* vol. 26, no. 1, pp. 97-107, 2013.

[5] Y.-S. Chung, "Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees," *Accident Analysis & Prevention,* vol. 61, pp. 107-118, 61.

[6] M. Chen, S. Mao and Y. Liu, "Big data: A survey," *Mobile networks and applications,* vol. 19, no. 2, pp. 171-209, 2014.

[7] L. Liling, S. Shrestha and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques," in *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, IEEE, 2017.

[8] J. Abellán, G. López and J. D. OñA, "Analysis of traffic accident severity using decision rules via decision trees," *Expert Systems with Applications,* vol. 40, no. 15, pp. 6047-6054, 2013.

[9] R. Kohavi, "Data mining and visualization," in *In Sixth Annual Symposium on Frontiers of Engineering*, 2001.

[10] S. K. Barai, "Data mining applications in transportation engineering," *Transport,* vol. 18, no. 5, pp. 216-223, 2003.

[11] B. Depaire, G. Wets and K. Vanhoof, "Traffic accident segmentation by means of latent class clustering," *Accident Analysis & Prevention,* vol. 40, no. 4, pp. 1257-1266, 2008.

[12] S. Y. Sohn and S. H. Lee, "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea," *Safety Science,* vol. 41, no. 1, pp. 1-14, 2003.

[13] H. Ghomi, M. Bagheri, L. Fu and L. F. Miranda-Moreno, "Analyzing injury severity factors at highway railway grade crossing accidents involving vulnerable road users: A comparative study," *Traffic injury prevention,* vol. 17, no. 8, pp. 833-841, 2016.

[14] Y. Kang and A. Khattak, "Cluster-based approach to analyzing crash injury severity at highway–rail grade crossings," *Transportation research record,* vol. 2608, no. 1, pp. 58-69, 2017.

[15] P. Lu and D. Tolliver, "Accident prediction model for public highway-rail grade crossings," *Accident Analysis & Prevention,* vol. 90, pp. 73-81, 2016.

[16] D. E. Brown, "Text mining the contributors to rail accidents," *IEEE Transactions on Intelligent Transportation Systems,* vol. 17, no. 2, pp. 346-355, 2015.

[17] A. Mirabadi and S. Sharifian, "Application of association rules in Iranian Railways (RAI) accident data analysis.," *Safety Science,* vol. 48, no. 10, pp. 1427-1435, 2010.

[18] T. Beshah and S. Hill, "Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia.," *In 2010 AAAI Spring symposium series.,* 2010.

[19] S. R. Mousa, P. R. Bakhit, O. A. Osama and S. Ishak, "A comparative analysis of tree-based ensemble methods for detecting imminent lane change maneuvers in connected vehicle environments," *Transportation Research Record,* vol. 2672, no. 42, pp. 268-279, 2018.

[20] S.-R. Hu, C.-S. Li and C.-K. Lee, "Model crash frequency at highway–railroad grade crossings using negative binomial regression," *Journal of the Chinese Institute of Engineers ,* vol. 35, no. 7, pp. 841-852, 2012.

[21] X. Liu, "Statistical causal analysis of freight-train derailments in the United States," *Journal of transportation engineering, Part A: Systems,* vol. 143, no. 2, p. 04016007, 2017.

[22] X. Liu, M. R. Saat, X. Qin and C. P. Barkan, "Analysis of US freight-train derailment severity using zero-truncated negative binomial regression and quantile regression," *Accident Analysis & Prevention ,* vol. 59, pp. 87-93, 2013.

[23] L. Xiang, M. R. Saat and C. P. Barkan, "Analysis of causes of major train derailment and their effect on accident rates," *Transportation Research Record,* vol. 2289, no. 1, pp. 154-163, 2012.

[24] M. M. Rahman, N. Haq and R. M. Rahman, "Machine learning facilitated rice prediction in Bangladesh.," in *In 2014 Annual Global Online Conference on Information and Computer Technology*, IEEE, 2014.

[25] E. G. Ross, N. H. Shah, R. L. Dalman, K. T. Nead, J. P. Cooke and N. J. Leeper, "The use of machine learning for the identification of peripheral artery disease and future mortality risk.," *Journal of vascular surgery,* vol. 64, no. 5, pp. 1515-1522, 2016.

[26] SAS Institute Inc., "Getting Started with SAS® Text Miner 12.1," 2012. [Online]. Available: https://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf. [Accessed 18 August 2017].

[27] N. García-Pedrajas, J. A. Romero Del Castillo and G. Cerruela-García, "A proposal for local k values for K-nearest neighbor rule," *IEEE transactions on neural networks and learning systems,* vol. 28, no. 2, pp. 470-475, 2015.

[28] Y. Xie, Y. Wang, A. Nallanathan and L. Wang, "An improved K-nearest-neighbor indoor localization method based on spearman distance.," *IEEE signal processing letters,* vol. 23, no. 3, pp. 351-355, 2016.

[29] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[30] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News,* vol. 2, no. 3, pp. 18-22, 2002.

[31] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, no. 2, pp. 123-140, 1996.

[32] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions.," *Machine learning,* vol. 37, no. 3, pp. 297-336, 37.

[33] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics,* pp. 1189-1232, 2001.

[34] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis,* vol. 38, no. 4, pp. 367-378, 2002.

[35] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho and K. Chen, "Xgboost: extreme gradient boosting," *R package version 0.4-2,* vol. 1, no. 4, pp. 1-4, 2015.

[36] R. Bridgelall, "Lecture Notes: Introduction to Support Vector Machines," 02 September 2017. [Online]. Available: https://docslib.org/doc/8549921/introduction-to-support-vector-machines. [Accessed 05 June 2018].

[37] G. Mountrakis, J. lm and C. Ogole, "Support vector machines in remote sensing: A review.," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 66, no. 3, pp. 247-259, 2011.

[38] A. Jahangiri and H. A. Rakha, "Applying machine learning techniques to transportation mode recognition using mobile phone sensor data," *IEEE transactions on intelligent transportation systems,* vol. 16, no. 5, pp. 2406-2417, 2015.

[39] J. Yoonsuh and J. Hu, "AK-fold averaging cross-validation procedure," *Journal of nonparametric statistics,* vol. 27, no. 2, pp. 167-179, 2015.

[40] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition,* vol. 48, no. 9, p. 48, 2839-2846.

[41] X. Liu, R. M. Saat and C. PL Barkan, "Freight-train derailment rates for railroad safety and risk analysis," *Accident Analysis & Prevention,* vol. 98, pp. 1-9, 2017.

[42] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research,* vol. 3, no. Mar, pp. 1157-1182, 2003.

[43] S. D. McCauliff, J. M. Jenkins, J. Catanzarite, C. J. Burke, J. L. Coughlin, J. D. Twicken, P. Tenenbaum, S. Seader, J. Li and M. Cote, "Automatic classification of Kepler planetary transit candidates," *The Astrophysical Journal ,* vol. 806, no. 1, p. 6, 2015.

[44] T. Chen, T. He, M. Benesty and Y. Tang, "Understand your dataset with XGBoost," R-Studio, [Online]. Available: https://cran.r-

project.org/web/packages/xgboost/vignettes/discoverYourData.html#understand-your-dataset-with-xgboost. [Accessed 18 May 2018].

[45] B. M. Greenwell, "pdp: An R package for constructing partial dependence plots," *The R Journal*, vol. 9, no. 1, pp. 421-436, 2017.

[46] Y.-S. Chung, "Factor complexity of accident occurrence: An empirical demonstration using boosted regression trees.," in *3rd International Conference on Road Safety and Simulation.*, 2011.

[47] Cambridge Systematics, Inc., "National Rail Freight Infrastructure Capacity and Investment Study," September 2007. [Online]. Available: http://www.coaltrainfacts.org/docs/natl_freight_capacity_study.pdf. [Accessed 05 May 2018].

5. SUMMARY

The 49 Code of Federal Regulation (CFR) §213.233 includes guidelines for maintaining each railroad track class type and frequency of inspections. These standards include continuous inspections as frequent as twice weekly for most track classes, which requires efficient allocation and utilization of resources. In addition, increasing rail traffic exacerbates track conditions more rapidly and makes continuous monitoring more elusive. Hence, to comply with FRA's track safety standards and maintain safe traffic conditions, the railroad industry uses various NDE technologies on most tracks in operation each year. These NDE technologies include ultrasonic testing, magnetic particle testing, eddy current testing, microwave, millimeter-wave testing, radiography testing, penetrant testing, acoustic emission testing, and thermography which are explained in detail by ISU-Centers for Nondestructive Evaluations [1]. NDE technologies provide additional support to the visual inspection techniques for identifying and locating track abnormalities [2]. However, the size and cost of these technologies currently limit their deployment to specially constructed automated inspection vehicles that locate internal rail flaws and irregular track geometry, track modulus, and gauge restraint [3]. Moreover, NDE technologies often fail to detect defects, which increases the safety risk of workers and may lead to catastrophic failures and requires track closure [4].

Many railroad companies, along with the NDE technologies, have adopted machine-vision inspection technologies to detect anomalies in the track structure [5]. Such technologies use optical sensors, which require light emitters to illuminate the surface, and image sensors to capture the reflected light. Light Detection and Ranging (LIDAR) is an optical laser measurement technique that uses ultraviolet and near-infrared light to generate high-resolution images for detecting track surface anomalies. To calculate distances and track geometric

parameters, the sensor emits a narrow beam of laser light towards the desired object and measures the time taken by the pulse to reflect off the target and return to the device [6]. Automated algorithms filter and extract the required information from high-resolution 2D or 3D images to estimate modulus by measuring the amount of rail displacement from a tangential horizontal plane above the wheel contact point. The laser scanning action produces transversal surface resolution at a relatively high speed but decreases the longitudinal resolution with increasing train speed which is a significant shortcoming. Manufacturers try to overcome this limitation by increasing the inspection speed beyond 40 mph, which requires extensive and expensive construction. Other shortcomings include the physical limitations in signal bandwidth, signal-to-noise ratio, power consumption, sample rate, potentially higher false positives, and slow processing speed provide a diminishing return on the investment of such systems.

Another famous technique machine vision which uses an optical sensor with image capturing to identify and characterize surface abnormalities. The technique uses high-quality cameras to create a stereoscopic vision or 3D images for detailed analysis. Some systems combine an infrared filter with the high-quality image capturing to detect cold wheels, hot wheels, and hot journals that may cause problems. The main advantages of using machine vision-based inspection include greater objectivity and consistency than manual vision. However, there are numerous limitations of this system. For instance, this technique requires a large storage capacity, ample light source with a sun shield, and computationally intensive image processing. Xenon lights or lasers can improve some lightning-related problems but add cost, bulk, and power consumption. All these methods are useful for assessing irregular track geometry but they require expensive sensors, cameras, and other infrastructure, thus resulting in high maintenance cost, greater energy requirements, and low robustness.

Railroads are moving towards real-time condition monitoring that provides in-service measurements of the railway components because they can detect faults while in revenue service operation [6]. The practical application of condition monitoring of the train dynamics is made either by employing track-based or vehicle-based sensors. The track-based sensors focus on monitoring the local area infrastructure and passing train characteristics, whereas rolling stock-based sensors continuously monitor the traveled infrastructure and vehicle for defects. Both these methods have their advantages and potential shortcomings, but the rolling stock-based sensors are more commonly used for tracking surface condition monitoring for their effectiveness. Modern rolling stock is fitted with high-capacity communication buses, inertial sensors, gyroscopes, and accelerometers. It also incorporates global navigation satellite systems, such as Global Positioning System (GPS), to record and identify the locations of the detected track anomalies. Moreover, these technologies require advanced processing units for collecting, processing, filtering, and managing signal data [7].

Implementing low-cost sensors and GPS receivers on rolling stock poses significant challenges and limitations in detecting signals at high accuracy and high precision. First, the non-uniform sample rate of an accelerometer causes a problem in signal detection, which reduces the signal-to-noise ratio and increases the possibility of false-positive and false-negative results. Second, GPS receivers do not always provide reliable and accurate position information because of the following five reasons [2].

1. Standard low-cost GPS receivers provide position updates approximately once per second. Given an inertial sensor that samples at 64 hertz on average, the GPS coordinate will update after every group of 64 inertial samples. Therefore, the system

will, on average, tag blocks of 64 inertial samples with the same GPS coordinates. This causes low position resolution and consequently signal misalignment.

2. Position updates in a traversal are spatially asynchronous. That means the GPS updates from an individual traversal will be at different geospatial points along the path. Therefore, some position updates will not tag some signal peaks.
3. The geospatial position error is in two dimensions, causing deviation in position updates from the travel path.
4. The standard deviation of the position estimates from GPS receivers is three to five meters along the travel direction.
5. The non-Line-Of-Sight (non-LOS) condition triggered by clouds, trees, or tunnels could block the reception of GPS signals in some locations, causing non-uniform update rates.

RAILS has the potential to overcome the problems listed above. The technique uses distance interpolation, heuristic, and correlation alignment to align the signals from multiple traversals. The advanced ensemble averaging aligned the signal data from multiple traversals to increase the rate of anomaly detection while minimizing the false positive and false negative. Hence, the system becomes more reliable, scalable, and effective. sections I, II, and III already discussed the other benefits, functionalities, and practices of RAILS.

This study provides and highlights additional features of RAILS by developing a new method to categorize the accidents based on responsible factors and prioritize the accidents based on the potential for causing financial damages. Categorizing such incidents would help to create a database that prioritizes issues and suggest possible countermeasures based on the problems. This study has been conducted in three phases, focusing on comparing and developing new

supportive methods that could help the proposed technology achieve its objective. The objective of the first phase was to develop a probabilistic method for comparing the performance of RAILS and its advantages over current NDE methods. For the analysis, the study used data from multiple scans of a track segment and applied probabilistic models based on the theory of operations. The findings suggest that the proposed methodology has 165% better chances of detecting TRS-related faults with only two trains passing per day, for the scenario of a first-pass probability detection of only 20%. One of the benefits of RAILS is to classify the defects based on the potential symptoms so that specialists can focus inspections on critical areas without closing the lines for a longer period. The second phase of this research would help in achieving this objective. The second phase aims to develop a methodology for classifying and categorizing accidents, determining the potential reasons behind causing these events, and identifying additional possible threats associated with these accidents causing factors. The proposed methodology will help develop a database based on different factors responsible for causing accidents and highlight other associated threats with the key areas needing focus. The study applied and compared many different non-parametric models to achieve the study's objective. The result suggested that studying the accident-causing factors at the micro-level is significant in identifying category-based factors responsible for causing accidents, which further helps develop core-level countermeasures.

Phase I and Phase II of this research would help connect PIEs disturbance with the accident-causing factors to create a database. This will help achieve the main objective of classifying the potential defects based on the signal variations and identifying some additional threats at an early stage that otherwise could have led to another issue in the future. In short, this phase will help the proposed technology develop a methodology to change the system from find

and fix" to "predict and prevent." The third phase of this research focused on identifying and categorizing the major contributing factors to financial losses from freight train accidents. Ranking the factors based on financial damage severity would help agencies prioritize the countermeasures that have the potential for causing severe damages and have railroads allocate budgets better, provide efficient use of resources, and prioritize safety investments.

Each section above discussed the limitation and challenges related to individual phases of this research. However, the research has overall limitations related to nonparametric machine learning algorithms. For instance, (i) the data in this study is limited to estimate the mapping function for some of the non-parametric models. In future, adding more data could provide better results. (ii) The models had far more parameters to train making the algorithms slower. (iii) Although cross validation helped to avoid the problem of overfitting, nonparametric machine learning algorithms always carry the overfitting risks, and it is even harder to analyze the reason behind specific predictions. (iv) The lack of uniformity in the qualitative and quantitative methods affected the accuracy of the results. All these shortcomings can be overcome in the future by incorporating more data and using advanced methodologies like deep learning.

In summary, in a short experiment, RAILS technique can provide anomaly detection with higher accuracy, precision, and the consistency while minimizing the false positive and false negative. Subsequently, this technology can leverage the PTC network to communicate track and roadbed problems and reduce derailment risks. The benefit cost analysis study conducted by Bridgelall *et al*. [8] showed that RAILS can add additional benefits to the current PTC system and thus increase the ROI from the overall PTC investment. FRA and railroads can use this benefit-cost analysis to help analyze the tradeoff between technology costs, their potential benefits in accident prevention, and the payback period with different discount rate scenarios [8]

5.1. Reference

[1] ISU-Center for Nondestructive Evaluation, "Nondestructive Evaluation Techniques," IOWA State University, 2021. [Online]. Available: https://www.nde-ed.org/NDETechniques/index.xhtml. [Accessed 31 May 2022].

[2] P. Lu, R. Bridgelall, D. Tolliver, B. Bhardwaj and N. Dhingra, "Track Surface Irregularity Position Localization With Smartphone-Based Solution. No. MPC-551," Mountain-Plains Consortium, Fargo, 2021.

[3] P. Lu, R. Bridgelall, D. Tolliver, L. Chia and B. Bhardwaj, "Intelligent Transportation Systems Approach to Railroad Infrastructure Performance Evaluation: Track Surface Abnormality Identification with Smartphone-Based App," Mountain-Plains Consortium-MPC 19-384, Frago, 2019.

[4] M. Hong, Q. Wang, Z. Su and L. Cheng, "In situ health monitoring for bogie systems of CRH380 train on Beijing–Shanghai high-speed railway," *Mechanical Systems and Signal Processing*, vol. 45, no. 2, pp. 378-395, 2014.

[5] L. Al-Nazer, T. Raslear, C. Patrick, J. Gertler, J. Choros, J. Gordon and B. Marquis, "Track Inspection Time Study (No. DOT/FRA/ORD-11/15)," FRA- Office of Railroad, Washington, 2011.

[6] A. Falamarzi, S. Moridpour and M. Nazem, "A Review on Existing Sensors and Devices for Inspecting Railway Infrastructure," *Jurnal Kejuruteraan,* vol. 31, no. 1, pp. 1-10, 2019.

[7] R. W. Ngigi, C. Pislaru, A. Ball and F. Gi, "Modern techniques for condition monitoring of railway vehicle dynamics," *Journal of Physics: Conference Series,* vol. 364, no. 1, p. 012016, 2012.

[8] R. Bridgelall, P. Lu, D. D. Tolliver, N. Dhingra and B. Bhardwaj, "Benefit Cost Analysis of Railroad Track Monitoring Using Sensors Onboard Revenue Service Trains," Mountain-Plains Consortium MPC 21-446, Fargo, 2021.

APPENDIX A. CAUSE I LIME RESULTS AND WORD ASSOCIATION FOR EACH SUB-CATEGORY

Lime Results	word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)
Correlation for C110						
journal	burn (0.46)	overheat (0.27)	failure (0.20)			
burn	journal (0.46)	mile (0.26)	side (0.26)	drop (0.26)	remain (0.26)	
roller	bear (0.55)	overheat (0.32)	failure (0.27)	defective (0.21)		
detector	equipment (0.45)	stop (0.42)	drag (0.39)	hot (0.31)	box (0.29)	defect (0.26)
hot	box (0.77)	detector (0.31)	receive (0.25)	alert (0.22)	defect (0.22)	
Correlation for C111						
Sill	break (0.48)	old (0.38)	center (0.35)	apart (0.26)		
plate	rigid (0.59)	center (0.42)	movement (0.42)	equipment (0.37)		
draft	head (0.31)	fall (0.31)	break (0.21)			
old	break (0.43)	force (0.41)	roll (0.41)	sill (0.38)	action (0.33)	separate (0.27)
bar	draw (1.00)	pin (061)	air (0.55)	condition (0.46)	hose (0.46)	fail (0.40)
Correlation for C112						
hose	air (0.78)	Separation (0.56)	uncouple (0.29)			
valve	malfunction (0.81)	tread (0.37)	stick (032)	buildup (0.23)	brake (0.22)	
Slack	action (0.65)	adjuster (0.42)	airbrake (0.34)			
rod	bleed (0.88)	gap (0.83)	switch (0.77)	point (0.77)	bend (0.88)	
bleed	gap (0.94)	switch (0.88)	rod (0.88)	point (0.82)	emergency (0.54)	

Lime Results	word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)
Correlation for C113						
gear	draft (0.85)	broken (0.8)	block (0.68)	force (0.48)	brake (0.26)	
yoke	broken (0.72)	drawbar (0.62)	system (0.6)	fail (0.37)	drop (0.23)	
drawbar	Yoke (0.62)	fall (0.42)	cause (0.33)	derailment (0.28)		
draft	gear (0.85)	broken (0.65)	block (0.48)	switch (0.3)	coupler (0.25)	
miss	retainer (0.36)	key (0.31)	pincross (0.29)	inspection (0.28)		
Correlation for C118						
Stiff	bolster (0.45)	improper (0.41)	swivel (0.39)	truck (0.28)		
insufficient	clearnace (0.57)	bear (0.51)				
climb	flange (0.41)	flange (0.39)	top (0.3)	curve (0.27)	rail (0.22)	
friction	wear (0.48)	limit (0.25)	undesired (0.23)	wedge (0.22)	plate (0.21)	
bear	clearance (0.68)	insufficient (0.51)	side (0.33)	excessive (0.25)		
Correlation for C119						
buildup	slag (0.45)	tread (0.4)	excessive (0.25)			
tread	build (0.52)	buildup (0.4)				
thin	flange (0.36)	pick (0.22)				
rim	brake (0.39)	traverse (0.3)	railcar (0.3)	single (0.23)		
flange	wear (0.45)	thin (0.36)	point (0.29)	sharp (0.26)	adjacent (0.23)	Switch (0.21)

APPENDIX B. CAUSE 2 LIME RESULTS AND WORD ASSOCIATION FOR EACH SUB-CATEGORY

Lime Results	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)
			Correlation for C220			
extreme	velocity (0.71)	environmental (0.65)	condition (0.46)	wind (0.43)	single (0.26)	
velocity	extreme (0.71)	environmental (0.46)	wind (0.45)	traverse (0.3)	impact (0.22)	
tornado	warning (0.44)	pass (0.28)	blow (0.23)	warn (0.22)		
weather	severe (0.32)	warn (0.28)	gust (0.28)	temperature (0.25)	wind (0.24)	strong (0.21)
gust	wind (0.42)	effect (0.40)	mph (0.35)	weather (0.28)	high (0.21)	
			Correlation for C221			
improperly	load (0.34)					
load	improperly (0.34)	empty (0.27)	shift (0.26)	car (0.22)		
harmonic	rock (0.55)	level (0.35)	speed (0.29)	lateralvertical (0.25)		
interaction	lateralvertical (0.82)	force (0.4)	harmonic (0.21)	rock (0.2)		
lateralvertical	interaction (0.82)	forces (0.39)	harmonic (0.25)			

APPENDIX C. CAUSE 3 LIME RESULTS AND WORD ASSOCIATION FOR EACH SUB-CATEGORY

Lime Results	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)
			Correlation for C330		
gapped	point (0.62)	switch (0.42)	account (0.28)	stock (0.21)	
stock	rail (0.73)	point (0.56)	switch (0.43)	gapped (0.21)	worn (0.21)
rod	connect (0.76)	yarding (0.22)	old (0.21)	switch (0.2)	
stand	defective (0.43)	switch (0.32)	strike (0.22)	adjacent (0.21)	
pass	switch (0.52)	point (0.38)	break (0.26)	strike (0.21)	
			Correlation for C332		
break	rail (0.62)	bar (0.41)	fracture (0.31)	veritcal (0.26)	piece (0.25)
bar	joint (0.57)	break (0.41)	defect (0.34)	angle (0.21)	
fracture	detail (0.77)	break (0.31)	rail (0.25)	curve (0.21)	
split	vertical(0.62)	head (0.25)			
piece	rail(0.56)	break (0.25)	section (0.21)	foot (0.2)	
			Correlation for C333		
settle	roadbed (0.49)	single (0.38)	soft (0.37)	material (0.35)	traverse (0.33)
temporary	speed (0.45)	mph (0.33)	right (0.3)	encounter (0.3)	bridge (0.21)
washout	short (0.35)	encounter (0.35)	flood (0.32)	heavy (0.26)	rain (0.22)
flood	short (0.37)	wash (0.35)	washout (0.32)	rain (0.27)	heavy (0.2)
soft	roadbed (0.54)	settle (0.37)	bed (0.26)	track (0.21)	
			Correlation for C334		
thermal	misalignment (0.82)	traverse (0.32)			
buckle	track (0.41)	sunkink (0.35)	alignment (0.32)	track (0.25)	
wide	gage (0.67)	gauge (0.5)	miss (0.38)	defective (0.26)	crossties (0.22)
alignment	irregular (0.53)	buckle (0.32)	traverse (0.26)	single (0.23)	
defective	crossties (0.49)	miss (0.46)	wide (0.26)	tie (0.2)	

APPENDIX D. CAUSE 5 LIME RESULTS AND WORD ASSOCIATION FOR EACH SUB-CATEGORY

Lime Results	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)
Correlation for C550							
thermal	misalignment (0.82)	traverse (0.32)					
buckle	track (0.41)	sunkink (0.35)	alignment (0.32)	track (0.25)			
wide	gage (0.67)	gauge (0.5)	miss (0.38)	defective (0.26)	crossties (0.22)		
alignment	irregular (0.53)	buckle (0.32)	traverse (0.26)	single (0.23)			
defective	crossties (0.49)	miss (0.46)	wide (0.26)	tie (0.2)			
Correlation for C553							
failure	control (0.72)	improperly (0.42)	remove (0.31)	derail (0.27)	shove (0.23)		
Shove	track (0.33)	protect (0.32)	end (0.3)	control (0.24)	failure (0.23)		
split	reverse (0.27)	point (0.22)	switch (0.2)				
protect	shove (0.32)	point (0.26)	fail (0.23)	move (0.22)			
remove	derail (0.47)	failure (0.31)	derail (0.29)	move (0.21)			
Correlation for C556							
speed	excessive (0.44)	rock (0.37)	harmonic (0.37)	comply (0.37)	excced (0.36)	mph (0.29)	restriction (0.35)
improperly	line (0.66)	switch (0.63)	failure (0.58)	comply (0.52)	speed (0.22)		
comply	failure (0.92)	restrict (0.74)	improperly (0.52)	speed (0.37)			
harmonic	rock (0.77)	mph (0.43)	speed (0.37)	exceed (0.25)	curve (0.23)		
mph	restriction (0.48)	harmonic (0.43)	rock (0.4)	mile (0.37)	speed (0.29)	high (0.28)	exceed (0.21)
Correlation for C557							
Switch	line (0.48)	point (0.4)	run (0.37)	through (0.3)	crossover (0.21)		
through	run (0.74)	previously (0.31)	switch (0.31)	crossover (0.26)			
line	Switch (0.48)	movement (0.42)	improperly (0.24)	clear (0.21)			
pull	track (0.33)	start (0.33)	car (0.28)	back (0.25)	tie (0.21)	block (0.2)	
latch	not (0.29)	lock (0.28)	secure (0.26)	switch (0.22)	properly (0.21)		

Lime Results	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)	Word (Corr)
Correlation for C558							
buff	force (0.49)	excessive (0.34)	action (0.22)				
brake	use (0.53)	dynamic (0.48)	automatic (0.45)	Slack (0.24)	block (0.2)		
descend	grade (0.42)	bridge (0.26)	train (0.22)	trail (0.2)			
excessive	action (0.39)	slack (0.37)	buff (0.34)	force (0.29)			
dynamic	break (0.48)	slow (0.3)	speed (0.27)	throttle (0.25)			

APPENDIX E. FACTORS RESPONSIBLE FOR CAUSING TRS-ACCIDENTS AND ACCIDENT COUNTS

S.No	Code	Description	Count
1	T102	Cross level of track irregular (not at joints)	161
2	T220	Broken Rail - Transverse/compound fissure	371
3	T110	Wide gage (due to defective or missing crossties)	1057
4	T101	Cross level of track irregular (at joints)	118
5	T319	Switch point gapped (between switch point and stock rail)	180
6	T311	Switch damaged or out of adjustment	195
7	T404	Catenary system defect	253
8	T001	Roadbed settled or soft	185
9	T111	Wide gage (due to defective or missing spikes or other rail fasteners)	300
10	T221	Broken Rail - Vertical split head	257
11	T207	Broken Rail - Detail fracture from shelling or head check	480
12	T315	Switch rod worn, bent, broken, or disconnected	12
13	T201	Broken Rail - Bolt hole crack or break	90
14	T399	Other frog, switch and track appliance defects (Provide detail	91
15	T403	Engineering design or construction	64
16	T210	Broken Rail - Head and web separation (outside joint bar limits)	241
17	T206	Defective spikes or missing spikes or other rail fasteners (use code T111 if results in wide gage)	77
18	T299	Other rail and joint bar defects (Provide detailed description in narrative)	78
19	T301	Derail, defective	11
20	T309	Switch (hand operated) stand mechanism broken, loose, or worn	55
21	T205	Defective or missing crossties (use code T110 if results in wide gage)	106
22	T314	Switch point worn or broken	483
23	T214	Joint bar broken (insulated)	12
24	T002	Washout/rain/slide/flood/snow/ice damage to track	52
25	T199	Other track geometry defects (Provide detailed description in narrative)	87
26	T112	Wide gage (due to loose, broken, or defective gage rods)	41
27	T212	Broken Rail - Horizontal split head	56
28	T317	Turnout frog (self guarded), worn or broken	22
29	T305	Retarder worn, broken, or malfunctioning	63
30	T103	Deviation from uniform top of rail profile	41
31	T308	Stock rail worn, broken or disconnected	35
32	T499	Other way and structure defect (Provide detailed description in narrative)	37
33	T202	.Broken Rail - Base	200
34	T208	Broken Rail - Engine burn fracture	10
35	T222	Worn rail	52
36	T313	Switch out of adjustment because of insufficient rail anchoring	45
37	T303	Guard rail loose/broken or mislocated	38
38	T211	Broken Rail - Head and web separation (within joint bar limits)	35
39	T099	Other roadbed defects (Provide detailed description in narrative)	18
40	T217	Mismatched rail-head contour	47
41	T204	Broken Rail - Weld (field)	40
42	T108	Track alignment irregular (other than buckled/sunkink)	116
43	T401	Bridge misalignment or failure	26
44	T113	Wide gage (due to worn rails)	118
45	T203	Broken Rail - Weld (plant)	12
46	T316	Turnout frog (rigid) worn, or broken	30
47	T109	Track alignment irregular (buckled/sunkink)	221
48	T213	Joint bar broken (compromise)	19
49	T106	Superelevation improper, excessive, or insufficient	43

S.No	Code	Description	Count
50	T307	Spring/power switch mechanism malfunction	17
51	T216	Joint bolts, broken, or missing	21
52	T215	Joint bar broken (noninsulated)	27
53	T310	Switch connecting or operating rod is broken or defective	24
54	T402	Flangeway clogged	15
55	T304	Railroad crossing frog, worn or broken	7
56	T312	Switch lug/crank broken	17
57	T302	Expansion joint failed or malfunctioned	2
58	T223	Rail Condition - Dry rail, freshly ground rail.	5
59	T104	Disturbed ballast section	4
60	T218	Broken Rail - Piped rail	2
61	T219	Rail defect with joint bar repair	7
62	T318	Turnout frog (spring) worn, or broken	4
63	T107	Superelevation runoff improper	6
64	T306	Retarder yard skate defective	12
65	T105	Insufficient ballast section	4
		Total	6,555